PROPER DATASET VALUATION BY POINTWISE MUTUAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Data plays a central role in the development of modern artificial intelligence, with high-quality data emerging as a key driver of model performance. This has prompted the development of various data curation methods in recent years. However, measuring the effectiveness of these data curation techniques remains a major challenge. Traditional evaluation methods, which assess a trained model's performance on specific benchmarks, risk promoting practices that merely make the data more similar to the test data. This issue exemplifies Goodhart's law: when a measure becomes a target, it ceases to be a good measure. To address this, we propose an information-theoretic framework for evaluating data curation methods, where dataset quality is measured by its informativeness about the true model parameters using the Blackwell ordering. We compare informativeness by the *Shannon mutual information* of the evaluated data and the test data, and we propose a novel method for estimating the mutual information of datasets by training Bayesian models on embedded data and computing the mutual information from the model's parameter posteriors. Experiments on real-world data demonstrate that our mutual information-based evaluation assigns appropriately lower scores to data curation strategies that reduce dataset informativeness, while traditional test score-based evaluation methods may favor data curation strategies that overfit to the test set but compromise the training data's informativeness.

1 INTRODUCTION

Data plays a central role in the development of modern artificial intelligence, where the large volume and high quality of the data used in training are critical to model performance (Brown et al., 2020; Peebles & Xie, 2023; Team et al., 2024; MetaAI, 2024). As AI systems continue to grow larger and the computational costs of training escalate, the focus is shifting from simply expanding model and dataset sizes to enhancing the quality of the data itself. This shift has prompted the development of various data curation strategies, including data filtering (Gunasekar et al., 2023; Li et al., 2023; Fang et al., 2023; Pouget et al., 2024), duplicate removal (Kandpal et al., 2022), data augmentation (Muennighoff et al., 2024), and synthetic data generation (Liu et al., 2024).

However, ensuring the effectiveness of these curation techniques remains a major challenge (Li et al., 2024; Weber et al., 2024). The standard evaluation approach involves training a model on a curated dataset and measuring its performance against benchmark test sets (Li et al., 2024; Albalak et al., 2024; Team et al., 2024; MetaAI, 2024). This methodology, though common, can inadvertently encourage undesirable curation practices that optimize performance on specific benchmarks, yet risk overfitting to the test data and undermining the model's ability to generalize to new data. For instance, as noted by (Pouget et al., 2024), popular pre-training methods often filter datasets to emphasize English-language image-text pairs in order to maximize performance on western-oriented benchmarks like ImageNet and COCO. While this may improve performance on those benchmarks, it degrades performance on global datasets. This illustrates a critical issue: as highlighted by Goodhart's law, when a measure becomes a target, it ceases to be a good measure.

An important question, therefore, is how to distinguish between data curation methods that simply boost a trained model's performance on specific benchmarks and those that genuinely enhance data quality and improve the model's ability to generalize to new data. In this work, we propose an alternative information-theoretic framework that may help make this distinction: rather than measuring the test score of a trained model on specific test sets, we evaluate the *informativeness* of a dataset for a given machine learning task. To achieve this, we adopt the well-known Blackwell ordering Blackwell et al. (1951) to compare the informativeness of datasets. A data curation method is considered effective if it increases the dataset's informativeness about the true model parameters, while it is deemed *strategic* if it decreases the dataset's informativeness, according to the Blackwell ordering.

To quantify informativeness, we propose using the *Shannon mutual information* (MI) of the curated dataset and the test dataset as a metric. This mutual information metric is effective in identifying data curation methods that reduce informativeness. Specifically, if a curation method decreases the dataset's informativeness according to the Blackwell ordering, it must lead to a decrease in the mutual information. However, estimating mutual information of *datasets* presents a challenge in practice. Existing techniques can reliably estimate the mutual information of two random variables in up to tens of dimensions but fail in higher dimensions (Gowri et al., 2024).[1] This is particularly problematic, as datasets are inherently high-dimensional due to the many data points they contain.

Our main technical contribution is a novel method for estimating the mutual information of two datasets. We exploit a dataset's capacity to train a machine learning model and compute the mutual information through the *posterior distributions* of the model parameters. To get such posteriors, we reduce model size by utilize embeddings from pre-trained foundation models. Our method consists of two steps: (1) using pre-trained foundation models to embed data examples, and (2) training relatively small Bayesian models, such as Bayesian logistic regression, on the resulting embeddings to estimate the mutual information of the datasets.

We demonstrate the effectiveness of our method through experiments on real-world datasets, including MNIST and CIFAR. Our experiments reveal that the test score-based evaluation method may favor data curation strategies that make the dataset more similar to the test data but reduce its informativeness about the true model parameters. In contrast, our mutual information-based evaluation assigns appropriately lower scores to such strategies. This is because our method accurately estimates the mutual information of datasets, as verified by our experiments.

To summarize, our contribution is threefold:

- We propose an information-theoretic framework for evaluating data curation methods, where dataset quality is measured by its informativeness about the true model parameters using the well-established Blackwell ordering.
- A novel method is introduced for estimating the mutual information of datasets by training a Bayesian model on embedded data and computing the MI from the parameter posteriors.
- Experiments on real-world data show that our mutual information-based evaluation, unlike the test score-based evaluation, assigns appropriately lower scores to informativeness-reducing data curation methods by accurately estimating the mutual information of datasets.

1.1 Related work

Data curation. The success of large language models (LLMs) is fundamentally anchored in the quality of their training datasets, which underscoring the critical need for advancements in data curation to ensure optimal training efficiency, cost-effectiveness and robust generalization. Data filtering (Gunasekar et al., 2023; Li et al., 2023; Fang et al., 2023; Pouget et al., 2024; Xie et al., 2023) aims to select data points to include in the training dataset from a large pool of raw data, often guided by various heuristics. Duplicate removal (Kandpal et al., 2022) focuses on repeated occurrences and impact of sequences within training datasets. The findings underscore the importance of sequence-level deduplication in training efficiency and model privacy without sacrificing model performance. Data augmentation (Muennighoff et al., 2024; Törnberg, 2023) generates new training samples from the original dataset to enhance its diversity and volume while preserving its core characteristics while synthetic data generation (Liu et al., 2024) creates totally new data that closely resemble the distribution of real data. Data mixing (Xie et al., 2024; Liu et al., 2025) determines the weight of each domain's dataset to optimize performance across all domains. Data distillation

[1]Considering that the intrinsic dimension of typical images ranges from approximately 20 to 43, as reported by (Pope et al., 2021), these nonparametric methods will not be reliable when applied to datasets containing more than just a few images.

(Sachdeva & McAuley, 2023) aims to create compact, high-fidelity data summaries that capture the most essential knowledge from a given target dataset.

Data point valuation. The assessment of data point value has been actively studied in the data valuation literature. A standard approach is to measure the change in the test accuracy after removing a single training data point of interest. Data Shapley by Ghorbani & Zou (2019) deploys the Shapley value from cooperative game theory to ML settings, and several variants that improve its computational efficiency or relax underlying conditions have been proposed (Jia et al., 2019a; Kwon & Zou, 2021; Wang & Jia, 2022; Wang et al., 2024a). An alternative common approach utilizes the influence function introduced in robust statistics (Koh & Liang, 2017; Feldman & Zhang, 2020). This method provides a mathematically rigorous interpretation of data values and has been implemented in various applications, such as image classification and sentiment analysis, or text-to-image generation (Park et al., 2023; Kwon et al., 2023). Other algorithm-agnostic and task-agnostic methods have also been explored, such as (Just et al., 2023; Xu et al., 2021). We refer the readers to Jiang et al. (2023) for a comprehensive and detailed review.

Dataset valuation. Beyond evaluating individual data points, various methods have been proposed for dataset evaluation. The standard approach involves training a model on a curated dataset and measuring its performance on benchmark test sets (Li et al., 2024; Albalak et al., 2024). Garrido-Lucero et al. (2024) leverages estimated Shapley values for efficient dataset valuation. Mohammadi Amiri et al. (2023) focus on intrinsic, task-agnostic dataset valuation by estimating data diversity and relevance without requiring a validation set. However, none of these methods provide the information-theoretic guarantees as we do.

Peer prediction approach. Our method is also connected to the *peer prediction* literature Miller et al. (2005); Prelec (2004); Jurca & Faltings (2008); Radanovic & Faltings (2013; 2014); Witkowski & Parkes (2012); Kong & Schoenebeck (2018a); Schoenebeck & Yu (2020b), which studies eliciting truthful information without ground truth. Among these, Kong & Schoenebeck (2018b); Chen et al. (2020); Schoenebeck & Yu (2020a) are most relevant. Kong & Schoenebeck (2018b) proposed a mutual information-based peer prediction method using two agents' predictions about a latent label, later adapted for data valuation by Chen et al. (2020). However, their approach computes pointwise mutual information through a complex integral involving the product of two posteriors divided by the prior (see Appendix B for details). Schoenebeck & Yu (2020a) also estimates mutual information but is restricted to discrete variables or specific continuous distributions.

Mutual information estimation. Mutual information (MI) is a key concept in data science that measures the statistical dependence between random variables. Non-parametric methods, such as binning, likelihood-ratio estimators with support vector machines, and kernel-density estimators, are commonly used (Fraser & Swinney, 1986; Darbellay & Vajda, 1999; Kraskov et al., 2004) for estimating mutual information, but these approaches often do not scale well with sample size and data dimension (Gao et al., 2015). Variational methods, such as MINE (Belghazi et al., 2018b) and InfoNCE (Oord et al., 2018), have become popular alternatives. Recent work by Gowri et al. (2024) shows that while standard MI estimators perform well in up to tens of dimensions, they are not reliable in higher dimensions when the available data is limited. To mitigate this, they suggest reducing dimensionality with pre-trained models before MI estimation, which improves scalability.

2 MODEL

Consider a machine learning task with a model parameterized by $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. We assume the Bayesian perspective, where $\boldsymbol{\theta}$ is drawn from an underlying prior distribution $p(\boldsymbol{\theta})$. Suppose we have a test dataset $T = (\mathbf{x}_T^{(1)}, \dots, \mathbf{x}_T^{(N_T)})$, consisting of N_T i.i.d. data points drawn from an underlying distribution $p(\mathbf{x}_T|\boldsymbol{\theta})$, and an original dataset $D = (\mathbf{x}_D^{(1)}, \dots, \mathbf{x}_D^{(N_D)})$, with N_D i.i.d. data points from an underlying $p(\mathbf{x}_D|\boldsymbol{\theta})$. The two datasets may not follow the same distribution, so $p(\mathbf{x}_D|\boldsymbol{\theta})$ need not equal $p(\mathbf{x}_T|\boldsymbol{\theta})$. Denote the support of D and T by $\mathcal{D}$ and $\mathcal{T}$, respectively.

We aim to evaluate different data curation methods, which can be seen as functions applied to the original dataset, possibly incorporating additional information to improve the data. This additional information is represented by a random variable A, which may be correlated with both the model parameter $\boldsymbol{\theta}$ and the dataset D.

Definition 2.1 (Data curation method)**.** Let A be a random variable representing additional information for data curation, and let $\mathcal{A}$ be the support of A. A data curation method with additional information A is a function $f : \mathcal{A} \times \mathcal{D} \to \mathcal{D}$ that outputs a modified dataset $f(A, D)$ given A and an original dataset $D \in \mathcal{D}$. The space of such functions is denoted by $\mathcal{F}$.

Below are several examples of data curation methods:

- **Adding new data.** A simple data curation method is adding new data, where $A \in \mathcal{D}$ represents the new data and $f(A, D) = D \cup A$.
- **Deleting data.** It is also common to select a subset of data and remove the others, as seen in coreset selection Mirzasoleiman et al. (2020), data filtering with quality signals (Gunasekar et al., 2023; Li et al., 2023; Fang et al., 2023; Pouget et al., 2024), data deduplication (Kandpal et al., 2022), and removing low-quality or out-of-domain data Northcutt et al. (2021); Ghorbani & Zou (2019); Jia et al. (2019b). In data deletion, the additional information can be represented as a random vector $A \in \{0, 1\}^{N_D}$ indicating whether each data point is retained or removed.
- **Reweighting data.** Another commonly used method is resampling data points with different weights Xie et al. (2024); Xu et al. (2024). In this case, the additional information can represented by a random vector $A \in \mathbb{N}^{N_D}$ indicating the number of copies of each data point in the final dataset.

To distinguish between methods that merely adapt the dataset to be more similar to the test data and those that introduce meaningful improvements, we employ the Blackwell ordering of informativeness.

Definition 2.2 (Blackwell order of informativeness Blackwell et al. (1951))**.** If random variables $X \to Y \to Z$ form a Markov chain, then Z is *less informative* than Y about X.

In particular, suppose we have a data curation method $f(A, D)$ that reduces the informativeness of D about the true model parameter $\boldsymbol{\theta}$ in the Blackwell order, i.e., $\boldsymbol{\theta} \to D \to f(A, D)$ forms a Markov chain. Then by Blackwell's theorem, the best model trained on D can achieve an expected loss that is at least as low as the best model trained on $f(A, D)$.

Theorem 2.3 (Informal, Blackwell et al. (1951))**.** *Suppose $\boldsymbol{\theta} \to D \to f(A, D)$ forms a Markov chain. Consider the decision problem of selecting a hypothesis/trained model h from a hypothesis/model class $\mathcal{H}$ to minimize the expected loss using a dataset. Then, the minimum expected loss achievable using D is at least as low as that achievable using $f(A, D)$.*

We defer the formal version of this theorem to Theorem A.5.

We thus define such curation methods that reduce informativeness as *strategic data curation methods*. A data curation method is considered *strategic* if the resulting dataset is less informative about the true model parameter $\boldsymbol{\theta}$ according to the Blackwell ordering.

Definition 2.4 (Strategic data curation)**.** A data curation method $f(\cdot)$ is strategic if the curated dataset $f(A, D)$ is less informative about $\boldsymbol{\theta}$ than the original dataset D. Formally, $\boldsymbol{\theta} \to D \to f(A, D)$ forms a Markov chain.

Below are several examples of strategic curation methods:

- **Adding fake data.** When adding new data A, if A consists of i.i.d. data points from $p(\mathbf{x}_D|\boldsymbol{\theta})$, then $f(A, D)$ is more informative because it does not form a Markov chain. However, if A contains randomly generated fake data, $f(A, D)$ becomes strategic, as $\boldsymbol{\theta} \to D \to f(A, D)$ forms a Markov chain.
- **Deleting or reweighting data without additional signals.** When deleting or reweighting data, if $A \in \mathbb{N}^{N_D}$ is guided by some additional quality or relevance signal, such as oracle information identifying incorrect/irrelevant labels, the filtered/reweighted dataset can be more informative. Conversely, if A is decided solely from the observed dataset D without utilizing new signals—i.e., when there exist functions $h_D(A)$ that determine the distribution of A given a dataset D—the resulting dataset $f(A, D)$ will be less informative. Because we have $p(A|D, \boldsymbol{\theta}) = h_D(A) = p(A|D)$, indicating that $\boldsymbol{\theta} \to D \to A$ forms a Markov chain, and thus $\boldsymbol{\theta} \to D \to f(A, D)$ forms a Markov chain.
- **Deleting or reweighting data by non-essential features.** In addition, when deleting or reweighting data, if $A \in \mathbb{N}^{N_D}$ is based on some non-essential feature that is non-predictive of the label, the resulting dataset will be less informative. To be more specific, suppose a data point $\mathbf{x} = (\mathbf{z}, y)$ in D consists of a label y and essential features $\mathbf{z}$. Suppose there is some non-essential feature z_N that satisfies $p(y|\boldsymbol{\theta}, \mathbf{z}, z_N) = p(y|\boldsymbol{\theta}, \mathbf{z})$ and is non-predictive

of y conditioned on $\mathbf{z}$, as illustrated in the graphical model in Figure 1. Then if the vector $A \in \mathbb{N}^{N_D}$ is decided by this non-essential feature of the data points, $z_N^{(1)}, \ldots, z_N^{(N_D)}$ (as well as D), then the resulting dataset will be less informative because $z_N^{(1)}, \ldots, z_N^{(N_D)}$ are independent of $\boldsymbol{\theta}$ conditioned on D (as the path between $\boldsymbol{\theta}$ and z_N is blocked by $\mathbf{z}$ and d-separation implies conditional independence).

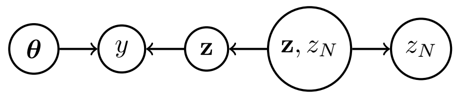

Figure 1: Graphical model for non-essential features.

For simplicity, we sometimes omit the dependency on A and use $f(D)$ to represent a data curation method.

We assess a data curation method by assigning it a score, with the goal of distinguishing between methods that increase or reduce informativeness. A scoring function for data curation methods is defined as follows.

Definition 2.5 (Scoring function for data curation methods)**.** A scoring function for data curation methods $S : \mathcal{F} \to \mathbb{R}$ assigns a score $S(f)$ to a data curation method $f(\cdot)$, given access to the original data D and test data T.

Our goal is to design a scoring function that does not encourage strategic data curation methods. Specifically, we seek a function that assigns lower scores to strategic methods than to the case of no modification.

Definition 2.6 (Strategy-proof scoring functions)**.** A scoring function $S(f)$ for data curation methods is strategy-proof if it ensures that strategic data curation methods always receive a score no higher than the identity function $f(D) \equiv D$, while $S(f)$ itself is non-constant.

3 PMI Scoring Function

We propose a strategy-proof scoring function that measures the *Shannon mutual information* (MI) of the curated datasets and test datasets. To estimate the mutual information, we leverage pre-trained models to embed data points and then build a Bayesian model, based on which we introduce a closed-form formula for approximating the *pointwise mutual information* (PMI) of datasets. In this section, we omit the dependency on A and use $f(D)$ to represent a data curation method for simplicity.

3.1 Method

We first introduce the key steps of our approach.

Mutual information as the metric. Due to the data processing inequality, the simplest metric that would yield a strategy-proof scoring function, if computable, is the *Shannon mutual information* of the model parameter $\boldsymbol{\theta}$ and the curated dataset $f(D) = \widehat{D}$, denoted by $I(\boldsymbol{\theta}, f(D))$.

Lemma 3.1 (Data processing inequality)**.** *If $\boldsymbol{\theta} \to D \to \widehat{D}$ form a Markov chain, then $I(\boldsymbol{\theta}, D) \geq I(\boldsymbol{\theta}, \widehat{D})$, where $I(X, Y)$ is the Shannon mutual information of X and Y. Therefore if we use $I(\boldsymbol{\theta}, f(D))$ to score $f(\cdot)$, a strategic $f(\cdot)$ will not receive a score higher than the identity function $f(D) \equiv D$.*

However, since the underlying true model parameter is not observable, we propose using the Shannon mutual information of the curated dataset and the observable test dataset T as the scoring function, which serves as a strategy-proof scoring function as well.

Proposition 3.2. *The* Shannon mutual information *of the curated dataset $f(D) = \widehat{D}$ and the test dataset T, if computable, is a strategy-proof scoring function. The Shannon mutual information $I(\widehat{D}, T) = \mathbb{E}_{\widehat{D},T}\left[\log \frac{p(\widehat{D},T)}{p(\widehat{D})p(T)}\right]$ is defined as the expectation of the* pointwise mutual information *$PMI(\widehat{D}, T) = \log \frac{p(\widehat{D},T)}{p(\widehat{D})p(T)}$, where the expectation is taken over the joint distribution $p(\widehat{D}, T)$ that is induced by the data generating process described in Section 2.*

The proof is deferred to Appendix C.1.

Bayesian modeling on embedded data. Then the problem boils down to estimating the mutual information of two *datasets*. Estimating mutual information of high-dimensional variables is challenging in practice. Existing techniques such as (Kraskov et al., 2004; Belghazi et al., 2018b; Oord et al., 2018) can reliably estimate MI in up to tens of dimensions, but fail in higher dimensions (Gowri et al., 2024). However, a dataset contains many data points, which inevitably boosts the dimension even with low-dimensional representation of data points. As a result, our problem of estimating mutual information between datasets introduces significant new challenges. To address this, we propose a novel method that leverages a dataset's ability to train a *machine learning model*.

The proposed method proceeds as follows. First, recall that the mutual information is the expectation of the pointwise mutual information. We generate k dataset pairs $(D_1, T_1), \ldots, (D_k, T_k)$ and use the average pointwise mutual information $\frac{1}{k}\sum_{i=1}^{k} PMI(f(D_i), T_i)$ to estimate the mutual information $I(f(D), T)$.

Next, we leverage widely-used large pretrained models to generate embeddings of the data, which are then used to train smaller Bayesian models with a specified prior $p(\boldsymbol{\theta})$ and specified likelihoods $p(\mathbf{x}_D|\boldsymbol{\theta})$, $p(\mathbf{x}_T|\boldsymbol{\theta})$ (such as Bayesian logistic regression, Bayesian linear regression, or a Bayesian multilayer perceptron). We assume that the true data generating process is well modeled by this Bayesian model on data embeddings.

Assumption 3.3. We assume that the true data generating process is adequately captured by applying a tractable Bayesian model parametrized by $\boldsymbol{\theta}$ to embeddings generated by a pretrained model, with a specified prior $p(\boldsymbol{\theta})$ and specified likelihoods $p(\mathbf{x}_D|\boldsymbol{\theta})$, $p(\mathbf{x}_T|\boldsymbol{\theta})$ for the resulting embeddings.

For simplicity, we still use D and T to represent the embedded data and then utilize this Bayesian model on embedded D, T to estimate the PMI of datasets.

Closed-form approximation of pointwise mutual information. Even with a Bayesian model, computing $PMI(f(D_i), T_i) = \log \frac{p(f(D_i), T_i)}{p(f(D_i))p(T_i)} = \log \frac{p(T_i|f(D_i))}{p(T_i)}$ remains non-trivial, particularly because the marginal probability $p(f(D_i)), p(T_i)$ is often intractable (which leads to intractable posterior $p(\boldsymbol{\theta}|\cdot)$ as well) for most of the Bayesian models. Extensive research in Bayesian machine learning has focused on estimating the posterior distribution of model parameters $p(\boldsymbol{\theta}|\cdot)$. For instance, methods such as Laplace approximation, variational inference aim to approximate the posterior $p(\boldsymbol{\theta}|\cdot)$ by tractable distributions. But even with a tractable approximated posterior $p(\boldsymbol{\theta}|\cdot)$, the *posterior predictive* $p(T_i|f(D_i)) = \int_\theta p(T_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|f(D_i))\,d\boldsymbol{\theta}$ is still intractable for most models including logistic regression. The computation of the *posterior predictive* $p(T_i|f(D_i))$ requires further approximation such as Monte Carlo integration or likelihood function approximation (see Section 4.1 for detailed discussion and experiments). Building on the vast literature on approximating $p(\boldsymbol{\theta}|\cdot)$, our main technical contribution is a closed-form formula for the PMI when $p(\boldsymbol{\theta}|\cdot)$ is approximated by a tractable distribution, bypassing the further approximation of the posterior predictive.

Theorem 3.4 (PMI dataset score)**.** *Let $\widehat{D}_i = f(D_i)$ be the curated datasets, and let $p(\boldsymbol{\theta}|X)$ be the posterior of $\boldsymbol{\theta}$ given a dataset X. Then the pointwise mutual information $PMI(\widehat{D}_i, T_i)$ can be computed as*

$$\begin{aligned} PMI(\widehat{D}_i, T_i) =& U_\eta(\widehat{D}_i, T_i) \\ &:= \log \frac{p(\boldsymbol{\theta} = \eta|\widehat{D}_i) \cdot p(\boldsymbol{\theta} = \eta|T_i)}{p(\boldsymbol{\theta} = \eta) \cdot p(\boldsymbol{\theta} = \eta|\widehat{D}_i, T_i)}, \end{aligned} \tag{1}$$

where η is an arbitrary parameter value in Θ.[2]

The proof of Theorem 3.4 only relies on Bayes' rule and we defer the proof to Appendix C.2.

Our PMI dataset score can be easily computed as long as the posteriors and the prior are approximated by tractable distributions. This makes it applicable to a wide range of commonly-used Bayesian

[2]Note that the value of $U_\eta(\widehat{D}_i, T_i)$ stays the same across all η values, so η can be chosen arbitrarily. Theorem 3.4 mainly builds on the following equation $\frac{p(T|D)}{p(T)} = \frac{p(\boldsymbol{\theta}=\eta|D)\cdot p(\boldsymbol{\theta}=\eta|T)}{p(\boldsymbol{\theta}=\eta)\cdot p(\boldsymbol{\theta}=\eta|D,T)}$ when D and T are independent conditional on $\boldsymbol{\theta}$, which might be of independent interest. Our expression in Theorem 3.4 also allows new interpretations of PMI, which we defer to Appendix D.1.

neural networks, including Gaussian approximation Daxberger et al. (2021); Yang et al. (2023); Blundell et al. (2015); Wang et al. (2024b), Gaussian mixture approximation Blundell et al. (2015), and Dirichlet approximation Hobbhahn et al. (2022).

Algorithm and convergence rate. Combining all the steps, our PMI scoring function for data curation methods can be computed as in Algorithm 1.

Algorithm 1 PMI scoring function

Require: Datasets $(D_1, T_1), \ldots, (D_k, T_k)$, an data curation method $f(\cdot)$ for evaluation, a pre-trained model used to embed the data points, a Bayesian model for the embedded data with tractable approximated posteriors $p(\boldsymbol{\theta}|\cdot)$, a vector $\eta \in \Theta$.

Ensure: A score for the curation method $f(\cdot)$

1: Apply the curation method $f(\cdot)$ on $D_1, \ldots, D_k$ and get the curated datasets $\widehat{D}_1, \ldots, \widehat{D}_k$

2: Use the pre-trained model to embed the datasets $\widehat{D}_1, \ldots, \widehat{D}_k$ and $T_1, \ldots, T_k$.

3: For each pair of embedded $(\widehat{D}_i, T_i)$, compute the pointwise mutual information $U_\eta(\widehat{D}_i, T_i)$ via the approximated posteriors of the Bayesian model parameters $p(\boldsymbol{\theta}|\cdot)$ as in Equation (1).

4: Return $\frac{1}{k}\sum_{i=1}^{k} U_\eta(\widehat{D}_i, T_i)$.

Based on the previous analysis, the algorithm outputs an unbiased estimator of the target metric $I(f(D), T)$, which converges to the true value of $I(f(D), T)$ as k increases.

Corollary 3.5. *The output of Algorithm 1 provides an unbiased estimator of* $I(f(D), T)$. *Assuming that the posteriors are in an exponential family and the datasets have bounded sufficient statistics, we have* $\Pr\left(\left|\frac{1}{k}\sum_{i=1}^{k} U_\eta(\widehat{D}_i, T_i) - I(\widehat{D}, T)\right| \le \varepsilon\right) \ge 1 - \delta$ *when* $k = O(\log(1/\delta)/\varepsilon^2)$, *and we have the expected square error of the estimator decreases as* $O(1/k)$.

The proof is deferred to Appendix C.3. Compared to commonly used MI estimators, our concentration bound is independent of the variable dimension, unlike the bound in Belghazi et al. (2018a), which scales as $O\left(\frac{d\log(\sqrt{d}/\varepsilon)+d+\log(1/\delta)}{\varepsilon^2}\right)$, where d is the variable dimension. Additionally, unlike the methods in (Belghazi et al., 2018a; Kraskov et al., 2004; Oord et al., 2018; Song & Ermon, 2019), our approach guarantees not only consistency but also unbiasedness. We further show the advantages of our method in Section 4.1

4 Experiments

We evaluate the accuracy of our MI estimator and its ability to assess dataset informativeness through experiments on real-world data. Our results demonstrate that the PMI scoring function remains effective even when employing the simple Bayesian logistic regression model with Gaussian approximation (outlined in Appendix C.4) for Bayesian modeling. The Gaussian posterior approximation can be efficiently computed by training a standard logistic regression model with L2 regularization or by employing the Laplace approximation method in Daxberger et al. (2021).

4.1 Accuracy of Mutual Information Estimation

We evaluate our method on resampled real-world data.

Dataset generation. We resample datasets from MNIST and estimate their mutual information, where the exact value of mutual information is unknown but their relative rankings can be inferred. To assess the accuracy of our method, we measure the rank correlation between the estimated and true rankings. The setup is as follows. We randomly sample dataset pairs containing images of 0s and 1s from MNIST. First, we randomly select two correlated numbers, $r_D, r_T \in \{0.2, 0.8\}$, distributed as in the following table where ρ is a number between 0.25 and 0.5.

$P(r_D, r_T)$	$r_D = 0.2$	$r_D = 0.8$
$r_T = 0.2$	ρ	$\frac{1}{2} - \rho$
$r_T = 0.8$	$\frac{1}{2} - \rho$	ρ

These numbers represent the proportion of data points with label 0 in datasets D and T, respectively. We then generate D and T by randomly sampling images from MNIST to match the specified label proportions r_D, r_T, resulting in two correlated datasets.

Fact 1. *The mutual information of the generated datasets $I(D,T)$ increases in ρ for $0.25 \leq \rho \leq 0.5$.*

The proof is deferred to Appendix E.1.

Baseline method. Gowri et al. (2024) demonstrated that commonly-used nonparametric methods reliably estimate MI in up to tens of dimensions, but fail in higher dimensions.[3] Consequently, such methods are unsuitable for estimating dataset-level mutual information. We thus focus on parametric methods for approximating $PMI(D,T) = \log \frac{p(T|D)}{p(T)}$, which requires estimating the posterior predictive $p(T|D) = \int_\theta p(T|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta}$. For classification problems, several approaches exist for posterior predictive approximation, including Monte Carlo integration, probit approximation (Gibbs, 1998), and Laplace bridge (Hobbhahn et al., 2022). However, probit approximation and Laplace bridge cannot be used as they only provide posterior predictives for individual data points, $\int_\theta p(\mathbf{x}_T^{(i)}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta}$, whereas we need posterior predictives for an entire dataset, $\int_\theta \prod_i p(\mathbf{x}_T^{(i)}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta}$, where $\mathbf{x}_T^{(i)}$ is the i-th data point in T. As a result, Monte Carlo integration stands as the only viable baseline for our problem.

Setting. We use our method and the Monte Carlo baseline to estimate the mutual information of D, T and assess their accuracy by the rank correlation between the estimated rankings the true ρ rankings. We consider ten values of ρ, corresponding to mutual information values ranging from 0.1 to 1.0 in increments of 0.1. For each ρ, we estimate the mutual information with both our PMI formula and Monte Carlo integration, averaged over the 1,000 dataset pairs. To compute our PMI formula, we train logistic regression models on D and T with L2 regularization parameterized by C, which corresponds to a Gaussian prior $N(0, C \cdot \mathbf{I})$. For Monte Carlo integration, we adopt the same logistic regression model and sample 1000 points from the posterior $\mathbf{p}(\boldsymbol{\theta}|D)$ to estimate the posterior predictive $p(T|D)$.

Results. As shown in Table 1, our PMI estimator consistently achieves significantly higher Kendall τ rank correlation than the baseline, regardless of the choice of regularization strength C. This demonstrates that our method provides far more accurate mutual information estimates, and its ranking estimates are robust to prior misspecification. Additionally, our method runs much faster than the baseline. This indicates that our approach not only provides more accurate results but is also computationally more efficient.

Method	τ	**Runtime (min)**
PMI ($C = 1$)	0.956	75
PMI ($C = 100$)	0.911	76
PMI ($C = 1000$)	0.911	75
Baseline	0.600	739

Table 1: Comparison of Kendall's τ rank correlation and runtime across different methods. PMI ($C = c$) refers to our PMI-based mutual information estimation method, where C denotes the regularization parameter for L2 regularization in logistic regression, corresponding to a Gaussian prior $N(0, C \cdot \mathbf{I})$. The baseline method employs Monte Carlo integration with 1,000 samples per estimate and an optimally selected regularization strength ($C = 100$). For each ρ, both PMI and MC integration are averaged over 1,000 pairs of correlated datasets (D, T), each with 100 images and reduced to 100 dimensions via Principal Component Analysis. Runtime is measured as the total time (in minutes) required to complete all experiments across 10 ρ values on the same machine.

4.2 Evaluating data curation methods

We next use popular datasets to test our PMI scoring function in evaluating data curation methods. We show that the PMI scoring function is effective in distinguishing between strategic and non-strategic curation methods, whereas evaluating curation methods using test scores could promote strategic methods that do not add new information but merely make the data more similar to the test data.

[3] Given that the intrinsic dimension of MNIST images is approximately 10, as reported by (Pope et al., 2021), these nonparametric methods fail when applied to datasets containing more than 10 images.

Data curation methods and dataset generation. There are numerous data curation methods available for evaluation. We select three that can be clearly classified as strategic or non-strategic. To evaluate these methods, we apply these methods to dataset pairs randomly sampled from the training and test sets of Colored MNIST Arjovsky et al. (2020) and Corrupted CIFAR Hendrycks & Dietterich (2019).

- **Data filtering:** We consider the removal of mislabeled data, assuming access to oracle information about the correctness of each label. We consider such data filtering as a non-strategic curation methods that is expected to receive a score higher score than no modification. To generate datasets for filtering, we randomly sample $T_1, \ldots, T_k$ from the test set, and sample datasets from the training set and flip the labels of some data points to generate $D_1, \ldots, D_k$. We compare the scores before and after the removal of these mislabeled data points.
- **Strategic data duplication or removal by non-essential features:** We then consider the duplication/removal of a data subset without using quality or relevance signals but only makes the data more similar to the test data based on non-essential features, such as the brightness of an image (in Corrupted CIFAR) or the color of a figure (in Colored MNIST). This results in a strategic curation method, which should receive a score lower than that of $f(D) \equiv D$. We generate datasets for duplication or removal as follows. Let $\mathbf{z}_E$ represent the essential features from the original MNIST/CIFAR, $z_N \in \{0, 1\}$ be a binary non-essential features introduced in Colored MNIST/Corrupted CIFAR (e.g. color and brightness), and $y \in \{0, 1\}$ be a binary label. We sample pairs of D and T with the same essential feature distribution $p_D(\mathbf{z}_E, y) = p_T(\mathbf{z}_E, y)$ but different compositions of non-essential features $p_D(z_N = 0|\mathbf{z}_E, y) \neq p_T(z_N = 0|\mathbf{z}_E, y)$. We then consider data duplication/removal on D based on the non-essential feature z_N that aligns $p_D(z_N|\mathbf{z}_E, y)$ with $p_T(z_N|\mathbf{z}_E, y)$. Conditioned on D, such duplication/removal is independent of the true $\boldsymbol{\theta}$, making it a strategic curation method.

For both cases, we generate the smallest datasets that achieve reasonable accuracy $\sim 80\% - 90\%$ to avoid overlap.

Scoring functions. We compare our PMI scoring function (Algorithm 1) to the test accuracy baseline that trains a model on the curated dataset and evaluate its accuracy on the test set. Specifically, we define the test accuracy scoring function as: $S_{TS}(f) = \frac{1}{k}\sum_{i=1}^{k} \text{Acc}(\theta(f(D_i)), T_i)$, where $\text{Acc}(\theta(D), T)$ represents the accuracy of the model trained on D when evaluated on the test set T. To compute our PMI score in Algorithm 1, we train different models for Colored MNIST and Corrupted CIFAR. For Colored MNIST, we directly use a logistic regression model. For Corrupted CIFAR, we use pre-trained ResNet18 to extract image embeddings and subsequently train a logistic regression. The logistic regression models are trained using L2 regularization with parameter C, which corresponds to a Gaussian prior $N(0, C \cdot \mathbf{I})$. These models are subsequently employed to evaluate the test accuracy scoring function.

Results. Table 2 and Table 3 present the **changes** in the PMI scoring function and test accuracy after applying the three data curation methods. Across both datasets, our PMI score effectively distinguishes between strategic and non-strategic curation methods: data filtering increases the PMI score, whereas data duplication or removal based on non-essential features leads to a decrease. In contrast, the test accuracy metric fails to detect strategic data duplication and removal, always assigning them higher scores. This suggests that relying solely on test accuracy may inadvertently promote strategic methods that do not introduce new information but merely make the training data more similar to the test data. Furthermore, our findings are robust to prior misspecification (varying choices of C), and we observe the same pattern across different data distributions, as detailed in Appendix E.2.

5 Discussion and Future Work

We propose an information-theoretic framework for evaluating data curation methods that measures data informativeness by the mutual information of the curated data and the test data. We discuss several potential directions for future work. Firstly, a key open problem is to develop principled method for selecting dataset pairs that most effectively estimate mutual information. We have observed that the PMI scoring function can fail when the datasets D_i, T_i are too small to train

C	Operation	PMI Score Change	Accuracy Change(%)
10	**Data Denoising**	6.4621 ± 0.8463	0.91 ± 0.06
50	**Data Denoising**	8.1367 ± 1.0031	0.78 ± 0.11
100	**Data Denoising**	12.7610 ± 1.2069	0.14 ± 0.01
200	**Data Denoising**	13.5874 ± 1.0423	0.31 ± 0.02
10	**Data Duplication**	-2.8150 ± 1.1563	3.84 ± 0.55
50	**Data Duplication**	-1.8428 ± 1.0567	3.30 ± 0.65
100	**Data Duplication**	-1.8762 ± 0.9954	2.75 ± 0.54
200	**Data Duplication**	-2.5927 ± 1.1437	1.43 ± 0.29
10	**Data Removal**	-5.2692 ± 0.7825	0.31 ± 0.02
50	**Data Removal**	-5.5265 ± 0.9823	0.28 ± 0.03
100	**Data Removal**	-6.5826 ± 1.0437	0.34 ± 0.04
200	**Data Removal**	-13.9614 ± 2.0497	0.54 ± 0.04

Table 2: **Changes** in PMI score function and test accuracy after applying three data curation methods to the Colored MNIST dataset. **C** denotes the regularization parameter for L2 regularization in the trained logistic regression models, corresponding to a Gaussian prior $N(0, C \cdot \mathbf{I})$. The training and the test sets consist of $200 - 400$ samples. The **Denoising** method removes flipped data points, while **Duplication** aligns the distribution of non-essential features in the training set with the test set by duplicating a subset of samples. **Removal** achieves the same alignment by discarding data points. We compute the mean changes in PMI scores and test accuracy by averaging results over 1,000 trials, while the variances are further estimated from 10 repeated runs. Details of the experimental setup and results for a different data distribution are provided in Appendix E.2.1.

C	Operation	PMI Score Change	Accuracy Change (%)
10000	**Data Denoising**	1.8112±0.1408	7.24±0.07
30000	**Data Denoising**	1.6553±0.1288	7.25±0.17
50000	**Data Denoising**	1.5311±0.1452	7.29±0.10
100000	**Data Denoising**	1.2920±0.1578	7.37±0.14
10000	**Data Duplication**	-0.6288±0.0275	0.53±0.03
30000	**Data Duplication**	-0.8258±0.0311	0.53±0.03
50000	**Data Duplication**	-0.9025±0.0385	0.58±0.03
100000	**Data Duplication**	-0.9987±0.0501	0.59±0.03
10000	**Data Removal**	-3.8205±0.0892	0.69±0.07
30000	**Data Removal**	-4.4238±0.0847	0.77±0.10
50000	**Data Removal**	-4.6780±0.1171	0.82±0.13
100000	**Data Removal**	-5.0191±0.0969	0.79±0.16

Table 3: **Changes** in PMI score function and test accuracy after applying three data curation methods to the Corrpted CIFAR dataset. **C** denotes the regularization parameter for L2 regularization in the trained logistic regression models, corresponding to a Gaussian prior $N(0, C \cdot \mathbf{I})$. The training and the test sets consist of $120 - 180$ samples. The experiments were repeated 1,000 times to compute the mean changes in PMI scores and test accuracy, and this process was repeated 10 times to estimate the variances. Details of the experimental setup and results for a different data distribution are provided in Appendix E.2.2.

effective models, as well as when they are too large, resulting in significant overlap between datasets that violates the independence assumption. Secondly, the selection of the prior is also crucial. While we observe that the PMI scoring function is robust to prior misspecifications in terms of ranking mutual information, the absolute accuracy of its MI estimates is highly sensitive to the choice of prior. Thirdly, our experiments focus on the simple logistic regression for Bayesian modeling. It remains an open question whether mutual information estimation could be improved by more advanced Bayesian neural networks.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL `https://arxiv.org/abs/1907.02893`.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018a.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018b.

David Blackwell et al. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 26, 1951.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yiling Chen, Yiheng Shen, and Shuran Zheng. Truthful data acquisition via peer prediction. *Advances in Neural Information Processing Systems*, 33:18194–18204, 2020.

Georges A Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Andrew M Fraser and Harry L Swinney. Independent coordinates for strange attractors from mutual information. *Physical review A*, 33(2):1134, 1986.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pp. 277–286. PMLR, 2015.

Felipe Garrido-Lucero, Benjamin Heymann, Maxime Vono, Patrick Loiseau, and Vianney Perchet. Du-shapley: A shapley value proxy for efficient dataset valuation, 2024. URL `https://arxiv.org/abs/2306.02071`.

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

Mark N Gibbs. *Bayesian Gaussian processes for regression and classification*. PhD thesis, Citeseer, 1998.

Gokul Gowri, Xiao-Kang Lun, Allon M Klein, and Peng Yin. Approximating mutual information of high-dimensional variables using learned representations. *arXiv preprint arXiv:2409.02732*, 2024.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

Marius Hobbhahn, Agustinus Kristiadi, and Philipp Hennig. Fast predictive uncertainty for classification with bayesian deep networks. In *Uncertainty in Artificial Intelligence*, pp. 822–832. PMLR, 2022.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas J Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619*, 2019a.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1167–1176. PMLR, 2019b.

Kevin Fu Jiang, Weixin Liang, James Zou, and Yongchan Kwon. Opendataval: a unified benchmark for data valuation. *arXiv preprint arXiv:2306.10577*, 2023.

Radu Jurca and Boi Faltings. Incentives for expressing opinions in online polls. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 119–128, 2008.

Hoang Anh Just, Feiyang Kang, Jiachen T. Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. Lava: Data valuation without pre-specified learning algorithms, 2023. URL https://arxiv.org/abs/2305.00054.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2022.

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

Yuqing Kong and Grant Schoenebeck. Equilibrium selection in information elicitation without verification via information monotonicity. In *9th Innovations in Theoretical Computer Science Conference*, 2018a.

Yuqing Kong and Grant Schoenebeck. Water from two rocks: Maximizing the mutual information. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, pp. 177–194, New York, NY, USA, 2018b. Association for Computing Machinery. ISBN 9781450358293. doi: 10.1145/3219166.3219194. URL https://doi.org/10.1145/3219166.3219194.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*, 2021.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. *arXiv preprint arXiv:2310.00902*, 2023.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training, 2025. URL https://arxiv.org/abs/2407.01492.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*, 2024.

MetaAI. Llama 3: Advancing open foundation models, 2024. URL https://ai.meta.com/blog/meta-llama-3/.

N. Miller, P. Resnick, and R. Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, pp. 1359–1373, 2005.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Mohammad Mohammadi Amiri, Frederic Berdoz, and Ramesh Raskar. Fundamentals of task-agnostic data valuation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9226–9234, Jun. 2023. doi: 10.1609/aaai.v37i8.26106. URL https://ojs.aaai.org/index.php/AAAI/article/view/26106.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. 2012.

Frank Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. No filter: Cultural and socioeconomic diversityin contrastive vision-language models. *arXiv preprint arXiv:2405.13777*, 2024.

D. Prelec. A Bayesian Truth Serum for subjective data. *Science*, 306(5695):462–466, 2004.

Goran Radanovic and Boi Faltings. A robust bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI" 13)*, number EPFL-CONF-197486, pp. 833–839, 2013.

Goran Radanovic and Boi Faltings. Incentives for truthful information elicitation of continuous signals. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI" 14)*, number EPFL-CONF-215878, pp. 770–776, 2014.

Noveen Sachdeva and Julian McAuley. Data distillation: A survey, 2023. URL https://arxiv.org/abs/2301.04272.

Grant Schoenebeck and Fang-Yi Yu. Learning and strongly truthful multi-task peer prediction: A variational approach, 2020a.

Grant Schoenebeck and Fang-Yi Yu. Two strongly truthful mechanisms for three heterogeneous agents answering one question. In *International Conference on Web and Internet Economics*. Springer, 2020b.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Petter Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning, 2023. URL https://arxiv.org/abs/2304.06588.

Jiachen T Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. *arXiv preprint arXiv:2406.11011*, 2024a.

Tianhao Wang and Ruoxi Jia. Data banzhaf: A data valuation framework with maximal robustness to learning stochasticity. *arXiv preprint arXiv:2205.15466*, 2022.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models. *arXiv preprint arXiv:2406.11675*, 2024b.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. Redpajama: an open dataset for training large language models. *arXiv preprint arXiv:2411.12372*, 2024.

Jens Witkowski and David C. Parkes. Peer prediction without a common prior. In Boi Faltings, Kevin Leyton-Brown, and Panos Ipeirotis (eds.), *Proceedings of the 13th ACM Conference on Electronic Commerce, EC 2012, Valencia, Spain, June 4-8, 2012*, pp. 964–981. ACM, 2012. doi: 10.1145/2229012.2229085. URL https://doi.org/10.1145/2229012.2229085.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling, 2023. URL https://arxiv.org/abs/2302.03169.

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. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36, 2024.

Xinnuo Xu, Minyoung Kim, Royson Lee, Brais Martinez, and Timothy Hospedales. A bayesian approach to data point selection, 2024. URL https://arxiv.org/abs/2411.03768.

Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Validation free and replication robust volume-based data valuation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10837–10848. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/59a3adea76fadcb6dd9e54c96fc155d1-Paper.pdf.

Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*, 2023.

A Blackwell ordering

We begin by providing background on the Blackwell order of information structures. We first introduce the formal definitions of decision-making problems and information structures.

Definition A.1. A decision-making problem under uncertainty is defined by the following components:

- **State Space** (Ω): A set of possible states of the world, denoted $\omega \in \Omega$.

- **Action Space** (A): A set of possible actions or decisions, denoted $a \in A$.
- **Utility Function** (u): A function $u : A \times \Omega \to \mathbb{R}$ that quantifies the payoff of taking action a in state ω.
- **Prior Belief** (P): A probability distribution over Ω, representing the decision-maker's initial beliefs. And the corresponding random variable for the state is denoted by W.

An information structure reveals some signal about the state of the world ω.

Definition A.2. An information structure S consists of a pair $(\mathcal{Y}, \pi)$, where:

- $\mathcal{Y}$ is a set of possible signals or observations.
- $\pi : \Omega \to \Delta(\mathcal{Y})$ is a Markov kernel specifying the conditional probability $\pi(y|\omega)$ of observing signal y given state ω. The corresponding random variable representing the signal is denoted by Y.

The decision-maker observes a signal y from the information structure and updates their beliefs about the state ω using Bayes' rule. Based on the updated beliefs, they choose an action a to maximize their expected utility.

The Blackwell order provides a way to compare two information structures in terms of their informativeness, which is defined as follows.

Definition A.3 (Blackwell et al. (1951))**.** Let $S_1 = (\mathcal{Y}_1, \pi_1)$ and $S_2 = (\mathcal{Y}_2, \pi_2)$ be two information structures over a common state space Ω, with the corresponding signals represented by random variables Y_1 and Y_2. We say that S_1 is *more informative* than S_2 in the Blackwell order, if there exists a Markov kernel $\kappa : \mathcal{Y}_1 \to \Delta(\mathcal{Y}_2)$ such that:

$$\pi_2(y_2|\omega) = \sum_{y_1 \in Y_1} \kappa(y_2|y_1)\pi_1(y_1|\omega) \quad \forall y_2 \in \mathcal{Y}_2, \omega \in \Omega,$$

or equivalently $W \to Y_1 \to Y_2$ forms a Markov chain, where W is the random variable representing the state.

In particular, if an information structure S_1 is more informative than S_2 in the Blackwell order, then, by Blackwell's theorem on decision-making superiority, the decision-maker can achieve at least as high an expected utility using S_1 as they can using S_2 for any decision-making problem.

Theorem A.4 (Blackwell's theorem on decision-making superiority Blackwell et al. (1951))**.** *Let $S_1 = (\mathcal{Y}_1, \pi_1)$ and $S_2 = (\mathcal{Y}_2, \pi_2)$ be two information structures over a common state space Ω with the corresponding signals represented by random variables Y_1 and Y_2. The following statements are equivalent:*

1. ***Blackwell Informativeness**: S_1 is more informative than S_2, or equivalently, $W \to Y_1 \to Y_2$ forms a Markov chain.*
2. ***Decision-Making Superiority**: For any decision-making problem (Ω, A, u, P), the maximum expected utility achievable using S_1 is at least as high as that achievable using S_2. Formally:*
$$\max_{a_1 : \mathcal{Y}_1 \to A} \mathbb{E}[u(a_1(y_1), \omega)] \geq \max_{a_2 : \mathcal{Y}_2 \to A} \mathbb{E}[u(a_2(y_2), \omega)],$$
where the expectations are taken over $\omega \sim P$, $y_1 \sim \pi_1(\cdot|\omega)$, and $y_2 \sim \pi_2(\cdot|\omega)$.

We can then apply Blackwell's theorem on decision-making superiority to the problem of data valuation in machine learning. Consider the true underlying model parameter $\boldsymbol{\theta}$ as the state of the world and the dataset D as a signal about $\boldsymbol{\theta}$. Suppose we aim to use D to select a hypothesis or trained model h from a hypothesis/model class $\mathcal{H}$, which serves as the action space. The utility function $u(h, \boldsymbol{\theta})$ represents the negative expected loss when the true model parameter is $\boldsymbol{\theta}$ and the hypothesis/model h is chosen:

$$u(h, \boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x}, y \sim p(\mathbf{x}, y|\boldsymbol{\theta})}[l(h(\mathbf{x}), y)] \triangleq -L(h, \boldsymbol{\theta}),$$

where $l(\cdot)$ is a loss function.

Now, suppose we have a data curation strategy $f(D)$ that reduces the informativeness of the dataset D about $\boldsymbol{\theta}$ in the Blackwell order, i.e., $\boldsymbol{\theta} \to D \to f(D)$ forms a Markov chain. By Blackwell's theorem on decision-making superiority, the decision-maker can achieve at least as low an expected loss using the original dataset D as they can using the curated dataset $f(D)$.

Theorem A.5. *Let $\boldsymbol{\theta}$ be the true underlying model parameter, D_1 be a dataset consisting of data points $(\mathbf{x}, y)$ drawn from $p(\mathbf{x}, y|\boldsymbol{\theta})$, and D_2 be a less informative dataset such that $\boldsymbol{\theta} \to D_1 \to D_2$ forms a Markov chain. Consider the decision problem of selecting a hypothesis/trained model h from a hypothesis/model class $\mathcal{H}$ to minimize the expected loss $\mathbb{E}[l(h(\mathbf{x}), y)]$ using a dataset. Then, the minimum expected loss achievable using D_1 is at least as low as that achievable using D_2. Formally:*

$$\min_{h_1:\mathcal{D}\to\mathcal{H}} \mathbb{E}[L(h_1(D_1), \boldsymbol{\theta})] \leq \min_{h_2:\mathcal{D}\to\mathcal{H}} \mathbb{E}[L(h_2(D_2), \boldsymbol{\theta})],$$

where $L(h, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},y\sim p(\mathbf{x},y|\boldsymbol{\theta})}[l(h(\mathbf{x}), y)]$ represents the expected loss when the true parameter is $\boldsymbol{\theta}$ and the model h is chosen. The expectation is taken over $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$, D_1, and D_2.

B INTEGRAL PMI SCORE

Kong & Schoenebeck (2018b) proposes a method to compute the PMI.

Theorem B.1 (Integral PMI score (Kong & Schoenebeck, 2018b))**.** *The pointwise mutual information $PMI(d, t) = \log \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D = d)p(\boldsymbol{\theta}|T = t)/p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. Therefore the data valuation function $U(d, t) = \log \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D = d)p(\boldsymbol{\theta}|T = t)/p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ is truthful.*

Nonetheless, this integral formulation remains computationally challenging for many basic Bayesian machine learning scenarios. Chen et al. (2020) introduced a theoretical framework for evaluating the integral score specifically within exponential family distributions; however, applying their approach is non-trivial. Computing their normalization function $g(\cdot)$ may necessitate solving a non-trivial integral.

For completeness, we provide a stand-alone proof for Theorem B.1.

Theorem B.2 (Kong & Schoenebeck (2018b); Chen et al. (2020))**.** *Let D and T be two datasets that are independent conditional on $\boldsymbol{\theta}$, i.e.,*

$$p(D, T|\boldsymbol{\theta}) = p(D|\boldsymbol{\theta})p(T|\boldsymbol{\theta}),$$

then the valuation function

$$U(d, t) = \log \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D = d)p(\boldsymbol{\theta}|T = t)/p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

is truthful.

Proof. This is basically because when D and T are conditionally independent, we have

$$\begin{aligned}
U(d', t) &= \log \int_{\boldsymbol{\theta}} \frac{p(\boldsymbol{\theta}|D = d')p(\boldsymbol{\theta}|T = t)}{p(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\
&= \log \int_{\boldsymbol{\theta}} \frac{p(d'|\boldsymbol{\theta})p(t|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(d')p(t)}\, d\boldsymbol{\theta} \\
&= \log \frac{\int_{\boldsymbol{\theta}} p(d', t, \boldsymbol{\theta})\, d\boldsymbol{\theta}}{p(d')p(t)} \\
&= \log \frac{p(d', t)}{p(d')p(t)} \\
&= \log \frac{p(t|D = d')}{p(t)} \\
&= \log p(t|D = d') - \log p(t),
\end{aligned}$$

which is just the log scoring rule. If the data provider manipulates the dataset and report $f(d) = d' \neq d$, then we have

$$
\begin{aligned}
&\mathbb{E}_T[U(d,T)|D=d] - \mathbb{E}_T[U(d',T)|D=d] \\
=& \sum_{t\in\mathcal{T}} p(t|D=d)\log p(t|D=d) - \sum_{t\in\mathcal{T}} p(t|D=d)\log p(t|D=d') \\
=& \sum_{t\in\mathcal{T}} p(t|D=d)\log \frac{p(t|D=d)}{p(t|D=d')} \\
=& D_{KL}\big(p(t|D=d), p(t|D=d')\big) \\
\geq& \, 0.
\end{aligned}
$$

□

Chen et al. (2020) proposed a theoretical framework for computing this integral score for exponential family distributions.

Definition B.3 (Exponential family Murphy (2012))**.** A likehihood function $p(\mathbf{x}|\boldsymbol{\theta})$, for $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^m$ is said to be in the *exponential family* in canonical form if it is of the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp\left[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})\right] \quad \text{or} \quad p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp\left[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})\right] \tag{2}$$

Here $\boldsymbol{\phi}(x) \in \mathbb{R}^m$ is called a vector of *sufficient statistics*, $Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^n} h(\mathbf{x}) \exp\left[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})\right]$ is called the *partition function*, $A(\boldsymbol{\theta}) = \ln Z(\boldsymbol{\theta})$ is called the *log partition function*.

If the posterior distributions $p(\boldsymbol{\theta}|\mathbf{x})$ are in the same probability distribution family as the prior probability distribution $p(\boldsymbol{\theta})$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior.

Definition B.4 (Conjugate prior for the exponential family Murphy (2012))**.** For a likelihood function in the exponential family $p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp\left[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})\right]$. The conjugate prior for $\boldsymbol{\theta}$ with parameters $\nu_0, \overline{\boldsymbol{\tau}}_0$ is of the form

$$p(\boldsymbol{\theta}) = \mathcal{P}(\boldsymbol{\theta}|\nu_0, \overline{\boldsymbol{\tau}}_0) = g(\nu_0, \overline{\boldsymbol{\tau}}_0) \exp\left[\nu_0 \boldsymbol{\theta}^T \overline{\boldsymbol{\tau}}_0 - \nu_0 A(\boldsymbol{\theta})\right]. \tag{3}$$

Let $\overline{\boldsymbol{s}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}(x_i)$. Then the posterior of $\boldsymbol{\theta}$ can be represented in the same form as the prior

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto \exp\left[\boldsymbol{\theta}^T(\nu_0\overline{\boldsymbol{\tau}}_0 + n\overline{\boldsymbol{s}}) - (\nu_0 + n)A(\boldsymbol{\theta})\right] = \mathcal{P}\big(\boldsymbol{\theta}|\nu_0 + n, \frac{\nu_0\overline{\boldsymbol{\tau}}_0 + n\overline{\boldsymbol{s}}}{\nu_0 + n}\big),$$

where $\mathcal{P}\big(\boldsymbol{\theta}|\nu_0 + n, \frac{\nu_0\overline{\boldsymbol{\tau}}_0 + n\overline{\boldsymbol{s}}}{\nu_0 + n}\big)$ is the conjugate prior with parameters $\nu_0 + n$ and $\frac{\nu_0\overline{\boldsymbol{\tau}}_0 + n\overline{\boldsymbol{s}}}{\nu_0 + n}$.

Then if the prior and the posteriors are in an exponential family, the integral PMI score can be expressed as follows using the normalization function $g(\cdot)$.

Lemma B.5. *If the model distributions are in an exponential family, so that the prior and all the posterior of $\boldsymbol{\theta}$ can be written in the form*

$$p(\boldsymbol{\theta}) = \mathcal{P}(\boldsymbol{\theta}|\nu_0, \overline{\boldsymbol{\tau}}_0) = g(\nu_0, \overline{\boldsymbol{\tau}}_0) \exp\left[\nu_0 \boldsymbol{\theta}^T \overline{\boldsymbol{\tau}}_0 - \nu_0 A(\boldsymbol{\theta})\right],$$

$p(\boldsymbol{\theta}|D) = \mathcal{P}(\boldsymbol{\theta}|\nu_D, \overline{\boldsymbol{\tau}}_D)$ *and* $p(\boldsymbol{\theta}|T) = \mathcal{P}(\boldsymbol{\theta}|\nu_T, \overline{\boldsymbol{\tau}}_T)$*, then the pointwise mutual information can be expressed as*

$$PMI(D,T) = \frac{g(\nu_D, \overline{\boldsymbol{\tau}}_D) g(\nu_T, \overline{\boldsymbol{\tau}}_T)}{g(\nu_0, \overline{\boldsymbol{\tau}}_0) g(\nu_D + \nu_T - \nu_0, \frac{\nu_D \overline{\boldsymbol{\tau}}_D + \nu_T \overline{\boldsymbol{\tau}}_T - \nu_0 \overline{\boldsymbol{\tau}}_0}{\nu_D + \nu_T - \nu_0})}.$$

However, finding the function $g(\cdot)$ is not straightforward and may involve solving a complex integral.

C Missing proofs in Section 3

C.1 Proof of Proposition 3.2

Let D and T be two datasets induced by the data generating process described in Section 2, and let $f(D)$ be any strategic data curation method so that $\boldsymbol{\theta} \to D \to f(D)$ forms a Markov chain. We want to show that the Shannon mutual information $I(f(D), T)$, if computable, is a desirable scoring function, in other words, $I(f(D), T) \leq I(D, T)$. Due to Lemma 3.1, it suffices to prove that $T \to D \to f(D)$ forms a Markov chain.

Since $\boldsymbol{\theta} \to D \to f(D)$ forms a Markov chain, which means that $\boldsymbol{\theta}$ and $f(D)$ are independent conditioned on D, and D and T are independent conditioned on $\boldsymbol{\theta}$ by the data generating process, it follows that T and $f(D)$ are independent conditioned on D,

$$
\begin{aligned}
& p(T, f(D)|D) \\
=& \int_{\boldsymbol{\theta}} p(T, f(D), \boldsymbol{\theta}|D)\, d\boldsymbol{\theta} \\
=& \int_{\boldsymbol{\theta}} p(T, f(D)|\boldsymbol{\theta}, D) p(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} \\
=& \int_{\boldsymbol{\theta}} p(T|f(D), \boldsymbol{\theta}, D) p(f(D)|\boldsymbol{\theta}, D) p(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} \\
=& \int_{\boldsymbol{\theta}} p(T|\boldsymbol{\theta}) p(f(D)|D) p(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} \\
=& p(f(D)|D) \int_{\boldsymbol{\theta}} p(T|\boldsymbol{\theta}) p(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} \\
=& p(f(D)|D) p(T|D).
\end{aligned}
$$

Therefore $T \to D \to f(D)$ forms a Markov chain as well, and by Lemma 3.1, the Shannon mutual information of the curated dataset and the test dataset $I(f(D), T)$ will be a desirable scoring function if computable.

C.2 Proof of Theorem 3.4

To prove the theorem, we first prove the following lemma.

Lemma C.1. *Let D and T be two random variables that are independent conditional on random variable $\boldsymbol{\theta}$, that is, $p(D, T|\boldsymbol{\theta}) = p(D|\boldsymbol{\theta})p(T|\boldsymbol{\theta})$. Then we have for any $\eta \in \Theta$, $d \in \mathcal{D}$, and $t \in \mathcal{T}$,*

$$
\frac{p(T = t|D = d)}{p(T = t)} = \frac{p(\boldsymbol{\theta} = \eta|D = d) \cdot p(\boldsymbol{\theta} = \eta|T = t)}{p(\boldsymbol{\theta} = \eta) \cdot p(\boldsymbol{\theta} = \eta|D = d, T = t)}.
$$

The proof of Lemma C.1 mainly relies on Bayes' rule and the conditional independence condition.

Proof. Since D, T are independent conditional on $\boldsymbol{\theta}$, for any $\eta \in \Theta$ we have

$$
\begin{aligned}
& p(\boldsymbol{\theta} = \eta|D = d, T = t) \\
=& \frac{p(D = d, T = t|\boldsymbol{\theta} = \eta) \cdot p(\boldsymbol{\theta} = \eta)}{p(D = d, T = t)} \\
=& \frac{p(D = d|\boldsymbol{\theta} = \eta) \cdot p(T = t|\boldsymbol{\theta} = \eta) \cdot p(\boldsymbol{\theta} = \eta)}{p(D = d, T = t)} \\
=& \frac{p(\boldsymbol{\theta} = \eta|D = d) \cdot p(\boldsymbol{\theta} = \eta|T = t) \cdot p(D = d) \cdot p(T = t)}{p(\boldsymbol{\theta} = \eta) \cdot p(D = d, T = t)}.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\frac{p(\boldsymbol{\theta} = \eta|D = d) \cdot p(\boldsymbol{\theta} = \eta|T = t)}{p(\boldsymbol{\theta} = \eta) \cdot p(\boldsymbol{\theta} = \eta|D = d, T = t)} &= \frac{p(D = d, T = t)}{p(D = d) \cdot p(T = t)} \\
&= \frac{p(T = t|D = d)}{p(T = t)}.
\end{aligned}
$$

□

With this equation, we can apply the logarithmic scoring rule to get a truthful valuation function, which gives the valuation function in Theorem 3.4. The proof is as follows.

Proof. According to Lemma C.1, $U(d,t) = \log p(T = t|D = d)/P(T = t)$. Then the expected score is maximized by reporting d because

$$\begin{aligned}
&\mathbb{E}_T[U_\eta(d,T)|D = d] - \mathbb{E}_T[U_\eta(d',T)|D = d] \\
&= \int_t p(t|D = d)\log p(t|D = d)\,dt - \int_t p(t|D = d)\log p(t|D = d')\,dt \\
&= \int_t p(t|D = d)\log \frac{p(t|D = d)}{p(t|D = d')}\,dt \\
&= D_{KL}\big(p(t|D = d), p(t|D = d')\big) \\
&\geq 0.
\end{aligned}$$

And when truthful reporting, the expected score $\mathbb{E}[U_\eta(D,T)]$ is just the Shannon mutual information $I(D,T) = \mathbb{E}_{D,T}\left[\log \frac{p(D,T)}{p(D)p(T)}\right]$. □

C.3 Proof of Corollary 3.5

When the posteriors are in an exponential family and the datasets have bounded sufficient statistics, the PMI will be bounded such that $U_\eta(\widehat{D}_i, T_i) \leq M$. Then the concentration bound can be easily derived using the Chernoff bound. The expected square error is just the variance of the estimator since the estimator is unbiased, which decrease as $O(1/k)$.

C.4 Logistic regression with Gaussian approximation

Consider logistic regression with likelihood function $p(y|\mathbf{x},\boldsymbol{\theta}) = \mathrm{Ber}(y|\mathrm{Sigm}(\boldsymbol{\theta}^T\mathbf{x}))$, and consider Bayesian logistic regression with Gaussian approximation (see Murphy (2012) Chapter 8) where a Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}(0,\Sigma_0)$ is assumed. Then given a dataset $(\mathbf{X},\mathbf{y})$, (where matrix $\mathbf{X}$ is the input data with each column being a data feature and vector $\mathbf{y}$ is the observed labels,) the approximate posterior is given by $p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) \approx \mathcal{N}(\mu,\Sigma)$ with

$$\mu = \arg\min_{\mathbf{w}} E(\mathbf{w}), \quad \Sigma^{-1} = \nabla^2 E(\mathbf{w})|\mu,$$

where $E(\mathbf{w}) = -(\log p(\mathbf{y}|\mathbf{X},\mathbf{w}) + \log p(\mathbf{w}))$. Then μ can be solved by gradient descent and the Hessian matrix Σ^{-1} can be computed in closed form. In particular, if we pick $\Sigma_0 = \mathbf{I}$, then we have $\Sigma^{-1} = \mathbf{X}^T\mathbf{S}\mathbf{X} + \mathbf{I}$, where $\mathbf{S} = \mathrm{diag}\big(\mathrm{Sigm}(\mu^T\mathbf{x}_i)(1 - \mathrm{Sigm}(\mu^T\mathbf{x}_i))\big)$. Therefore as long as the data collector knows the prior $\mathcal{N}(0,\Sigma_0)$, she will be able to compute the posterior given any dataset, and thus our PMI score can be computed by Corollary C.2. Again, we do not need to assume the distribution of the feature $p(\mathbf{x}|\boldsymbol{\theta})$ and our PMI score can be used when the test data and the evaluated data have different feature distributions.

Gaussian models. We provide the closed-form solution for the widely-used Gaussian models below. Consider a Gaussian model with a normal prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mu_0,\Sigma_0)$ and normally distributed posteriors $p(\boldsymbol{\theta}|D = d) = \mathcal{N}(\mu_a,\Sigma_a),\ p(\boldsymbol{\theta}|T = t) = \mathcal{N}(\mu_b,\Sigma_b),\ p(\boldsymbol{\theta}|D = d, T = t) = \mathcal{N}(\mu_{ab},\Sigma_{ab})$. We demonstrate that to compute our PMI score, it is sufficient to evaluate just two posteriors: $p(\boldsymbol{\theta}|D = d) = \mathcal{N}(\mu_a,\Sigma_a)$ and $p(\boldsymbol{\theta}|T = t) = \mathcal{N}(\mu_b,\Sigma_b)$. The parameters of the joint posterior μ_{ab},Σ_{ab} can be derived from $\mu_a,\Sigma_a,\mu_b,\Sigma_b$. Consequently, even if data providers are unable to share the entire dataset due to privacy concerns, the PMI score can still be computed as long as the data provider submits μ_a and Σ_a.

Corollary C.2. *Suppose we have $p(\boldsymbol{\theta}|D = d) = \mathcal{N}(\mu_a,\Sigma_a)$, $p(\boldsymbol{\theta}|T = t) = \mathcal{N}(\mu_b,\Sigma_b)$, and the prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mu_0,\Sigma_0)$, then our PMI score equals*

$$U_\eta(d,t) = \frac{1}{2}\left(\log\frac{\det(\Sigma_0)\det(\widetilde{\Sigma})}{\det(\Sigma_a)\det(\Sigma_b)} + \mu_0^T\Sigma_0^{-1}\mu_0 + \widetilde{\mu}^T\widetilde{\Sigma}^{-1}\widetilde{\mu} - \mu_a^T\Sigma_a^{-1}\mu_a - \mu_b^T\Sigma_b^{-1}\mu_b\right),$$

where $\widetilde{\Sigma} = (\Sigma_a^{-1} + \Sigma_b^{-1} - \Sigma_0^{-1})^{-1}$ *and* $\widetilde{\mu} = \widetilde{\Sigma}\left(\Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b - \Sigma_0^{-1}\mu_0\right)$. *In addition, we have* $p(\boldsymbol{\theta}|D = d, T = t) = \mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma})$.

Proof. We consider Gaussian models with posteriors $p(\boldsymbol{\theta}|D = d) = \mathcal{N}(\mu_a, \Sigma_a)$, $p(\boldsymbol{\theta}|T = t) = \mathcal{N}(\mu_b, \Sigma_b)$, $p(\boldsymbol{\theta}|D = d, T = t) = \mathcal{N}(\mu_{ab}, \Sigma_{ab})$, and the prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mu_0, \Sigma_0)$. Then the PMI score with $\eta = 0$ is equal to

$$U_0(d,t) = \frac{1}{2}\left(\log\frac{\det(\Sigma_0)\det(\Sigma_{ab})}{\det(\Sigma_a)\det(\Sigma_b)} + \mu_0^T\Sigma_0^{-1}\mu_0 + \mu_{ab}^T\Sigma_{ab}^{-1}\mu_{ab} - \mu_a^T\Sigma_a^{-1}\mu_a - \mu_b^T\Sigma_b^{-1}\mu_b\right). \tag{4}$$

Then it suffices to prove that $\Sigma_{ab} = (\Sigma_a^{-1} + \Sigma_b^{-1} - \Sigma_0^{-1})^{-1}$ and $\mu_{ab} = \Sigma_{ab}\left(\Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b - \Sigma_0^{-1}\mu_0\right)$. Due to conditional independence and according to the proof of Lemma C.1, we have

$$\begin{aligned} p(\boldsymbol{\theta}|d,t) &\propto \frac{p(\boldsymbol{\theta}|d)p(\boldsymbol{\theta}|t)}{p(\boldsymbol{\theta})} \\ &= \frac{\mathcal{N}(\boldsymbol{\theta};\mu_a,\Sigma_a)\mathcal{N}(\boldsymbol{\theta};\mu_b,\Sigma_b)}{\mathcal{N}(\boldsymbol{\theta};\mu_0,\Sigma_0)} \\ &\propto \exp\Big(-\frac{1}{2}g(\boldsymbol{\theta})\Big) \end{aligned}$$

where

$$g(\boldsymbol{\theta}) := (\boldsymbol{\theta} - \mu_a)^T\Sigma_a^{-1}(\boldsymbol{\theta} - \mu_a) + (\boldsymbol{\theta} - \mu_b)^T\Sigma_b^{-1}(\boldsymbol{\theta} - \mu_b) - (\boldsymbol{\theta} - \mu_0)^T\Sigma_0^{-1}(\boldsymbol{\theta} - \mu_0)$$

Here, $g(\boldsymbol{\theta})$ can be further simplified as $g(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \widetilde{\mu})^T\widetilde{\Sigma}^{-1}(\boldsymbol{\theta} - \widetilde{\mu}) + Z_2$ where

$$\begin{aligned} \widetilde{\Sigma} &= (\Sigma_a^{-1} + \Sigma_b^{-1} - \Sigma_0^{-1})^{-1} \\ \widetilde{\mu} &= \widetilde{\Sigma}\left(\Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b - \Sigma_0^{-1}\mu_0\right) \\ Z_2 &= \mu_a^T\Sigma_a^{-1}\mu_a + \mu_b^T\Sigma_b^{-1}\mu_b - \mu_0^T\Sigma_0^{-1}\mu_0 - \widetilde{\mu}^T\widetilde{\Sigma}^{-1}\widetilde{\mu}. \end{aligned}$$

Then $p(\boldsymbol{\theta}|d,t)$ must be the Gaussian distribution with mean $\widetilde{\mu}$ and covariance matrix $\widetilde{\Sigma}$.

□

D Interpretation of PMI

Our expression in Theorem 3.4 uncovers the relationship between the PMI of two datasets and the predictions they induce about $\boldsymbol{\theta}$. Using this expression, we demonstrate that the PMI of two datasets can be decomposed into the sum of two terms: (1) a term that measures the similarity between the outcomes obtained from two datasets, i.e., $p(\boldsymbol{\theta}|D)$ and $p(\boldsymbol{\theta}|T)$; (2) a term that measures how much D, T boost the confidence of our estimation of $\boldsymbol{\theta}$.

We first present the interpretation for Gaussian models and then extend it to general distributions. When the prior $p(\boldsymbol{\theta})$ is uninformative compared to $p(\boldsymbol{\theta}|d)$ and $p(\boldsymbol{\theta}|t)$, the PMI dataset score for Gaussian models can be represented as the sum of two terms: (1) a term quantifying the similarity between $p(\boldsymbol{\theta}|D)$ and $p(\boldsymbol{\theta}|T)$, characterized by the *dual skew G-Jensen-Shannon divergence* (Nielsen, 2019) between $p(\boldsymbol{\theta}|D)$ and $p(\boldsymbol{\theta}|T)$; (2) a term assessing how much D, T boost the confidence of our estimation of $\boldsymbol{\theta}$, which is equal to how much d and t reduce the (logarithm of the generalized) variance of our belief about $\boldsymbol{\theta}$.

Given two distributions p and q, the dual skew G-Jensen-Shannon divergence between p and q is their total KL divergence to their geometric mean.

Definition D.1 (Dual skew G-Jensen-Shannon divergence (Nielsen, 2019)). The dual skew G-Jensen-Shannon divergence of two distributions p, q for parameter $\alpha \in [0, 1]$ is defined as $JS_*^{G_\alpha}(p\|q) = (1-\alpha)D_{KL}(G_\alpha(p,q)\|p) + \alpha \cdot D_{KL}(G_\alpha(p,q)\|q)$, where $G_\alpha(p,q)$ is the weighted geometric mean of p and q with $G_\alpha(p,q)(x) \propto p(x)^{1-\alpha}q(x)^\alpha$.

Then the PMI dataset score can be expressed as follows.

Theorem D.2. *When the prior $p(\boldsymbol{\theta})$ is uninformative, our PMI dataset score for Gaussian models has*

$$U(d,t) = \frac{1}{2}\log\frac{|\Sigma_0|}{|\widetilde{\Sigma}|} - 2\cdot JS_*^{G_\alpha}(\mathcal{N}(\mu_a,\Sigma_a)\|\mathcal{N}(\mu_b,\Sigma_b)) - k\log 2$$

with $\alpha = 1/2$, where $\mathcal{N}(\mu_a,\Sigma_a) = p(\boldsymbol{\theta}|d)$, $\mathcal{N}(\mu_b,\Sigma_b) = p(\boldsymbol{\theta}|t)$, and $\mathcal{N}(\widetilde{\mu},\widetilde{\Sigma}) = p(\boldsymbol{\theta}|d,t)$.

The negative dual skew G-Jensen-Shannon divergence indicates the similarity between $p(\boldsymbol{\theta}|D)$ and $p(\boldsymbol{\theta}|T)$. Besides the constant term $-k\log 2$, the term $\frac{1}{2}\log|\Sigma_0|/|\widetilde{\Sigma}| = \frac{1}{2}(\log|\Sigma_0| - \log|\widetilde{\Sigma}|)$ corresponds to the difference in (the logarithm of) the generalized variances of $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|d,t)$, as the determinant of the covariance matrix is the generalized variance of a Gaussian distribution. In other words, it could be interpreted as how much d and t reduce the uncertainty or increase the confidence of our estimation. Therefore $\frac{1}{2}\log|\Sigma_0|/|\widetilde{\Sigma}|$ can be interpreted as how much datasets d and t reduce uncertainty and increase confidence in our estimation.

For general distributions, if we similarly define $D_{\text{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|d)) + D_{\text{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|t))$ as the divergence and $D_{\text{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}))$ as the confidence increase, the approximation holds at equality. See Appendix D.3 for the proof and the details. In addition, this KL divergence representation can be interpreted as the "mutual information" of d and t regarding $\boldsymbol{\theta}$. Due to space constraints, we discuss this interpretation in Appendix D.1.

D.1 Interpretation by pointwise mutual parameter information

Firstly, our score can be represented as d and t's mutual information regarding $\boldsymbol{\theta}$, where the amount of information regarding $\boldsymbol{\theta}$ in a dataset is measured by how much the dataset decreases the KL divergence defined below.

Definition D.3 (Pointwise parameter information of datasets)**.** Given two datasets d, t, and a prior $p(\boldsymbol{\theta})$, define the *pointwise parameter information* of a dataset s as

$$PI_{d,t}(s) = D_{\text{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta})) - D_{\text{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|s)),$$

which represents how much observing s reduces the KL divergence to $p(\boldsymbol{\theta}|d,t)$ from our belief about $\boldsymbol{\theta}$. Similarly, we define the *conditional pointwise parameter information* of a dataset s given another dataset r as

$$PI_{d,t}(s|r) = D_{\text{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|r)) - D_{\text{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|s,r)),$$

which represents how much observing s reduces the KL divergence to $p(\boldsymbol{\theta}|d,t)$ if we have already observed r.

Then our score can be represented as "mutual information" similar to the Shannon mutual information $I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ with the entropy $H(\cdot)$ replaced by our pointwise parameter information.

Theorem D.4. *Our PMI score equals*

$$U_\eta(d,t) = PI_{d,t}(d) + PI_{d,t}(t) - PI_{d,t}(d\cup t) \triangleq PMI_{d,t}(d,t),$$

which we define as the pointwise mutual parameter information of d and t. In addition, we have

$$PMI_{d,t}(d,t) = PI_{d,t}(d) - PI_{d,t}(d|t) = PI_{d,t}(t) - PI_{d,t}(t|d).$$

See the proof in Appendix D.2. Theorem D.4 also suggests that our PMI score can be computed by computing/estimating KL divergence between the posteriors.

D.2 Proof of Theorem D.4

We prove the theorem by proving the following lemma.

Lemma D.5. *When D and T are independent conditional on $\boldsymbol{\theta}$, we have*

$$U_\eta(d,t) = D_{KL}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta})) - D_{KL}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|d)) - D_{KL}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|t)).$$

Proof. The right side of the equation equals

$$
\begin{aligned}
&D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta})) - D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|d)) - D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|t))\\
&= \int p(\boldsymbol{\theta}|d,t)\log\frac{p(\boldsymbol{\theta}|d,t)}{p(\boldsymbol{\theta})}\,d\boldsymbol{\theta} - \int p(\boldsymbol{\theta}|d,t)\log\frac{p(\boldsymbol{\theta}|d,t)}{p(\boldsymbol{\theta}|d)}\,d\boldsymbol{\theta} - \int p(\boldsymbol{\theta}|d,t)\log\frac{p(\boldsymbol{\theta}|d,t)}{p(\boldsymbol{\theta}|t)}\,d\boldsymbol{\theta}\\
&= \int p(\boldsymbol{\theta}|d,t)\log\frac{p(\boldsymbol{\theta}|d)p(\boldsymbol{\theta}|t)}{p(\boldsymbol{\theta}|d,t)p(\boldsymbol{\theta})}\,d\boldsymbol{\theta}\\
&= \log\frac{p(t|d)}{p(t)}\\
&= U_\eta(d,t).
\end{aligned}
$$

The third equation is due to Theorem 3.4, that is, we have $\frac{p(\boldsymbol{\theta}|d)p(\boldsymbol{\theta}|t)}{p(\boldsymbol{\theta}|d,t)p(\boldsymbol{\theta})} = \frac{p(t|d)}{p(t)}$ for all $\boldsymbol{\theta}$. □

Then according to our definition of pointwise parameter information, we have

$$
\begin{aligned}
U_\eta(d,t) &= D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta})) - D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|d)) - D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|t))\\
&= \big(D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta})) - D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|d))\big)\\
&\quad + \big(D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta})) - D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|t))\big)\\
&\quad - \big(D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta})) - D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|d,t))\big)\\
&= PI_{d,t}(d) + PI_{d,t}(t) - PI_{d,t}(d\cup t)\\
&\triangleq PMI_{d,t}(d,t).
\end{aligned}
$$

And by our definition of conditional pointwise parameter information, we have

$$
\begin{aligned}
U_\eta(d,t) &= \big(D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta})) - D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|d))\big)\\
&\quad - \big(D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|t)) - D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|d,t))\big)\\
&= PI_{d,t}(d) - PI_{d,t}(d|t).
\end{aligned}
$$

Similarly, we have $U_\eta(d,t) = PI_{d,t}(t) - PI_{d,t}(t|d)$.

D.3 Proof of Theorem D.2

Recall that the dual skew G-Jensen-Shannon divergence is defined as follows.

Definition D.6 (Dual skew G-Jensen-Shannon divergence (Nielsen, 2019))**.** The dual skew G-Jensen-Shannon divergence of two distributions p, q for parameter $\alpha \in [0,1]$ is defined as $JS_*^{G_\alpha}(p\|q) = (1-\alpha)D_{KL}(G_\alpha(p,q)\|p) + \alpha \cdot D_{KL}(G_\alpha(p,q)\|q)$, where $G_\alpha(p,q)$ is the weighted geometric mean of p and q with $G_\alpha(p,q)(x) \propto p(x)^{1-\alpha}q(x)^\alpha$.

Nielsen (2019) solved the dual skew G-Jensen-Shannon divergence JS_*^G between two multivariate Gaussian, which is equal to the following.

Lemma D.7 (Nielsen (2019) Corollary 1)**.** *The dual skew G-Jensen-Shannon divergence $JS_*^{G_\alpha}$ between two multivariate Gaussian $\mathcal{N}(\mu_1,\Sigma_1)$ and $\mathcal{N}(\mu_2,\Sigma_2)$ with $\alpha = \frac{1}{2}$ is equal to*

$$
JS_*^{G_\alpha}(\mathcal{N}(\mu_1,\Sigma_1)\|\mathcal{N}(\mu_2,\Sigma_2)) = \frac{1}{4}\left(\mu_1^T\Sigma_1^{-1}\mu_1 + \mu_2^T\Sigma_2^{-1}\mu_2 - 2\mu^T\Sigma^{-1}\mu + \log\frac{|\Sigma_1||\Sigma_2|}{|\Sigma|^2}\right),
$$

where $\Sigma = 2(\Sigma_1^{-1}+\Sigma_2^{-1})^{-1}$ and $\mu = \frac{1}{2}\Sigma(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2)$.

Then suppose we have $p(\boldsymbol{\theta}|D=d) = \mathcal{N}(\mu_a,\Sigma_a)$, $p(\boldsymbol{\theta}|T=t) = \mathcal{N}(\mu_b,\Sigma_b)$, the prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mu_0,\Sigma_0)$, and $p(\boldsymbol{\theta}|D=d,T=t) = \mathcal{N}(\widetilde{\mu},\widetilde{\Sigma})$ with $\widetilde{\Sigma} = (\Sigma_a^{-1}+\Sigma_b^{-1}-\Sigma_0^{-1})^{-1} \approx (\Sigma_a^{-1}+\Sigma_b^{-1})^{-1}$ and $\widetilde{\mu} = \widetilde{\Sigma}\left(\Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b - \Sigma_0^{-1}\mu_0\right) \approx \widetilde{\Sigma}\left(\Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b\right)$. By definition, we have

$$
JS_*^{G_\alpha}(\mathcal{N}(\mu_a,\Sigma_a)\|\mathcal{N}(\mu_b,\Sigma_b)) = \frac{1}{4}\left(\mu_a^T\Sigma_a^{-1}\mu_a + \mu_b^T\Sigma_b^{-1}\mu_b - 2\mu^T\Sigma^{-1}\mu + \log\frac{|\Sigma_a||\Sigma_b|}{|\Sigma|^2}\right),
$$

where $\Sigma = 2(\Sigma_a^{-1} + \Sigma_b^{-1})^{-1} \approx 2\widetilde{\Sigma}$ and $\mu = \frac{1}{2}\Sigma(\Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b) \approx \widetilde{\mu}$. Then $U_\eta(d,t)$ defined in Corollary C.2 has

$$\begin{aligned}
U_\eta(d,t) + 2 \cdot JS_*^{G_\alpha}(\mathcal{N}(\mu_a,\Sigma_a)\|\mathcal{N}(\mu_b,\Sigma_b)) \approx& \frac{1}{2}\log\frac{\det(\Sigma_0)\det(\widetilde{\Sigma})}{\det(\Sigma_a)\det(\Sigma_b)} + \frac{1}{2}\log\frac{\det(\Sigma_a)\det(\Sigma_b)}{\det(2\widetilde{\Sigma})^2} \\
=& \frac{1}{2}\log\frac{\det(\Sigma_0)\det(\widetilde{\Sigma})}{\det(2\widetilde{\Sigma})^2} \\
=& \frac{1}{2}\log\frac{\det(\Sigma_0)\det(\widetilde{\Sigma})}{4^k\cdot\det(\widetilde{\Sigma})^2} \\
=& \frac{1}{2}\log\frac{\det(\Sigma_0)}{\det(\widetilde{\Sigma})} - k\log 2.
\end{aligned}$$

For general distributions, we can get a similar interpretation using Lemma D.5. Similar to the definition of the dual skew G-Jensen-Shannon divergence, we define $\frac{1}{2}D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|d)) + \frac{1}{2}D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|t))$ as the divergence of $p(\boldsymbol{\theta}|d)$ and $p(\boldsymbol{\theta}|t)$, where $p(\boldsymbol{\theta}|d,t)$ is viewed as the geometric mean of $p(\boldsymbol{\theta}|d)$ and $p(\boldsymbol{\theta}|t)$. In addition, we define $D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}))$ as the counterpart of $\frac{1}{2}\log\frac{\det(\Sigma_0)}{\det(\widetilde{\Sigma})} - k\log 2$, representing confidence increase/uncertainty reduction. Then by Lemma D.5, the PMI dataset score $U_\eta(d,t)$ equals the confidence increase $D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}))$ minus the divergence $D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|d)) + D_{\mathrm{KL}}(p(\boldsymbol{\theta}|d,t)\|p(\boldsymbol{\theta}|t))$.

E SIMULATIONS

E.1 DETAILED EXPERIMENT SETUP IN SECTION 4.1

We randomly sample dataset pairs containing images of 0s and 1s from MNIST. First, we randomly select two correlated numbers, $r_D, r_T \in \{0.2, 0.8\}$, such that their mutual information can be computed.

	$r_D = 0.2$	$r_D = 0.8$
$r_T = 0.2$	ρ	$\frac{1}{2} - \rho$
$r_T = 0.8$	$\frac{1}{2} - \rho$	ρ

These numbers represent the proportion of data points with the label 0 in datasets D and T, respectively. Next, we generate a random vector $L_D \in \{0,1\}^{N_D}$, where each element is 0 with probability r_D, and a similar vector $L_T \in \{0,1\}^{N_T}$, where each element is 0 with probability r_T. Each value in L_D and L_T corresponds to a label for an image. To simplify analysis while preserving the overall dataset composition, we make a minor modification: we replace the last bit of L_D by $\oplus_{j=1}^{N_D-1} L_D(i) \oplus \mathbf{1}(r_D = 0.2)$ so that the XOR sum of L_D reveals the value of r_D. Similarly, we adjust the last bit of L_T so that the XOR sum of L_T reveals r_T. Finally, we replace each label in L_D and L_T with a randomly selected image matching the label, resulting in two correlated datasets D and T.

Fact 2. *The mutual information of the generated datasets $I(D,T) = I(r_D, r_T)$ increases in ρ for $0.25 \le \rho \le 0.5$.*

Proof. By Theorem 4 in (Gowri et al., 2024), $I(D,T) = I(L_D, L_T)$ assuming that $H(L_D|D) = 0$ and $H(L_T|T) = 0$. Again, since L_D and L_T fully reveals r_D and r_T, which means $H(r_D|L_D) = 0$ and $H(r_T|L_T) = 0$, Theorem 4 in (Gowri et al., 2024) implies that $I(L_D, L_T) = I(r_D, r_T)$. Therefore $I(D,T) = I(L_D, L_T) = I(r_D, r_T)$. □

To estimate the mutual information, we generate k dataset pairs $(D_1, T_1), \ldots, (D_k, T_k)$ and compute the average PMI using our formula Theorem 3.4 or Monte Carlo integration. We repeat the process for m times using different parameters $0.25 \le \rho_1, \ldots, \rho_m \le 0.5$ and estimate the ranking of mutual information (using our method or Monte Carlo integration). The accuracy of the methods is assessed by the rank correlation between the estimated and true rankings.

Setting. We selected ten values of ρ corresponding to mutual information values ranging from 0.1 to 1.0 (in increments of 0.1) for evaluation. For each ρ, we tested the performance of our PMI estimator under different regularization strengths C.

To fit the dataset, we employed the Bayesian logistic regression with Gaussian approximation outlined in Appendix C.4. We use the `LogisticRegression` function from the `sklearn` library. The model utilizes the L_2 norm as the regularization term, with the regularization strength controlled by C. The logistic regression model is configured with a maximum number of iterations set to 5000 (max_iter $= 5000$) and no intercept fitting (fit_intercept $=$ False), while all other parameters are set to their default values. The range of C is tuned via cross-validation.

Then for each ρ, we computed the estimated mutual information using our PMI formula, averaged over 1000 repeated trials, and calculated the Kendall τ rank correlation between the estimated mutual information rankings and the true rankings of ρ.

As a baseline, we used Monte Carlo integration to estimate the mutual information for each ρ. Each Monte Carlo integration involved sampling 1000 points. Similarly, we computed the Kendall τ rank correlation between the rankings derived from the baseline's estimated mutual information and the true ρ rankings.

E.2 Detailed experiment setup in Section 4.2

E.2.1 Colored MNIST

Experimental Settings:

In this study, we evaluate a logistic regression model with varying regularization strengths C on a colorized MNIST dataset under three scenarios: (1) Data Denoising, (2) Data Duplication, and (3) Data Removal. The training and test sets consist of samples from four categories: blue-label-0, blue-label-1, green-label-0, and green-label-1. `blue-label-0` refers to images with a blue background and a label of 0, `blue-label-1` refers to images with a blue background and a label of 1, `green-label-0` refers to images with a green background and a label of 0, and `green-label-1` refers to images with a green background and a label of 1.

The logistic regression model is implemented using the `LogisticRegression` function from the `sklearn` library. It employs the L_2 norm as the regularization term, with the strength of regularization controlled by C. The model is configured with a maximum number of iterations set to 5000 (max_iter $= 5000$) and no intercept fitting (fit_intercept $=$ False), while all other parameters are set to their default values. The range of C is tuned via cross-validation.

In all scenarios, the experiment is repeated 1,000 times for each value of C, and the mean changes in the PMI score and the test accuracy are computed. To compute our PMI scoring function, we employed the Bayesian logistic regression with Gaussian approximation outlined in Appendix C.4. This process is independently repeated 10 times, resulting in 10 groups of mean values (each group based on 1,000 repetitions). From these groups, the overall mean (averaged across all 10,000 experiments) and standard deviation (from the 10 groups) are calculated.

1. Data Denoising: For this scenario, we introduce noise by flipping the labels of 10 training samples prior to model training. After training, the mislabeled samples are corrected, and the model is retrained. Results are presented in Table 2 and Table 4.

2. Data Duplication: In this scenario, additional blue-label-0 and green-label-1 samples are duplicated in the training set to match the ratio of four categories of the test set. The model is retrained, and the changes in PMI Score, Loss, and Accuracy are recorded. Results are presented in Table 2 and Table 4.

3. Data Removal: Here, blue-label-0 and green-label-1 samples are removed from the training set to match the ratio of four categories of the test set. The model is retrained, and the changes in PMI Score, Loss, and Accuracy are recorded. Results are presented in Table 2 and Table 4.

C	Change in PMI Score	Change in Accuracy (%)
Data Denoising		
10	7.8126 ± 0.9157	0.92 ± 0.08
20	7.8239 ± 1.0087	0.74 ± 0.05
50	6.2547 ± 0.9763	1.29 ± 0.09
100	14.5329 ± 1.3924	0.20 ± 0.02
200	11.9261 ± 1.1762	0.09 ± 0.02
Data Duplication		
10	-2.2345 ± 1.1247	4.50 ± 0.74
20	-1.5895 ± 1.0426	2.95 ± 0.71
50	-1.3916 ± 1.0483	3.21 ± 0.68
100	-0.4248 ± 0.2519	1.61 ± 0.54
200	-1.7580 ± 0.8914	0.97 ± 0.34
Data Removal		
10	-7.2140 ± 0.9073	0.41 ± 0.03
20	-7.5783 ± 1.1306	0.19 ± 0.01
50	-6.5111 ± 1.1430	0.21 ± 0.02
100	-7.1336 ± 0.9251	0.20 ± 0.02
200	-14.1899 ± 1.7394	0.67 ± 0.03

Table 4: **Changes** in PMI score function and test accuracy after three data curation methods with different regularization strengths **C** in the Colored MNIST dataset. The original training set, sampled from a larger dataset, consists of samples from four categories, with sizes of 50 or 150 per category, and the test set has sizes of $50, 150, 150, 50$. To introduce noise, a certain percentage of the training labels are flipped. The **Denoising** method removes flipped data points, with a training set size of 50 samples per category and a test set size of 50 samples per category. The **Duplication** method adjusts the training set to match the test set's category ratios via duplication, resulting in sizes of $50, 100, 100, 50$ for the test categories. Finally, the **Removal** method reduces the training set size to match the test set category ratios, resulting in training sizes of $100, 50, 50, 100$. The experiment was repeated $1,000$ times to compute the mean changes in PMI scores and accuracy and repeated 10 times to compute final means and variances.

E.2.2 CORRUPTED CIFAR

Experimental Settings. We set up the following three experiments to compare the performance of our PMI score function against the standard evaluation approach on the corrupted CIFAR dataset in evaluating three different data curation methods. We choose two classes as labels 0 and 1 among all classes in the CIFAR-10 datasets and select two corruption types (`brightness` and `contrast`) as bias in the datasets. More details of corruption design can be found in Hendrycks & Dietterich (2019).

Using the label and bias of data, we sample with different ratios in four categories: `brightness-label-0`, `contrast-label-0`, `brightness-label-1`, `contrast-label-1`. `brightness-label-0` refers to images with `brightness` corruption and a label of 0. `contrast-label-0` refers to images with `contrast` corruption and a label of 0. `brightness-label-1` refers to images with `brightness` corruption and a label of 1. `contrast-label-1` refers to images with `contrast` corruption and a label of 1. We sample training sets with ratio 1:1:1:1 and test sets with ratio 1:2:2:1 or 1:3:3:1 with respect to four categories.

In each experiment, we extract image embeddings using ResNet18 pre-trained on ImageNet (with the last layer removed) and flip 10% labels of the sampled training dataset to introduce noise. Then we train logistic regression models on these embeddings with varying regularization strengths C ranging from 10000 to 100000. To further clarify, the logistic regression model is implemented using the `LogisticRegression` function from the `sklearn` library. It employs the L_2 norm as the regularization term, with the strength of regularization controlled by C. The model is configured with a maximum number of iterations set to 5000 (max_iter $=$ 5000) and no intercept fitting (fit_intercept $=$ False), while all other parameters are set to their default values. Here we add a dimension in embeddings where each entry is 1 and omit the bias term to integrate the bias into the weight vector. The range of C is tuned via cross-validation.

For each value of C, the experiment is repeated 1,000 times, and we compute the mean changes in PMI Score and test accuracy across these 1,000 runs. To compute our PMI we employed the Bayesian logistic regression with Gaussian approximation outlined in Appendix C.4. This process is independently repeated 10 times, producing 10 groups of mean values (each group based on 1,000 repetitions). From these 10 groups, we calculate the overall mean (averaging across all 10,000 experiments) and the standard deviation (calculated from the 10 groups of mean values). The results are summarized in Tables 3 and 5.

1. Data Denoising. In this experiment, we check the change of PMI score function and test accuracy after removing the mislabeled data. We directly remove the data points whose labels are flipped.

2. Data Duplication. In this experiment, we check the change of PMI score function and test accuracy after duplicating part of training dataset to match the ratio of four categories of test dataset which is a non-essential feature irrelevant to the model.

3. Data Removal. In this experiment, we check the change of PMI score function and test accuracy after removing part of training dataset to match the ratio of four categories of test dataset which is a non-essential feature irrelevant to the model.

C	Change in PMI Score	Change in Accuracy (%)
	Data Denoising	
10000	2.1175±0.1916	7.38±0.11
20000	1.9854±0.1680	7.29±0.14
30000	1.9894±0.2097	7.36±0.15
50000	1.8297±0.1332	7.26±0.14
100000	1.5816±0.1717	7.31±0.09
	Data Duplication	
10000	-0.2901±0.0757	0.84±0.07
20000	-0.3753±0.0793	0.86±0.06
30000	-0.5194±0.1152	0.84±0.06
50000	-0.5766±0.0758	0.86±0.06
100000	-0.8050±0.0973	0.86±0.05
	Data Removal	
10000	-6.1685±0.0608	1.83±0.12
20000	-6.8668±0.1173	2.00±0.14
30000	-7.3100±0.1357	1.85±0.12
50000	-7.8402±0.1086	1.86±0.14
100000	-8.3621±0.1293	1.92±0.09

Table 5: **Changes** in PMI score function and test accuracy after three data curation methods with different regularization strengths **C** in Corrpted CIFAR dataset. The original training set, sampled from a larger dataset, consists of images from four categories, each with size 30, and the test set has sizes $20, 60, 60, 20$. To introduce noise, 10% of the training labels are flipped. The **Denoising** method simply removes flipped data points, while **Duplication** and **Removal** adjust the training set to match the test set's category ratios via copy or delete operations, resulting in sizes of $30, 90, 90, 30$ and $10, 30, 30, 10$, respectively. The experiment, repeated 1,000 times to compute mean changes in PMI scores and accuracy and repeated 10 times to obtain final means and variances.