QUANTITATIVE DATA USAGE INFERENCE IN MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

How much of a given dataset was used to train a machine learning model? This is a critical question for data owners assessing the risk of unauthorized data usage and protecting their right (United States Code, 1976). However, previous work mistakenly treats this as a binary problem—inferring whether *all or none* or *any or none* of the data was used—which is fragile when faced with real, non-binary data usage risks. To address this, we propose a fine-grained analysis called Dataset Usage Cardinality Inference (DUCI), which estimates the exact proportion of data used. Our algorithm, leveraging debiased membership guesses, matches the performance of the optimal MLE approach (with a maximum error < 0.1) but with significantly lower (e.g., $300\times$ less) computational cost.

1 INTRODUCTION

The increasing legal conflicts over the unauthorized use of datasets to train artificial intelligence models (e.g., the New York Times' lawsuit against OpenAI (The New York Times Company, 2023)) signal the increasingly critical issues about the boundary between intellectual property infringement and fair use, as well as compliance with data privacy regulations (e.g., GDPR (European Parliament and Council of the European Union, 2016) and CCPA (California Legislature, 2018)). This has driven extensive research into data provenance and dataset ownership verification in machine learning (Li et al., 2023; 2020; Tang et al., 2023; Li et al., 2022b; Sun et al., 2022; Sablayrolles et al., 2020; Hu et al., 2022; Zou et al., 2022; Wenger et al., 2022; Maini et al., 2021; Song & Shmatikov, 2019), primarily focusing on implanting backdoor into target models by adding poisoning data to the protected dataset (Li et al., 2023; 2020; Tang et al., 2023; Li et al., 2022b) or thresholding average statistics over the dataset to conduct hypothesis testing (Maini et al., 2021; Song & Shmatikov, 2019).

However, all these methods are limited to exploring a binary question: whether an entire dataset were used in training a specific model? Such binary methods are observed to be **fragile under partial dataset utilization** (i.e., a practical setting where models are trained on a combination of subsets from multiple data sources). In Figure 1, we plot the original binary prediction given by two representative works (Li et al., 2023; Maini et al., 2021) and the continuous scores we retrieved from their methods (normalized to the range from 0 to 1). (See the implementation details in Appendix B.) While both methods achieve perfect prediction when predicting binary utilization (either none or all), their decision fluctuates in cases of partial usage (regardless of the threshold). For example, both methods will neglect small proportion of usage, and might classify a model as ”utilizing” the dataset when 50% is employed, yet deeming a model as ”not utilizing” when usage increases to 60%. These failures could cause severe social problems since the victim can easily bypass the detection by utilizing a small portion of the dataset. Thus, **fine-grained estimation of utilization proportions is necessary for ensuring a reliable and consistent dataset usage prediction.**

On the other hand, practical application of dataset usage detection requires to know the extent of dataset usage. For example, according to the Section 107 of the U.S. Copyright Act (United States Code, 1976), determining whether a use qualifies as fair use or copyright infringement must consider the “amount and substantiality of the portion used in relation to the copyrighted work” and “nature of the copyrighted work” (i.e., different use cases require different thresholds). However, none of the existing methods can be naively extended to predict the proportion of dataset usage due to the following challenges: **(1)** Finding a promising dataset-level score for the task is non-trivial. For example, a continuous score forcefully obtained from backdoor watermarking decisions has significant errors, as backdoor injection is not a linear process. **(2)** Even if a promising dataset-level score could be found, determining thresholds for various usage proportions is still challenging. This is because the distribution of scores across different proportions is unpredictable unless we obtain enough samples from the full distribution of all proportions—requiring the training of hundreds of reference models, a process that is both costly and sensitive (MLE Baselines in Table 1).

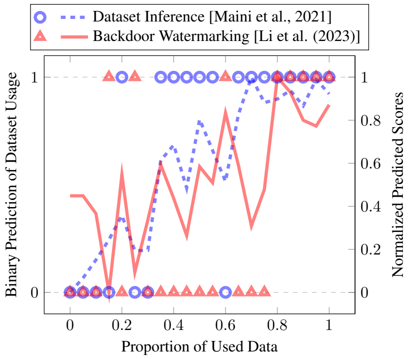

Figure 1: Prior (binary) methods fluctuate between 0 and 1.

All these motivate us to ask the question: ***Can we reliably and cost-effectively infer the proportion of the target dataset used in a model?*** Specifically, our goal is to identify a suitable continuous score and inferring dataset usage *without any modifications* to the data itself. This later constraint is critical for broadening the applicability of our methods to scenarios where *all* prior *binary* watermarking approaches (Guo et al., 2023; Li et al., 2023; Tang et al., 2023; Hu et al., 2022; Li et al., 2022b; Wenger et al., 2022; Sablayrolles et al., 2020) fall short: 1) scenarios where altering high-quality data is undesirable or where data integrity must be preserved (e.g., in healthcare applications); 2) situations involving previously released datasets that have likely already been incorporated in trained models.

Our contributions We identified the fragility of previous binary methods in dataset usage inference (Section 1). Thus, we introduce a new problem of Dataset Usage Cardinality Inference (DUCI) in machine learning, via a paradigm shift from binary formulations to a granular assessment of dataset utilization, and propose an efficient and powerful algorithm to precisely estimate the proportion of a dataset used in model training. Specifically,

1. We formally define the problem of DUCI and discuss the additional challenges in designing the dataset-level score (Section 3.2).
2. To construct a promising dataset-level score of DUCI task, we developed a novel debiasing procedure that is **applicable to all membership identification techniques** (equation 6).
3. We design an unbiased estimator for data usage by debiasing individual membership guesses (Section 4) and quantify uncertainty in DUCI by constructing **accurate** confidence intervals and validate them through extensive experiments (Figure 3).
4. Our experiments confirm the accuracy and cost-effectiveness of our method (Section 5.3): it provides reliable proportion estimation **maximum error** < 0.1 with as few as **one** reference model (Figure 2), and requires $630\times$ less computational budget than idealized methods while maintaining comparable performance (Table 1).
5. Our debiasing method can be extended to group-level(Table 2), and effectively address the challenge of practical quality-dependent dataset sampling schemes.
6. Our methods can also be applied to language models for practical tasks such as quantifying book copyright infringement (Table 4).

2 PRELIMINARIES

Notation We use $\mathcal{T}$ to denote the training algorithm that, in addition to the model architecture, may have access to auxiliary public information, e.g., a population data pool ρ. We consider the standard supervised learning setting, where a trained model $\theta : \chi \to \mathbb{R}^n$ maps an input $x \in \chi$ to an n-class prediction vector $\theta(x) \in [0, 1]^n$, where $\|\theta(x)\|_1 = 1$. We denote $\theta(x)_y$ as the predicted probability for the groundtruth class y where $y \in [n]$. (e.g., in text generation tasks, n is the vocabulary size.) Membership m of a data point x with regard to the training set D of a model θ is $m = 1$ if $x \in D$ (a member), and $m = 0$ if $x \notin D$ (a nonmember). We define the term *membership guess* $\hat{m}$ as the predicted membership status of a data point x by any algorithm $\mathcal{A}$ (denoted as membership identification algorithms) capable of inferring membership.

2.1 MEMBERSHIP INFERENCE BACKGROUND

Membership inference (MI), examines whether a particular data point x is part of the training set of a target model θ, serves as a natural choice of providing membership guess. We next introduce the state-of-the-art (SOTA) MI algorithm demonstrating the best performance in our experiments. Results for other SOTA MI algorithms are included in Appendix G.4 for comparison.

Robust Membership Inference Attack (RMIA) (Zarifzadeh et al., 2024) compares the target point x with M population data $z_m \sim \rho$ for $m = 1, \ldots, M$, using the ratio $\frac{P(x|\theta)}{P(x)}$ as signal. Here, $P(x|\theta) = \theta(x)_y$ and $P(x)$ is a normalization constant for x calculated by integrating over all models trained with the same structure and training data distribution (but not the same training dataset) as the target model θ (i.e., reference models):

$$\hat{m} = \text{MIA}(x;\theta) = \mathbb{1}_{[P_{z\sim\rho}(LR_\theta(x,z)\geq 1)\geq\beta(x)]} \text{ for threshold}\beta(x), \tag{1}$$

where $LR_\theta(x,z) = \left(\frac{P(x|\theta)}{P(x)}\right)\cdot\left(\frac{P(z|\theta)}{P(z)}\right)^{-1}$ and $P_{z\sim\rho}(LR_\theta(x,z)\geq 1) = \frac{1}{M}\sum_{m=1}^{M}\mathbb{1}_{[LR_\theta(x,z_m)\geq 1]}$. Additional discussions of related works are deferred to Appendix A.

3 DATASET USAGE CARDINALITY INFERENCE PROBLEM

In this section, we formulate the problem of Dataset Usage Cardinality Inference (DUCI) in machine learning and analyze the challenges of solving this problem.

3.1 DATASET USAGE AS A NON-BINARY INFERENCE PROBLEM

To precisely assess the degree of dataset utilization, we propose moving beyond the binary 'Yes' or 'No' paradigm in prior works. Instead, we employ a continuous metric, proportion $p \in [0, 1]$, to precisely measure the dataset utilization extent.

Problem formulation: Let X be a target dataset and θ a model (trained with an algorithm $\mathcal{T}$) that potentially incorporates data from X and other sources. Each data record x_i are sampled from X following certain unknown probability vector $\boldsymbol{p} = [p_i]_{i=1}^{|X|}$. The Dataset Usage Cardinality Inference (DUCI) algorithm $\mathcal{A}$—acting as an agent for the dataset owner with full access to X—aims to estimate (denoted as $\hat{p}$) the overall proportion $p = \frac{1}{|X|}\mathbb{E}[\sum_{i=1}^{|X|}\mathbf{1}_{\mathrm{x_i\ selected}}]$ of X used in training θ, given black-box access to θ and knowledge of the training algorithm $\mathcal{T}$. The complete pipeline is depicted as follows:

$$\begin{tikzcd}
\text{Dataset } X = [x_i]_{i=1}^{|X|} \arrow[d, "\text{independently include } x_i \text{ with probability } p_i"', "\text{results in } p"] & & X, \mathcal{T} \arrow[d] & \\
\text{TrainingAlgorithm } \mathcal{T} \arrow[r] & \theta \arrow[r, "\text{black-box}"] & \mathcal{A} \arrow[r] & \hat{p}
\end{tikzcd} \tag{2}$$

Objective We aim to design a DUCI algorithm $\mathcal{A}$ that achieves minimal (mean absolute) error across all possible data usage proportion, i.e.,

$$\min_{\mathcal{A}} \max_{p\in[0,1]} \mathbb{E}\left[|p - \hat{p}|\right], \tag{3}$$

where p denotes the ground-truth proportion, $\hat{p}$ denotes the estimated proportion, and the expectation is over the random trials of the dataset usage cardinality inference pipeline (Equation (2)).

3.2 CHALLENGES IN DATASET USAGE CARDINALITY INFERENCE

A natural strategy for inferring dataset usage is to sum the membership inference guesses over individual training data. However, we observe that this method suffers from poor performance (Table 1) due to inherent and systematic errors in identifying individual data point memberships.

Errors in optimal point-wise membership inference: Membership inference at the level of individual training data is known to suffer from high error when there are inherent randomness in the training pipelines (Ye et al., 2022) (such as data sampling and randomness of the training algorithm). When the training algorithm satisfy certain constraints in terms of output indistinguishability, one can prove that even *the theoretically most informed and optimal membership estimate* ($\hat{m}_i$) suffer from an error that increases with the level of indistinguishability (Kairouz et al., 2015; Steinke et al., 2024).

In practice, existing membership identification methods not only contend with these inherent errors but also encounter systematic errors due to the inability to capture the precise membership information from model outputs. For example, state-of-the-art MIAs (Carlini et al., 2022; Bertran et al., 2024; Zarifzadeh et al., 2024) on a standard Wide ResNet model trained on CIFAR-10, despite employing extensive resources such as 254 reference models, reveal inaccurate predictions, e.g., an overall AUC of approximately 72% with less than 2% True Positive Rate in areas with low False Positive Rates.

Challenge: biases arising from the accumulation of errors: The aforementioned per-point membership identification errors, particularly systematic errors, accumulate and introduce biases in solving the DUCI task. This bias is clearly illustrated in Table 1. In the next section, we will discuss strategies to address this challenge by mitigating these biases, and show that the debiased aggregated membership guesses can serve as a promising score for proportion estimation.

4 UNBIASED DATASET USAGE INFERENCE FROM AGGREGATION OF INDIVIDUAL GUESSES

Given the unavoidable errors in membership identification methods, how can we achieve a reliable (e.g., unbiased) estimation of the proportion?

Denote X as the dataset of interest, and $\mathcal{M}$ as the distribution of the ground-truth membership vector. By Section 3.1, i.e., the probability of each data record x_i being used is p_i, we have $\mathcal{M} = [\text{Bernoulli}(p_i)]_{i=1}^{|X|}$. The target model θ is then trained by randomly drawing a membership vector $\mathbf{m} = (m_1, \cdots, m_{|X|}) \sim \mathcal{M}$, and including the i-th data record of X (denoted as x_i) if $m_i = 1$ for each i. Note that since $m_i \sim \text{Bernoulli}(p_i)$, $\mathbb{E}[m_i] = p_i$ for any $i = 1, 2, \ldots, |X|$ over the random trials.

The intuition is that if we can identify an unbiased estimator $\hat{p}_i$ of the probability that each data record x_i is used over random trials, i.e., $\mathbb{E}[\hat{p}_i] = \mathbb{E}[m_i]$ for $i = 1, \ldots, |X|$, then we can construct an unbiased estimator $\hat{p}$ of the true proportion p by aggregation:

$$\hat{p} = \frac{1}{|X|} \sum_{i=1}^{|X|} \hat{p}_i. \tag{4}$$

This is by the additivity of expectation, i.e., $p = \frac{1}{|X|}\mathbb{E}[\sum_{i=1}^{|X|} m_i] = \frac{1}{|X|}\sum_{i=1}^{|X|} \mathbb{E}[\hat{p}_i]$. We next introduce how we identify the unbiased estimator $\hat{p}_i$ of $\mathbb{E}[m_i]$.

Construct $\hat{p}_i$ by debiasing individual MIA guesses Note that for any algorithm capable of providing membership guesses $\hat{m}_i$ for data record x_i (with ground-truth memberships m_i, where $m_i, \hat{m}_i \in \{0, 1\}$), the rule of total probability ensures the following relationship:

$$\mathbb{E}[\hat{m}_i] = P(\hat{m}_i = 1|m_i = 0) \cdot P(m_i = 0) + P(\hat{m}_i = 1|m_i = 1) \cdot P(m_i = 1) \tag{5}$$

where $P(\hat{m}_i = 1|m_i = 0)$ and $P(\hat{m}_i = 1|m_i = 1)$ are the true False Positive Rate (FPR) and True Positive Rate (TPR) for inferring the membership of each record in dataset X (constants across trials). Note that $P(m_i = 1) = 1 - P(m_i = 0) = \mathbb{E}[m_i]$. Rearranging the terms in Equation (5) yields:

$$\hat{p}_i = \frac{\hat{m}_i - P(\hat{m}_i = 1|m_i = 0)}{P(\hat{m}_i = 1|m_i = 1) - P(\hat{m}_i = 1|m_i = 0)}. \tag{6}$$

We assume TPR $\neq$ FPR, which is reasonable given that membership identification algorithms are designed to discriminate between member and non-member data points, not to act randomly. In our experiment, we use the state-of-the-art MIA methods to compute the MIA guesses $\hat{m}_i = \text{MIA}(x_i; \theta)$ (as introduced in Section 2), but note that our approach is *flexible and applicable to any membership prediction technique*. We next discuss how to perform the debiasing operation to compute equation 6 and use it for estimating dataset usage proportion p in Equation (4).

Debiasing method To compute the estimators in Equation (6), we require the probabilities $P(\hat{m}_i = 1|m_i = 0)$ (FPR$_i$) and $P(\hat{m}_i = 1|m_i = 1)$ (TPR$_i$). A direct, but computationally expensive, approach is to estimate these probabilities using N empirically trained reference models, where the training datasets are known. However, this method incurs a sampling error of order $\Omega\left(\frac{1}{\sqrt{N}}\right)$ (i.e., the approximation error of TPR_i and FPR_i can be notable when N is small). To reduce error and computational cost, we propose estimating FPR and TPR across the entire dataset X, which requires as few as a **single** reference model and can **reuse** models from the MIA process. This simplification is justified because the proportion p is a dataset-level statistic, i.e., $p = \frac{1}{|X|}\sum_{i=1}^{|X|} p_i$.[1] Denote $\theta_1, \ldots, \theta_N$ as N reference models (where N can be 1), each trained on randomly sampled halves of the dataset X, we can compute:

$$P(\hat{m}=1|m=0) = 2\sum_{j=1}^{N}\sum_{i=1}^{|X|}\frac{\mathbf{1}_{MIA(\theta_j,x_i)=1 \text{ and } x_i\notin\theta_j}}{N\cdot|X|} \tag{7}$$

$$P(\hat{m}=1|m=1) = 2\sum_{j=1}^{N}\sum_{i=1}^{|X|}\frac{\mathbf{1}_{MIA(\theta_j,x_i)=1 \text{ and } x_i\in\theta_j}}{N\cdot|X|} \tag{8}$$

The confidence interval analysis below reveals that this empirical average is only subject to $O(\frac{1}{\sqrt{N\cdot|X|}})$ sampling error (i.e., the standard deviation of the empirical estimation of $P(\hat{m} = 1|m = 0)$ and $P(\hat{m} = 1|m = 1)$ over the randomness of θ_j and MIA algorithm). See Appendix F for the proofs. Thus, employing larger datasets and more reference models enhances the accuracy of our debiasing method.

Confidence interval and uncertainty estimates for Dataset Usage Cardinality Inference Modern machine learning algorithms and their data sampling processes are highly randomized. To capture the intrinsic uncertainty and confidence in data usage inference, we resort to the Lyapunov Central Limit Theorem (CLT) (Billingsley, 2017) to compute confidence intervals for our estimation method via aggregate statistics. Specifically, if the individual statistics $\hat{p}_1, \cdots, \hat{p}_{|X|}$ in Equation (6) are independent[2] and the Lyapunov condition is satisfied (see the proofs in Appendix E.2), then their aggregation converge to a Gaussian distribution as $|X|$ increases. By applying Lyapunov CLT to Equation (4) and using the sample variance to estimate the average variance (see details in Appendix E), we obtain the following approximate 95% confidence interval for estimating p.

$$\hat{p} - t_{\alpha/2}\sqrt{\frac{s^2}{|X|}} \le p \le \hat{p} + t_{\alpha/2}\sqrt{\frac{s^2}{|X|}} \tag{9}$$

Here $\hat{p} = \frac{1}{|X|}\sum_{i=1}^{|X|}\hat{p}_i$, $\alpha = 0.05$, $t_{\alpha/2}$ is the t critical value, and $s^2 = \frac{1}{|X|-1}\sum_{i=1}^{|X|}(\hat{p}_i - \hat{p})^2$.

5 EXPERIMENTAL SETUP AND RESULTS

5.1 EVALUATION SETTINGS

We considered the general uniformly random sampling setting and special data selection setting, while ensuring that p is controllable over the random trials (as described in equation 3). The comparison experiments in Section 5.3 are conducted under the uniformly random sampling, and Section 5.4 extends the discussion to the special sampling scenarios. For uniformly random sampling, we examine all methods under dataset usage proportion $p \in [0, 1]$ with a granularity of 5%, i.e., $p = 0.00, 0.05, \cdots, 1.00$. For each p, we evaluate the (mean absolute) error $\mathbb{E}[|p - \hat{p}|]$ of proportion estimation over 30 random trials. In each evaluation trial, we train the target model θ on freshly sampled random p proportion of the protected dataset X, combined with randomly sampled remaining data from a population pool (that does not overlap with X). For specialized sampling, we constructed the data selection probability vector $\mathbf{p}$ for each ground-truth p using data selection strategies (Paul et al., 2021) and sample data from X for model training accordingly. We then compute the mean absolute error (MAE) in the same way.

---

[1] To clarify further, under dataset-level estimates $\text{TPR} = \frac{1}{|X|}\sum_i \text{TPR}_i$ and $\text{FPR} = \frac{1}{|X|}\sum_i \text{FPR}_i$, we have $\frac{1}{|X|}\sum_i \mathbb{E}[\hat{p}_i] = p + \frac{\text{Corr}_i(\text{TPR}_i - \text{FPR}_i, p_i)}{\text{TPR}-\text{FPR}}$ (See derivation in Appendix D). This suggests that for many practical samplings (e.g., i.i.d. sampling), this simplification yield an unbiased estimator for p. For special sampling methods, we can debias subgroups to achieve the unbiasedness (as discussed in Table 2).

[2] We experimentally validate that $\hat{p}_1, \cdots, \hat{p}_{|X|}$ are weakly independent (Figure 4) and use > 30 individual observations to compute valid confidence intervals via Lyapunov CLT.

Considering the evaluation cost of training target models on various proportions p, we mainly assess all methods in Section 5.3 across three model architectures: a standard 5-layer fully connected network (FC-5), a wide ResNet with width 2 (WRN28-2) (Zagoruyko & Komodakis, 2016), and ResNet-34 (He et al., 2016), as well as three benchmark datasets: CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and Tiny-ImageNet (Le & Yang, 2015). We further validated the applicability of our methods to language models through a book copyright infringement case study on GPT-2 (Radford et al., 2019), using the BookMIA dataset (Shi et al., 2023). To implement our unbiased proportion estimator, we use MIA to predict the membership guesses, i.e., $\hat{m}_1, \cdots, \hat{m}_{|X|}$. We report the best-performing MIA (i.e., RMIA in Section 2.1) and defer discussions on other MIAs to Appendix C and Appendix G.4. See Algorithm 1 for a pseudocode of our method and see Appendix G for more implementation details.

5.2 BASELINES

Vanilla Membership Inference Attacks (MIA) baselines To demonstrate the importance of debiasing, we compare our method against two intuitive baselines: direct averaging of binary MIA guesses and real-valued MIA scores outputted by existing MIA methods Zarifzadeh et al. (2024); Carlini et al. (2022). We assign N reference models with each trained on half of dataset X to MIA baselines:

- **MIA Guess:** This involves directly counting MIA membership predictions, where the proportion is calculated as the fraction of member predictions: $\hat{p} = \frac{1}{|X|}\sum_{i=1}^{|X|}\hat{m}_i$. For a fair comparison, we use the same MIA algorithms that were employed in our method.
- **MIA Score:** This takes the average of MIA confidence scores across all data points: $\hat{p} = \frac{1}{|X|}\sum_{i=1}^{|X|}\hat{c}_i$, where $\hat{c}_i$ is the confidence score of RMIA (i.e., probability before thresholding).

Maximum Likelihood Estimate (MLE) baselines When computation cost is not a concern, the below computationally expensive MLE estimators serve as an idealized baseline:

$$\hat{p}_{MLE}(o) = \arg\max_{q} P(o|p = q), \tag{10}$$

where $P(o|p = q)$ is the conditional probability of observations o (generated from the trained model θ) given that q proportion of the records in the target dataset X is used in training. This probability $P(o|p = q)$ can be empirically approximated via training many (N) reference models on freshly sampled random q-fraction of the target dataset X for each of 21 possible proportion value q, employing the same algorithm $\mathcal{T}$ that was used for training the target models (in total $21N$ reference models). Regarding the observation o in the MLE baselines, we consider two choices as follows.

- **Joint logits (Joint-Logit) of all records in target dataset (high-dimensional):** Building on prior research (Carlini et al., 2022) which demonstrated that scaled logits for individual data points can serve as effective observations and empirically follow a Gaussian distribution, we naturally extend this approach to a dataset-wide scale. We concatenate these point-wise logits into high-dimensional joint logits and model their distribution using a multivariate Gaussian.

  $$P(o|p = q) = P(\text{logit}(X;\theta)|\mathcal{N}(\boldsymbol{\mu_q}, \boldsymbol{\Sigma_q})), \tag{11}$$

  where $\text{logit}(X;\theta)$ is the the logits of model θ on all records in X, and μ_q and $\boldsymbol{\Sigma_q}$ are its empirical mean and covariance matrix across N reference models trained with q proportion of X.
- **Averaged logits (Avg-Logit) of all records in the target dataset (one-dimensional):** To reduce the dimensionality of the observation (and thus enable more accurate empirical mean and covariance estimation), we additionally consider the average of the logits over the dataset X, following the approach in Maini et al. (2021). The likelihood is then modeled as one-dimensional Gaussian.

  $$P(o|p = q) = P\left(\frac{1}{|X|}\sum_{x\in X}\text{logit}(x;\theta)\Big|\mathcal{N}(\mu_q, \sigma_q^2)\right),$$

  where μ_q and σ_q are the empirical mean and variance of the averaged logits of all records in X across N reference models trained with q proportion of X.

5.3 MAIN RESULTS

In this section, we compare the performance of our method (Algorithm 1) with the baselines in Section 5.2 in terms of estimation quality and efficiency.

Table 1: Maximum mean absolute error (MAE) ($\max_{p\in[0,1]} \mathbb{E}[|p-\hat{p}|]$) over 21 proportions, each derived from 30 trials, of all methods across various datasets and model architectures within the same computational budget (# reference models). The lowest error in each test is highlighted in bold. Grey rows are the idealized computationally expensive MLE baselines.

#Ref	Methods	CIFAR-10		CIFAR-100		Tiny-ImageNet	
		FC-5	WRN28-2	ResNet34	WRN28-2	ResNet34	WRN28-2
1	MIA Guess	0.3291	0.4819	0.1611	0.1676	0.0793	0.1478
	MIA Score	0.3619	0.4486	0.2950	0.3220	0.2909	0.3040
	Ours	**0.0678**	**0.1836**	**0.0534**	**0.0873**	**0.0879**	**0.0580**
2	MIA Guess	0.2617	0.4801	0.1637	0.1458	0.0859	0.1131
	MIA Score	0.3587	0.3865	0.2687	0.3074	0.2822	0.2952
	Ours	**0.0380**	**0.1787**	**0.0421**	**0.0851**	**0.0723**	**0.0463**
4	MIA Guess	0.2797	0.4350	0.1468	0.1401	0.0746	0.0889
	MIA Score	0.3571	0.3822	0.2571	0.2836	0.2593	0.2739
	Ours	**0.0283**	**0.1634**	**0.0325**	**0.0774**	**0.0332**	**0.0424**
8	MIA Guess	0.2831	0.3862	0.1493	0.119	0.0768	0.1099
	MIA Score	0.3022	0.3621	0.2552	0.2824	0.2588	0.2759
	Ours	**0.0296**	**0.1127**	**0.0180**	**0.0554**	**0.0147**	**0.0341**
42	MLE (Joint-Logit)	0.2450	0.2230	0.1983	0.2667	0.2050	0.2067
	MLE (Avg-Logit)	0.0533	0.1383	0.0567	0.0617	0.0383	0.0583
	MIA Guess	0.2516	0.2889	0.086	0.0797	0.0377	0.0867
	MIA Score	0.2948	0.3692	0.2530	0.2920	0.2576	0.2867
	Ours	**0.0271**	**0.0722**	**0.015**	**0.0339**	**0.0124**	**0.0179**
630	MLE (Joint-Logit)	0.1400	-	-	-	-	-
	MLE (Avg-Logit)	0.0283	-	-	-	-	-

Comparison under the same computation budget We first fix the total computational budgets to be the same for all methods and compare their estimation quality. Table 1 shows an overview of the comparison results under different datasets and model architectures. Note that for our debiasing process, single reference model is enough and we can directly reuse the reference models used for launching MIA. For the MIA Guess baseline, we report the best performance achieved across all candidate MIA methods. We first observe that our method consistently outperform all baselines under all settings under 42 reference models (which is the minimal number of reference models required for MLE baselines), whether on well-generalized setting such as WRN28-2 on CIFAR-10 or more challenging datasets like CIFAR-100 and Tiny-Imagenet. These latter datasets have fewer data records per class, thereby increasing their susceptibility to memorization and generally making them easier tasks for DUCI. We then tested the case with extremely limited computational budget (i.e., 1-8 reference models). This is challenging, as the scarcity of empirical samples for estimating $P(\hat{m}=1|m=0)$ and $P(\hat{m}=1|m=1)$ will lead to less accurate debiasing. Additionally, existing MIAs themselves also suffer from reduced strength when the number of reference models is small (Carlini et al., 2022; Zarifzadeh et al., 2024). Remarkably, our method achieves significantly smaller estimation error than vanilla MIA baselines even with as few as 1 model, i.e., *our method can provide reliable proportion estimates even under a severely restricted computational budget.*

Figure 2 further illustrates the detailed comparison under different amount of total computational budgets, i.e., different number of reference models, under the setting of FC-5 trained on CIFAR-10. Notably, the advantage of our method compared with MLE methods is significant when given less computation budget. This is because the accuracy of MLE methods relies heavily on the quality of approximating the $P(o|p)$ for all possible $p \in [0,1]$, a task that becomes challenging with a limited computation budget. Our results indicate that, under such constraints, the distribution estimation process required for MLE introduces more significant bias compared to our method. For all methods, we observe a general trend of reduced prediction error as the number of reference models increases. This improvement is expected because more reference models provide a more accurate estimation of the likelihood $P(o|p)$ for MLE methods, the distribution for the MIAs, and TPR and FPR estimation for our debiasing methods (Section 4), thereby reducing prediction errors.

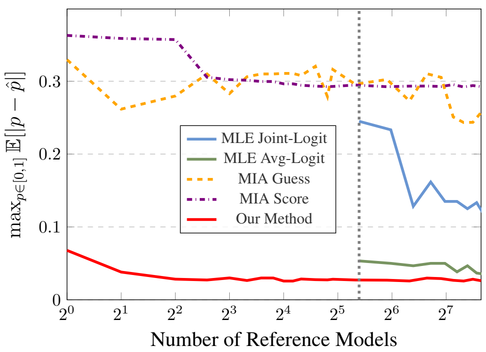

Figure 2: Maximum MAE of all methods under an increasing number of reference models. The grey dotted line represents the minimum computational budget required for MLE methods. Our method can yield meaningful predictions even using single reference model.

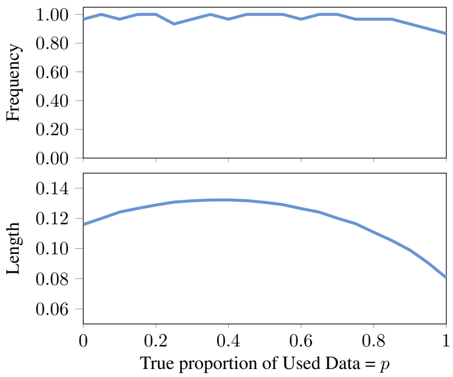

Figure 3: *(Confidence Interval)* **Top:** Frequency of the true proportion p falling within the predicted 95% confidence interval for different p. The frequency was computed across 30 interval predictions for each p. **Bottom:** Average width of the predicted interval over 30 tests for each p.

We also allocate a substantial budget—a total of $30 \times 21 = 630$ reference models—to the strongest baseline, the MLE methods (grey rows in Table 1), to fully exploit their potential. Note that MLE methods, which utilize complete proportion distribution information, can serve as idealized baselines when computational costs are unlimited. The comparison demonstrates that our method can achieve *comparable estimation quality* to MLE methods while incurring *significantly lower computational costs*, i.e., $100\times$ smaller. Due to the huge computational costs, we only presents the results of FC-5 models trained on CIFAR-10 for this comparison.

Confidence interval for dataset usage inference Our methodology also aims to deliver an interval estimate with statistical assurances to encapsulate the uncertainty inherent in dataset usage inference. To evaluate the efficacy of our interval estimation, we conducted tests on the same series of target models—30 trials for each proportion—employing a significance level of $\alpha = 0.05$. The confidence intervals were computed in accordance with Equation (22), where the degrees of freedom is $|X| - 1 = 499$, leading to a t-value of $t_{\alpha/2} = 1.96$.

Figure 3 (Top) displays the frequency of the true proportion p falling within the predicted 95% confidence interval. There is a consistently high frequency (over 0.9), which demonstrates the high coverage of our interval predictions. Moreover, the intervals generally have small length, as shown in Figure 3 (Bottom), suggesting that the overall uncertainty of our proportion prediction is small (around 0.1). Note that here we derive the additive confidence interval (absolute uncertainty) instead of the multiplicative confidence interval (relative uncertainty) because the error in Dataset Usage Inference task is not continuous. (e.g., For a small dataset in size 100, the unit of error is 1%.) Additionally, we observe that our confidence intervals exhibit greater length when $p = 0.5$. This is reasonable as the number of possible p-proportion subsets of the target dataset is the largest when $p = 0.5$, thus making dataset usage inference inherently more complex at $p = 0.5$.

5.4 DISCUSSIONS

We next discuss the properties, further improvement and potential application of our method.

Dataset usage inference under special sampling challenges. While our dataset-level debiasing methods (Equation (7) and Equation (8)) are statistically guaranteed to be unbiased under common sampling scenarios like i.i.d. and uniformly random sampling (see Footnote 1 for clarification), the question remains: ***can they also achieve near-unbiasedness under special or adversarial sampling?***

Table 2: Relative stability of our methods under special sampling, measured by Maximum mean absolute error (MAE) ($\max_{p\in[0,1]} \mathbb{E}[|p-\hat{p}|]$) over 21 proportions.

Sampling	MLE (Best)	MIA Guess	Ours
Random	0.0533	0.2933	**0.0619**
Significant First	0.6117	0.3876	**0.1089**
Insignificant First	0.6301	0.3787	**0.1077**

Table 3: Dataset usage estimation error decreases as the (relative) size of the target dataset increases for all methods. Errors are measured by Maximum Mean Absolute Error (MAE) $\left(\max_{p\in[0,1]} \mathbb{E}[|p-\hat{p}|]\right)$ across 21 proportions using WRN28-2 trained on CIFAR-10

$\lvert X \rvert$	MLE (Best)	MIA Guess	Ours
500	0.1383	0.2889	**0.0722**
5000	0.0849	0.268	**0.0528**
25000	0.08	0.2659	**0.0442**

To the best of our knowledge, all previous work considers a uniformly random (or i.i.d.) sampling scenario in method design (or evaluation) (Li et al., 2023; Tang et al., 2023; Maini et al., 2021). This means that when the target model employs special but unknown sampling, the sampling mismatch presents a significant challenge for all dataset usage inference techniques.

As noted in Footnote 1, the unbiasedness of our methods in special sampling hinges on the correlation term $\frac{\mathrm{Corr}_i(\mathrm{TPR}_i - \mathrm{FPR}_i, p_i)}{\mathrm{TPR}-\mathrm{FPR}}$. (Under i.i.d. or uniformly random sampling, this term is zero since $p_i = p$ for all $i = 1, 2, \ldots, |X|$.) For special sampling, we construct subgroups where $\mathrm{TPR}_i - \mathrm{FPR}_i \approx TPR' - FPR'$ and debias within each subgroup to achieve an unbiased estimator of p_i (see Figure 6 for empirical validation).

To evaluate the performance of all methods under the special sampling challenge, we conduct experiments where the model trainer applies dataset selection techniques (Guo et al., 2022; Sorscher et al., 2022) to select p-percentage of the target dataset to include in training. Specifically, we adopt an error-based coreset selection method, EL2N (Paul et al., 2021), to rank the target dataset by ”data significance”[3] using 10 trained models that have not seen the target dataset. Then, we constructed the probability vector $\boldsymbol{p}^{(p)}$ over the target dataset X according to this ”data significance” for each $p \in \{0.00, 0.05, \ldots, 1.00\}$. The data in the target dataset that used in the training of target models are therefore sampled according to each $\boldsymbol{p}^{(p)}$. By contrast, the reference models are trained under uniformly random $p\%$ subset of the target dataset[4]. These scenarios represent the least favor samplings for our methods, as the EL2N is closely related to $TPR_i - FPR_i$, which introduces strong correlation between $\boldsymbol{p}$ and the errors.

Table 2 shows that all other methods degrade in performance under special sampling challenges and exhibit similar performance across the two symmetric sampling methods. This confirms our earlier discussion on the distribution mismatch challenge from special sampling: they make mistakes occur early or late in the process. Nevertheless, the effectiveness of our debiasing method largely remains and yields better performance than all other methods. By contrast, the MLE method suffers the most significantly under the mismatch, since its superiority comes from fully recover the possible model distribution when data sampling scheme is known.

Dataset usage estimation error decreases as the ratio $\frac{|X|}{D_{\text{Train}}}$ increases. In our previous experiments, we focused on a challenging scenario where the owner's dataset (size 500) constitutes only a small portion of the target model's training set D_{Train}, i.e., the ratio $\frac{|X|}{D_{\text{Train}}}$ was small. We now examine the impact of the (relative) dataset size on the performance of dataset usage inference. To do this, we evaluated our method across various target dataset sizes, ranging from 500 to 25,000 (i.e., relative ratios from 0.02 to 1.00), while keeping the target model's training dataset fixed at 25,000 examples (half the CIFAR-10 training set).

Table 3 demonstrates that all methods benefit from the increasing relative dataset size. This is because, as the ratio $\frac{|X|}{D_{\text{Train}}}$ increases, the randomness from the remaining data in D_{Train} decreases, making dataset usage inference easier. For instance, in the extreme case where $\frac{|X|}{D_{\text{Train}}} = 1$, there is no randomness at 100% utilization. This observation highlights that *prior evaluations in the literature (Li et al., 2023; Maini et al., 2021)*, which *test whether a model was trained on private dataset A versus an alternative dataset B, are oversimplified and offer limited insights into real-world*

---

[3] A term we use cautiously, as EL2N can only be considered as a reflection of data value.

[4] Since the sampling method used in the target model is inaccessible, assuming random sampling is the a reasonable (if not the best) strategy a dataset owner can take for simulating the model trainer.

scenarios. Additionally, we noticed that while the improvement for MIA baselines is modest (as MIA does not leverage dataset-level statistics and benefits only from reduced randomness at larger ratios), our debiasing process sees a more substantial gain. This is because the sampling error in our debiasing is proportional to $\frac{1}{\sqrt{N\cdot|X|}}$ (see formal proof in Appendix F), meaning that as the target dataset size increases, our debiasing procedure becomes more accurate.

On (the possibility of) improving our method by leveraging correlation in membership prediction process Our method implicitly that the membership prediction process is ”independent” across records. Specifically, the prediction of membership probability for one record i does not affect the prediction for another record j, i.e., $\mathbb{E}[\hat{p}_i\hat{p}_j] = \mathbb{E}[\hat{p}_i]\mathbb{E}[\hat{p}_j]$. *Will the ignorance of “correlations” between* $\hat{p}_i$ *and* $\hat{p}_j$ *make our method sub-optimal?* We evaluate the magnitude of correlations between the predictions for pairs of records in the target dataset (no duplication) according to the pairwise second-order error (Equation (34) in Appendix H). Figure 4 show that most data pairs have near-zero covariance between their membership predictions. This aligns with the theoretical findings of small correlation between records in prior works (Pillutla et al., 2023). This trend suggests *limited potential* to improve our method by leveraging record correlations on *natural* dataset.

Case study: quantifying book copyright infringement Copyright infringement in books can be seen as a specific case of data usage inference: given a model likely infringing on book copyrights, the authors or publishing houses would want to determine the extent of content usage. We conduct experiments on GPT-2 (Radford et al., 2019) and 50 new books with first editions in 2023, unseen by GPT-2, from the BookMIA dataset (Shi et al., 2023). We fine-tune pre-trained GPT-2 on varying proportions of these books to create target models. The sequences in books are sampled **continuously**, **rather than randomly and uniformly**, from the copyrighted books in order to simulate more realistic scenarios. Results in Table 4 show that both membership inference baselines (MIA Guess and MIA Score in Section 5.2) struggle with varying extents of infringement, particularly when infringement is less significant. In contrast, our method estimates the extent of infringement much more reliably, with a maximum MAE (Equation (3)) of 0.168, compared to 0.335 for MIA Guess and 0.5 for MIA Score. Details of our experiment and further analysis are in Appendix I.

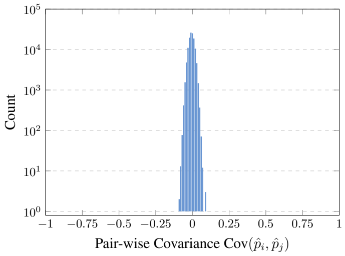

Figure 4: Histogram of the covariance Cov$(\hat{p}_i, \hat{p}_j)$ between data usage predictions of all data pairs i and j in the target dataset.

Table 4: *(Book copyright infringement)* Mean Absolute Error (MAE) $\mathbb{E}[|\hat{p}_i - p|]$ for all methods under different proportion p.

p	MIA Guess	MIA Score	Our Method
0.0	0.3350	0.5019	**0.1356**
0.2	0.2200	0.4041	**0.1379**
0.4	0.2875	0.3270	**0.1290**
0.6	0.2125	0.2245	**0.1684**
0.8	0.1200	0.1228	**0.0874**
1.0	0.1750	0.0091	0.0188
$\max_p$ MAE	0.3350	0.5019	**0.1684**

6 CONCLUSION

We identify and formally define the problem of Dataset Usage Cardinality Inference (DUCI), which quantifies dataset utilization—a significant issue that has been overlooked or misinterpreted in prior literature. To solve this, we propose a low-cost, reliable, and unbiased estimator for DUCI, along with confidence intervals for uncertainty. Our extensive experiments demonstrate the effectiveness of our approach, including its practical application in detecting book copyright infringement.

Limitations and future works We have thoroughly discussed the special sampling challenges in DUCI in Table 2 and demonstrated improved robustness of our methods compared to baselines. However, there is still room for further enhancement, particularly in achieving statistical guarantees under adversarial samplings. Another promising avenue is applying our methods to related tasks, such as dataset neighborhood inference. A potential solution for this could be extending our debiasing and aggregation methods, which are independent of specific MI designs, to orthogonal membership identification techniques like user-level inference Li et al. (2022a); Chen et al. (2023); Kandpal et al. (2023). Given that whether similar (but not copyrighted data) data the target models trained on should be members is still an open question, we leave these extensions for future work.

REFERENCES

Martin Bertran, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu. Scalable membership inference attacks via quantile regression. *arXiv preprint arXiv:2307.03694*, 2023.

Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems*, 36, 2024.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2017.

California Legislature. California consumer privacy act of 2018. Assembly Bill No. 375, Chapter 55, 2018. URL `https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375`.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Guangke Chen, Yedi Zhang, and Fu Song. Slmia-sr: Speaker-level membership inference attacks against speaker recognition systems. *arXiv preprint arXiv:2309.07983*, 2023.

European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). Official Journal of the European Union, L119, 1-88, 2016.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pp. 181–195. Springer, 2022.

Junfeng Guo, Yiming Li, Lixu Wang, Shu-Tao Xia, Heng Huang, Cong Liu, and Bo Li. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. *arXiv preprint arXiv:2310.14942*, 2023.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 705–714, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

Hongsheng Hu, Zoran Salcic, Gillian Dobbie, Jinjun Chen, Lichao Sun, and Xuyun Zhang. Membership inference via backdooring. *arXiv preprint arXiv:2206.04823*, 2022.

Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. Revisiting membership inference under realistic assumptions. *arXiv preprint arXiv:2005.10881*, 2020.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pp. 1376–1385. PMLR, 2015.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A Choquette-Choo, and Zheng Xu. User inference attacks on large language models. *arXiv preprint arXiv:2310.09266*, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Guoyao Li, Shahbaz Rezaei, and Xin Liu. User-level membership inference attack against metric embedding learning. *arXiv preprint arXiv:2203.02077*, 2022a.

Yiming Li, Ziqi Zhang, Jiawang Bai, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. Open-sourced dataset protection via backdoor watermarking. *arXiv preprint arXiv:2010.05821*, 2020.

Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. *Advances in Neural Information Processing Systems*, 35:13238–13250, 2022b.

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022c.

Yiming Li, Mingyan Zhu, Xue Yang, Yong Jiang, Tao Wei, and Shu-Tao Xia. Black-box dataset ownership verification via backdoor watermarking. *IEEE Transactions on Information Forensics and Security*, 2023.

Zheng Li and Yang Zhang. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 880–895, 2021.

Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership inference attacks by exploiting loss trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2085–2098, 2022.

Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018.

Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. A pragmatic approach to membership inferences on machine learning models. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 521–534. IEEE, 2020.

Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. *arXiv preprint arXiv:2104.10706*, 2021.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset? *arXiv preprint arXiv:2406.06443*, 2024.

Yuantian Miao, Xue Minhui, Chao Chen, Lei Pan, Jun Zhang, Benjamin Zi Hao Zhao, Dali Kaafar, and Yang Xiang. The audio auditor: user-level membership inference in internet of things voice services. *Proceedings on Privacy Enhancing Technologies*, 2021:209–228, 2021.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607, 2021.

Krishna Pillutla, Galen Andrew, Peter Kairouz, H Brendan McMahan, Alina Oprea, and Sewoong Oh. Unleashing the power of randomization in auditing differentially private ml. *arXiv preprint arXiv:2305.18447*, 2023.

Saurav Prat. Fine-tuning gpt-2 for text generation. `https://debuggercafe.com/fine-tuning-gpt2-for-text-generation/`, 2023. Accessed: 2024-05-21.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pp. 5558–5567. PMLR, 2019.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: tracing through training. In *International Conference on Machine Learning*, pp. 8326–8335. PMLR, 2020.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Congzheng Song and Vitaly Shmatikov. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 196–206, 2019.

Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2615–2632, 2021.

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

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhensu Sun, Xiaoning Du, Fu Song, Mingze Ni, and Li Li. Coprotector: Protect open-source code against unauthorized training usage with data poisoning. In *Proceedings of the ACM Web Conference 2022*, pp. 652–660, 2022.

Ruixiang Tang, Qizhang Feng, Ninghao Liu, Fan Yang, and Xia Hu. Did you train on my dataset? towards public dataset protection with clean-label backdoor watermarking. *arXiv preprint arXiv:2303.11470*, 2023.

The New York Times Company. Complaint. `https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf`, December 2023. Civil Action No. 1:23-cv-11195, United States District Court for the Southern District of New York.

United States Code. U.s. copyright act, title 17, section 107. `https://www.copyright.gov/title17/92chap1.html#107`, 1976. Fair Use.

Nikhil Vyas, Sham M Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International Conference on Machine Learning*, pp. 35277–35299. PMLR, 2023.

Yuxin Wen, Arpit Bansal, Hamid Kazemi, Eitan Borgnia, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Canary in a coalmine: Better membership inference with ensembled adversarial queries. *arXiv preprint arXiv:2210.10750*, 2022.

Emily Wenger, Xiuyu Li, Ben Y Zhao, and Vitaly Shmatikov. Data isotopes for data provenance in dnns. *arXiv preprint arXiv:2208.13893*, 2022.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. Recall: Membership inference via relative conditional log-likelihoods. *arXiv preprint arXiv:2406.15968*, 2024.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3093–3106, 2022.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *ICML*, 2024.

Zihang Zou, Boqing Gong, and Liqiang Wang. Anti-neuron watermarking: Protecting personal data against unauthorized neural networks. In *European Conference on Computer Vision*, pp. 449–465. Springer, 2022.

A RELATED WORKS

Membership inference attack Membership inference attacks (MIAs) Ye et al. (2022); Shokri et al. (2017) aim to determine whether a specific data entry is part of the training dataset of a target model. Initially gained traction in medical research Homer et al. (2008), MIAs were first introduced into machine learning by Shokri et al. (2017), subsequently inspiring numerous shadow model attacks Long et al. (2020); Sablayrolles et al. (2019); Song & Mittal (2021); Carlini et al. (2022). These shadow model attacks train a collection of *shadow models* (or *reference models*) to simulate the target model's behavior. All shadow models are trained with the same architecture as the target model using random subsets of data that follow the same distribution as the target model's training data, including or excluding the target data point, to estimate a statistical distribution of certain score function (such as but not limited to loss Sablayrolles et al. (2019); Yeom et al. (2018), confidence Salem et al. (2018), entropy Song & Mittal (2021), or loss trajectory Liu et al. (2022)) reflecting the data point's presence or absence in the training set. Aside from shadow model attacks, an alternative line of research eliminates the need to train multiple shadow models. Instead, they queries the target model with *population data* that is related to, or follows the same distribution as, the target point Wen et al. (2022); Bertran et al. (2023); Zarifzadeh et al. (2024); Jayaraman et al. (2020); Li & Zhang (2021); Long et al. (2018); Xie et al. (2024). These data are utilized as a reference for inferring membership. Individual MIAs cannot be directly adapted to solve the dataset utilization cardinality inference problem because they often cannot achieve perfect distinguishability, and their predictions vary drastically depending on the specific threshold or test used. Therefore, it is necessary to consider the uncertainty in the attack decision to reliably infer the extent of dataset usage. Moreover, the cost of conducting point-wise MIA tests for the entire dataset is another concern, which calls for an efficient adaptation.

A.0.1 DATASET OWNERSHIP VERIFICATION AND DATASET PROVENANCE TRACING

Dataset Ownership Verification (DOV) and Dataset Provenance Tracing are two related research area to our Dataset Usage Inference problem. Currently, all known dataset ownership verification techniques utilize dataset watermarking, which involves altering a subset of the data to embed a detectable signature Li et al. (2023; 2020); Tang et al. (2023); Li et al. (2022b); Sun et al. (2022); Sablayrolles et al. (2020); Hu et al. (2022); Zou et al. (2022); Wenger et al. (2022); Maini et al. (2021); Song & Shmatikov (2019). In addition to these, a handful of non-watermark-based dataset tracing techniques, although not specifically designed for the dataset ownership problem, can be adapted to address the dataset ownership problem or are related to this issue. These methods primarily encompass variants of membership inference techniques that operate beyond the individual level Song & Shmatikov (2019); Maini et al. (2021); Li et al. (2022a); Miao et al. (2021). Thi

Dataset watermarking Dataset watermarking Li et al. (2023); Tang et al. (2023); Li et al. (2020); Sun et al. (2022), predominantly utilizing poison-only backdoor attacks (that manipulate only the training data while keeping training components, model parameters, and structures intact Li et al. (2022c)), currently leads in the field of dataset ownership inference. Generally, these methods capitalize on the memorization capabilities of machine learning models during training to embed spurious features (information divergent from the true data distribution) into the model. This embedded knowledge is then used for subsequent verification: the detection of specific backdoor behaviors (targeted or untargeted) in the model's output during the verification phase indicates the model was trained with the protected dataset.

While the earlier watermarking techniques modified the target label of poisoned samples Li et al. (2020); Hu et al. (2022), contemporary and more sophisticated methods employ untargeted backdoor attacks Li et al. (2022b); Sun et al. (2022) or clean-label backdoor attacks Sablayrolles et al. (2020); Tang et al. (2023); Zou et al. (2022), which are significantly more stealthy due to their preservation of label consistency. Several works have extended their focus to different application scenarios, such as protecting open-source code Sun et al. (2022) or performing classification tasks without knowledge of labels Wenger et al. (2022). Additionally, recent works have attempted to minimize the impact on model utility when modifying samples Guo et al. (2023). However, all these modification-based methods are not suitable for scenarios where the goal is to protect the ownership of already released datasets, nor do they guarantee data utility in sensitive scenarios, such as with medical data.

Dataset Inference While several class-level membership inferences Li et al. (2022a); Miao et al. (2021) exist for object detection, such as voice or face recognition from an individual, these works are limited to their specific use cases and are not applicable to the broader scenarios considered in this paper: they require that the data group has a unique pseudo-label. For text data that lacks a clear boundary between records, Song & Shmatikov (2019) utilizes the long tail of the text data distribution within a user's document to carry out user-level MIA. The work most closely related to ours is Dataset Inference Maini et al. (2021), which hypothesizes that private data in general will have a larger distance to the decision boundary than public data, since the model always attempts to maximize this distance during training. It trains a linear binary classifier that takes in the *distance-to-boundary* for each data record to predict its membership. The underlying idea of this method is to use membership inference based on the average of a *loss-like signal* over the data group as a metric. However, all these methods still focus on solving the *all-or-none* binary question, and thus also fail to address the problem of inferring the dataset usage extent.

Another area of research in copyright infringement, such as Vyas et al. (2023), focuses on protecting generative models from producing copyrighted content, regardless of whether the model was trained on copyrighted data. Although this task may appear similar to ours, it focuses on different aspect of copyright and can be addressed quite differently—for example, by adding distance constraints to prevent outputs from closely resembling protected content. In contrast, our goal is to infer properties of a target model's training set based on its outputs. As this line of work is greatly distinct from ours, we did not discuss them in the scope of this paper.

B IMPLEMENTATION DETAILS FOR WATERMARKING AND DATASET INFERENCE

For both implementations, the target protected dataset X (for which we infer the usage) is of size 500. All models are Wide Residual Networks trained on 25,000 records from CIFAR-10 (that contains a randomly sampled fraction of the target dataset). This setting ensures that the target protected dataset is only a small subset of the target model's training set, making it a more practical scenario.

B.1 WATERMARKING

The watermarking baseline depicted in Figure 1 is an implementation of Algorithm 1 in (Li et al., 2023). Our objective is merely to demonstrate the limitations of watermarking methods in addressing the Dataset Usage Cardinality Inference (DUCI) problem. Consequently, we selected the most robust and straightforward backdoor watermarking techniques, disregarding stealthiness. We employed the targeted poisoned-label attack with a simple white square as the trigger pattern (examples are shown in Figure 5). Considering the random sampling inherent in the DUCI problem, we opted for a poison rate of 30% for the protected dataset X with a size of 500 (lower poison rates were tested, but the likelihood of a successful backdoor attack diminishes with reduced poison rates). We designated class 0 as the target label.

The following steps were undertaken:

1. We randomly sampled $m = 2000$ instances from the CIFAR10 test set, adhering to the methodology outlined in the original paper. Specifically, we only chose instances correctly classified by the benign model to minimize the impact of the model's intrinsic accuracy on the test. If the model exhibits low accuracy, it might erroneously categorize benign instances as poisoned, thereby influencing the confidence value.
2. Following Li et al. (2023), We obtained the confidence score at the ground-truth label for each benign sample x and its poisoned counterpart x' from the target model, denoted as $P_b = f(x)$ and $P_p = f(x')$, respectively.
3. We conducted a Pairwise T-Test: For each pair of samples, the difference in posterior probabilities $P_b - P_p$ was calculated. Utilizing these differences, a pairwise left tail T-test was performed to test the null hypothesis $H_0 : P_b = P_p + \tau$, where τ is a hyperparameter. We select the hyper-parameter τ as the one has best performance on the original task ($\tau = 0.02$).

B.2 DATASET INFERENCE

We implemented a simplified yet effective version of the approach described by Maini et al. (2021), replacing the random walk distance to the decision boundary of each data point with the direct logits of each data point queried on the target model. Empirically, we observed no performance degradation with this simplification, and the underlying principle remains intact[5]. This approach is based on a hypothesis that private data in general will have a larger distance (smaller loss and higher prediction confidence) to the decision boundary than public data, since the model always attempts to maximize this distance during training. Therefore, comparing the average distance/logits score across the entire data group with that of public data can help to make a decision. All subsequent steps adhere strictly to those outlined in the original paper:

1. We randomly sampled 40 data points each time from both the public and private datasets and queried the target model to compute their confidence scores, $\boldsymbol{c}$ (public) and $\boldsymbol{c_v}$ (private).
2. We repeated Step 1 for 10,000 iterations, calculating the mean values $\mu = \bar{\boldsymbol{c}}$ and $\mu_v = \bar{\boldsymbol{c_v}}$.
3. We conducted a Pairwise T-Test on the distributions of $\boldsymbol{c}$ and $\boldsymbol{c_v}$, calculating the p-value to test the null hypothesis (the dataset is not being used) $H_0 : \mu = \mu_v$.

---

[5] The distance-to-boundary in a linear model is essentially equivalent to the loss

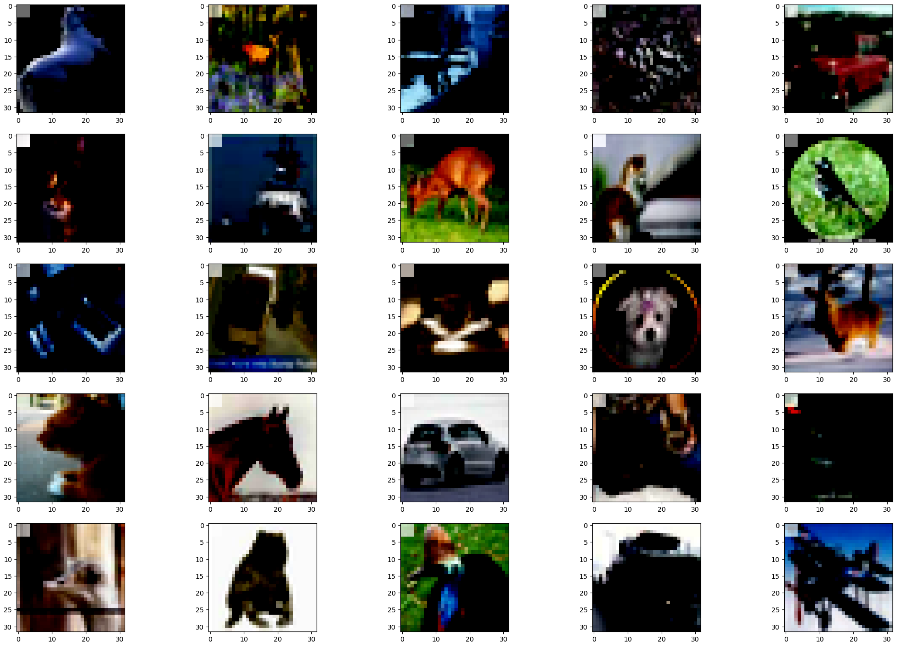

Figure 5: Visualization of poisoned data. Following standard method, we add the white square with random transparency as the trigger pattern.

C IMPLEMENTATION DETAILS OF OUR METHOD

We next introduce the implementation details of our method.

Algorithm 1 provides the pipeline of our method, where we use the existing membership inference algorithm, denoted as MIA, as the membership identification method to provide membership guesses $\hat{m}_i$ for all $x_i \in X$. We first trained N reference models, $\theta_1, \dots, \theta_N$, such that each reference models are trained on half of the dataset. Knowing the membership status of all $x_i \in X$ for these reference models, we can run MIA on them to get membership guesses for each x_i, and then compute the $P(\hat{m} = 1 | m = 0)$ (TPR) and $P(\hat{m} = 1 | m = 1)$ (FPR) as define in Equation (7) and Equation (8). The aggregation of debiased individual membership guesses gives us an unbiased estimator of the true proportion p.

C.1 SELECTION OF MEMBERSHIP INFERENCE ALGORITHMS

While any membership identification method that is capable of providing membership guesses can be applied to our method, we adopt two state-of-the-art MIA algorithm in this paper: Likelihood Ratio Attack (LiRA) Carlini et al. (2022) and Robust Membership Inference Attack (RMIA) Zarifzadeh et al. (2024). Given that both methods perform similarly with practical computational budget, and RMIA shows greater robustness under extremely limited computational resources, we present the RMIA results in the main paper and include results for both methods in the appendix.

Likelihood Ratio Attack (LiRA) Define the scaled logits for a point x of a model θ as $\text{logit}(x;\theta) = \log\left(\frac{\theta(x)_y}{1-\theta(x)_y}\right)$, where $\theta(x)_y$ is the value at the ground-truth class y of the prediction vector. LiRA Carlini et al. (2022) generates a membership guess $\hat{m}_i$ for a point x_i by conducting a likelihood ratio test, utilizing scaled logits as the test signal, as follows:

$$\hat{m}_i = \text{MIA}(x_i;\theta) = \mathbb{1}\left[\frac{P\left(\text{logit}(x_i;\theta) | \mathcal{N}(\mu_{in}, \sigma_{in}^2)\right)}{P\left(\text{logit}(x_i;\theta) | \mathcal{N}(\mu_{out}, \sigma_{out}^2)\right)} > \beta(x_i)\right] \tag{12}$$

where $\mathcal{N}(\mu_{in}, \sigma_{in}^2)$ and $\mathcal{N}(\mu_{out}, \sigma_{out}^2)$ are the Gaussian distributions of $\text{logit}(x_i;\theta)$ estimated on the N reference models trained with (in) and without (out) x_i, respectively.

Algorithm 1 Dataset Usage Inference: Infer the proportion p of data in a dataset X being used in the training of a given target model θ

```
Require: Dataset $X$, Training algorithm $\mathcal{T}$ (including the data distribution $\mathbb{D}$), Membership inference algorithm MIA, Target model $\theta$ (black-box access)
    Train reference models for debiasing (as few as one model)
for $j = 1$ to $N$ do
    $D \leftarrow \mathbb{D}$
    $\boldsymbol{m}^{(j)} \leftarrow \text{Bernoulli}(\frac{1}{2})^{|X|}$
    $\theta_j \leftarrow \mathcal{T}(D \cup \{\boldsymbol{m}^{(j)} \odot X\})$        ▶ Select data from $X$ by $\boldsymbol{m}^{(j)}$
end for
Save $\boldsymbol{m}^{1,\ldots,N} = \{\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N\}$
    Debias
Run MIA for $x_i \in X$ on $\theta_j$ for $j \in 1\ldots, \theta_N$ to estimate $P(\hat{m} = 1|m = 0)$ and $P(\hat{m} = 1|m = 1)$:
    $$P(\hat{m} = 1|m = 0) = 2\sum_{j=1}^{N}\sum_{i=1}^{|X|}\frac{\mathbf{1}_{\text{MIA}(\theta_j, x_i)=1 \text{ and } \boldsymbol{m}_i^{(j)}=0}}{N \cdot |X|}$$
    $$P(\hat{m} = 1|m = 1) = 2\sum_{j=1}^{N}\sum_{i=1}^{|X|}\frac{\mathbf{1}_{\text{MIA}(\theta_j, x_i)=1 \text{ and } \boldsymbol{m}_i^{(j)}=1}}{N \cdot |X|}$$
    Launch datase usage inference
for $i = 1$ to $|X|$ do
    $\hat{m}_i = \text{MIA}(x_i; \theta)$        ▶ Get individual membership guess on the target model
    Debiasing each individual membership guess $\hat{m}_i$:
        $$\hat{p}_i = \frac{\hat{m}_i - P(\hat{m} = 1|m = 0)}{P(\hat{m} = 1|m = 1) - P(\hat{m} = 1|m = 0)}$$
end for
Output the proportion estimation $\hat{p} = \sum_{i=1}^{|X|} \hat{p}_i$
```

The essential idea of LiRA is to compare the likelihood of observing the target model's signal given the distribution of models trained with the target point x against the distribution of models trained without x, where both distributions are estimated as Gaussian.

Robust Membership Inference Attack (RMIA)

$$\hat{m}_i = \text{MIA}(x_i; \theta) = \mathbb{1}\left[P_{z\sim\rho}(LR_\theta(x_i, z) \geq 1) \geq \beta(x_i)\right]$$
$$LR_\theta(x_i, z) = \left(\frac{P(x_i|\theta)}{P(x_i)}\right) \cdot \left(\frac{P(z|\theta)}{P(z)}\right)^{-1} \tag{13}$$

RMIA Zarifzadeh et al. (2024) compares the target point x_i with a number of population data $z \sim \rho$ from the same distribution, using the ratio $\frac{P(x_i|\theta)}{P(x_i)}$ as signal, where $P(x_i|\theta)$ is the prediction score (SoftMax) of output of the target model and $P(x_i)$ is the normalization term which is calculated by averaging $P(x_i|\theta_j)$ over all reference models θ_j for $j = 1, 2, \ldots, 2N$ trained with or without x_i, i.e., $P(x_i) = \frac{1}{N}\sum_{i=1}^{N} P(x_i|\theta_j) + \frac{1}{N}\sum_{i=N}^{2N} P(x_i|\theta_j)$ where $\theta_1, \theta_2, \ldots, \theta_{N-1}$ are models trained with x_i and $\theta_N, \theta_{N+1}, \ldots, \theta_{2N}$ are models not trained with x_i. For the offline setting, the $P(x_i)$ can be estimated only using models not trained with x_i, i.e., $\theta_N, \theta_{N+1}, \ldots, \theta_{2N}$. It can be extended to the offline setting using single reference model.

C.2 LiRA IMPLEMENTATION

Given the target points $x \in X$ and target model θ, we implement the LiRA following Carlini et al. (2022) as follows:

1. Train $2N$ reference models. Each are trained on uniformly and randomly sampled half of X together with uniformly and randomly sampled half of the population pool. Therefore, each $x_i \in X$ will have N reference models trained with it and the remaining N reference models trained without it.
2. For each x_i, estimate two Gaussian distributions on reference models: one distribution on models trained with data record x_i (denoted as $\mathcal{N}(\mu_{in}, \sigma^2_{in})$) and another distribution on models trained without x_i (denoted as $\mathcal{N}(\mu_{out}, \sigma^2_{out})$). The mean μ and variance σ^2 of each distributions are estimated on the scaled logits of each data record $x_i \in X$ over N reference models (trained with x or without x, respectively).
3. For each data record x_i, querying the target model to get the scaled logit $\text{logit}(x_i; \theta)$ and then compute the likelihood of observing it on two distributions: $P\left(\text{logit}(x_i; \theta)|\mathcal{N}(\mu_{in}, \sigma^2_{in})\right)$ and $P\left(\text{logit}(x_i; \theta)|\mathcal{N}(\mu_{out}, \sigma^2_{out})\right)$.
4. Compute the likelihood ratio $\text{LR}(x_i) = \frac{P\left(\text{logit}(x_i;\theta)|\mathcal{N}(\mu_{in}, \sigma^2_{in})\right)}{P(\text{logit}(x_i;\theta)|\mathcal{N}(\mu_{out}, \sigma^2_{out}))}$ for each data record x_i.
5. $\hat{m}_i = \mathbb{1}(\text{LR}(x_i) > \beta)$.

C.3 RMIA IMPLEMENTATION

In the original RMIA paper Zarifzadeh et al. (2024), the population data utilized in Equation (41) are the residual target data. Integrating RMIA scores directly into our aggregation approach would encounter issues of the differential rate of change as the true proportion of utilized data rises. To address this, we implemented the following adaptive version of RMIA using extra population data: Similarly, given the target models θ and the target points $x_i \in X$, we implement the RMIA as follows:

1. Train N reference models. Each are trained on uniformly and randomly sampled half of X together with uniformly and randomly sampled half of the population pool. Therefore, each $x_i \in X$ will have N_1 reference models trained with it and the remaining $N_2 = N - N_1$ reference models trained without it. (When $N >= 2$, we can ensure $N_1 = N_2$. When $N = 1$, we can do approximation as described below.)
2. For each $x_i \in X$, denote the reference models $\theta_1, \theta_2, \ldots, \theta_{N_1}$ as the models trained with x_i, and $\theta_{N_1+1}, \theta_{N_1+2}, \ldots, \theta_{N_1+N_2}$ as the models trained without x_i.

   (a) ($N >= 2$) Compute $\frac{P(x_i|\theta)}{P(x_i)} = \frac{f_\theta(x_i)_y}{\frac{1}{N_1}\sum_{j=1}^{N_1} f_{\theta_j}(x_i)_y + \frac{1}{N_2}\sum_{j=N_1+1}^{N} f_{\theta_j}(x_i)_y}$, where $f_\theta(x_i)_y$ are the model's confidence value at the ground-truth class.
   (b) ($N = 1$) We can use an linear approximation to compute the denominator. Assume x_i is not in the training set of θ_j. we can estimate the probability for it on a model trained with it by $af_{\theta_j}(x_i)_y + 1 - a$ We can estimate the the We select $a = 0.3$ according to Figure 6 in (Zarifzadeh et al., 2024). In this case, only N reference models are required for inference. (Note that the same N models can be shared among all x_i as the models trained without x_i.)

3. For $t = 1, 2, \ldots, T$:
   (a) Select a z^t from the population pool such that it is in the training set of N reference models and not in the training set of the remaining N reference models. Note that $z^t \notin X$ since it is selected from population pool.
   (b) Compute $\frac{p(z^t|\theta)}{p(z^t)} = \frac{f_\theta(z^t)_{y_z}}{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_{z^t}}(z^t)_{y_z} + \frac{1}{N}\sum_{i=1}^{N} f_{\theta_{-z^t}}(z^t)_{y_z}}$.
   (c) Compute $\text{LR}(x_i, z^t) = \left(\frac{p(x_i|\theta)}{p(x)}\right) \cdot \left(\frac{p(z^t|\theta)}{p(z^t)}\right)^{-1}$ for (x_i, z^t) pair.
4. For each $x_i \in X$, compute $P_z(\text{LR}(x_i, z) > 1) = \frac{\mathbb{1}[\text{LR}(x_i,z^t)>1]}{T}$
5. $\hat{m}_i = \mathbb{1}(P_z(\text{LR}(x_i, z) > 1) > \beta)$.

C.4 SELECTION OF THRESHOLDS β

For the choice of threshold in both LiRA and RMIA, we sweep over all possible threshold to select the β as the optimal threshold maximizing the Youden's index ($J = TPR - FPR$) Youden (1950) on reference modes, which provides the best balance between sensitivity (TPR) and specificity (1 - FPR). Depends on the computation of TPR and FPR in dataset usage cardinality inference problem, the threshold β can be computed either in the *individual-level* or *dataset-level*.

Individual-level threshold:

$$\hat{\beta}_i = \arg\max_{\beta}[P(\hat{m}_\beta(x_i;\theta) = 1|m = 1) - P(\hat{m}_\beta(x_i;\theta) = 1|m = 0)] \tag{14}$$

Here, $P(\hat{m}_\beta(x_i;\theta) = 1|m = 1)$ is the True Positive Rate (TPR) computed under threshold β for each data point x_i, and $P(\hat{m}_\beta(x_i;\theta) = 1|m = 0)$ is the False Positive Rate (FPR), similarly computed. Each of them is estimated on reference models.

Dataset-level threshold: The individual threshold may not be optimal when the number of samples is limited. In such cases, we can adopt dataset-level thresholds:

$$\hat{\beta} = \arg\max_{\beta}[P_X(\hat{m_\beta} = 1|m = 1) - P_X(\hat{m_\beta} = 1|m = 0)] \tag{15}$$

where $P_X(\hat{m}_\beta = 1|m = 1)$ is the True Positive Rate (TPR) as depicted in Equation (8) under threshold β, and $P(\hat{m}_\beta(x;\theta) = 1|m = 0)$ is the False Positive Rate (FPR) in Equation (7), computed in the same setting.

Our all results presented in the main paper are based on dataset-level threshold and dataset-level debiasing due to the smaller sampling error under limited reference models.

C.5 EFFECT OF MIA CHOICES

By focusing on the two variants of our methods, we observe that *RMIA combined with debiasing performs better than LiRA combined with debiasing*, especially when the number of reference models is small (Figure 2). This is possibly because by design, LiRA solely relies on computations on the reference models to perform the MIA. By contrast, RMIA explicitly incorporate population data in its test in addition to reference models, thus enjoying lower prediction errors, especially under limited number of reference models.

D ANALYSIS OF POTENTIAL ERRORS FROM SIMPLIFYING RECORD-LEVEL DEBIASING TO DATASET-LEVEL DEBIASING

In Equation (6), we demonstrated that applying a record-level debiasing process to membership predictions before aggregation yields an unbiased estimator, $\hat{p} = \frac{1}{|X|}\sum_{i=1}|X|\hat{p}_i$ for the ground-truth proportion p. However, due to the computational cost and notable sampling error when the number of reference models is small, we instead estimate the TPR and FPR across the entire dataset X. We next show that the simplification in the debiasing process will not introduce errors in many practical sampling scenarios, such as uniform or i.i.d. sampling (which are the common scenario considered in the long line of prior works in binary dataset inference literature listed in the Introduction section). Only in special cases where there is a strong correlation between the probability of sampling i-th point p_i and its $TPR_i - FPR_i$, there would be a error. However, this can be effectively mitigated by subgroup debiasing, as shown in Table 2.

In Equation (7) and Equation (8), we leverage dataset-level TPR and FPR (i.e., $\text{TPR} = \frac{1}{|X|}\sum_i \text{TPR}_i$ and $\text{FPR} = \frac{1}{|X|}\sum_i \text{FPR}_i$) to replace the individual TPR_i and FPR_i. This simplification avoids the computationally cost (or large sampling errors) of debiasing each $\hat{p}_i$. This simplification is justified because the proportion p is a dataset-level statistic, as analyzed below. To avoid confusion, we introduce $\tilde{p}$ and $\tilde{p}_i$ as the estimators under dataset-level debiasing, and we next prove $\tilde{p} = \frac{1}{|X|}\sum_i \tilde{p}_i = \frac{1}{|X|}\sum_i \frac{\hat{m}_i - \text{FPR}}{\text{TPR}-\text{FPR}}$ is an unbiased estimator of p whenever a correlation term between $\text{TPR}_i - \text{FPR}_i$ and p_i is zero: Given

$$\mathbb{E}[\tilde{p}] = \mathbb{E}\left[\frac{1}{|X|}\sum_i \tilde{p}_i\right] = \frac{1}{|X|}\sum_i \mathbb{E}[\tilde{p}_i] = \frac{1}{|X|}\sum_i \mathbb{E}\left[\frac{\hat{m}_i - \text{FPR}}{\text{TPR}-\text{FPR}}\right] \tag{16}$$

Plugging Equation (5) into Equation (16), we can get:

$$\mathbb{E}[\tilde{p}] = \frac{1}{\text{TPR}-\text{FPR}} \cdot \frac{1}{|X|}\sum_i \left(p_i \cdot \text{TPR}_i + (1-p_i)\cdot \text{FPR}_i - \text{FPR}\right) \tag{17}$$

$$= \frac{\frac{1}{|X|}\sum_i \left[p_i \cdot (\text{TPR}_i - \text{FPR}_i)\right]}{\frac{1}{|X|}\sum_i \left[\text{TPR}_i - \text{FPR}_i\right]} \tag{18}$$

Given $p = \frac{1}{|X|}\sum_i p_i$, note that

$$\frac{1}{|X|}\sum_i \left[p_i \cdot (\text{TPR}_i - \text{FPR}_i)\right] = \frac{1}{|X|}\sum_i p_i \cdot \frac{1}{|X|}\sum_i (\text{TPR}_i - \text{FPR}_i) + \text{Corr}_i(\text{TPR}_i - \text{FPR}_i, p_i).$$

(Here, we use the term "correlation" instead of "covariance" because $\text{TPR}_i - \text{FPR}_i$ and p_i are not random variables.) Thus, we have:

$$\mathbb{E}[\tilde{p}] = p + \frac{\text{Corr}_i(\text{TPR}_i - \text{FPR}_i, p_i)}{\text{TPR}-\text{FPR}}. \tag{19}$$

The correlation term suggests that for many practical sampling methods (e.g., uniform sampling, i.i.d. sampling), this simplification results in an unbiased estimator for p because the correlation is 0. For specialized sampling methods, subgroup debiasing can ensure (empirical) unbiasedness, as discussed in Table 2, by making $\text{TPR}_i - \text{FPR}_i$ constant within each subgroup. This ensures that the correlation term for each subgroup is 0, providing a group-level debiasing approach. Note that the term "correlation" (slightly abused here) is used in the context of how the value of p_i, a pre-fixed constant in the DUCI pipeline, is determined. This is distinct from the correlation between membership predictions in Figure 4.

E DERIVATION OF THE CONFIDENCE INTERVAL

Modern machine learning algorithms and their data sampling processes are highly randomized. To capture the intrinsic uncertainty and confidence in dataset usage inference, we resort to the Lyapunov Central Limit Theorem (CLT) Billingsley (2017) to compute confidence intervals for our estimation method via aggregate statistics. We first introduce the Lyapunov CLT.

Definition E.1 (Lyapunov condition Billingsley (2017))**.** If a sequence of independent random variables $\{X_i\}_{i=1}^n$ satisfy $\mathbb{E}[X_i] = \mu_i < \infty, \mathbb{E}[(X_i - \mu_i)^2] = \sigma_i^2 < \infty$, and for some $\delta > 0$:

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left[|X_i - \mu_i|^{2+\delta}\right] = 0, \tag{20}$$

where $s_n^2 = \sum_{i=1}^n \sigma_i^2 > 0$, the Lyapunov condition is said to hold for $\{X_i, i = 1, \ldots, n\}$.

Theorem E.2 (Lyapunov Central Limit Theorem Billingsley (2017))**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of independent random variables with means $\mu_i = \mathbb{E}[X_i]$ and variances $\sigma_i^2 = Var(X_i)$. If the Lyapunov condition is satisfied, the following sum converges in distribution to a standard normal distribution:*

$$\frac{\sum_{i=1}^n (X_i - \mu_i)}{s_n} \xrightarrow{d} \mathcal{N}(0,1).$$

where $s_n = \sqrt{\sum_{i=1}^n \sigma_i^2}$ and $\xrightarrow{d}$ denotes the convergence in distribution.

Theorem E.2 states that if the individual statistics $\hat{p}_1, \cdots, \hat{p}_{|X|}$ in Equation (6) are independent (without needing to be sampled from the same distribution), and the Lyapunov condition is satisfied, their aggregation $\sum_{i=1}^{|X|} \hat{p}_i$ converge to a Gaussian distribution as $|X|$ increases. Therefore, applying Lyapunov CLT to Equation (4), we obtain the approximate 95% confidence interval for estimating p.

$$\hat{p} - t_{\alpha/2}\sqrt{\frac{s_{|X|}^2}{|X|}} \le p \le \hat{p} + t_{\alpha/2}\sqrt{\frac{s_{|X|}^2}{|X|}}. \tag{21}$$

Here, $\hat{p}$ is as given in Equation (4), $\alpha = 0.05$, $t_{\alpha/2}$ is the t critical value, and $s_{|X|}^2 = \frac{1}{|X|}\sum_{i=1}^{|X|} \sigma_i^2$, where σ_i^2 is the population variance of $\hat{p}_i$ which can be empirically estimated over the random trials. However, considering the cost of launching this confidence interval inference, we porpose to use the sample variance to overestimate $\frac{s_{|X|}^2}{|X|}$ (see proof in Appendix E.1). This gives us the following approximate 95% confidence interval for estimating p.

$$\hat{p} - t_{\alpha/2}\sqrt{\frac{s^2}{|X|}} \le p \le \hat{p} + t_{\alpha/2}\sqrt{\frac{s^2}{|X|}}. \tag{22}$$

Similarly, $\hat{p}$ is as given in Equation (4), $\alpha = 0.05$, $t_{\alpha/2}$ is the t critical value, and $s^2 = \frac{1}{|X|-1}\sum_{i=1}^{|X|}(\hat{p}_i - \hat{p})^2$, where $\hat{p} = \frac{1}{|X|}\sum_{i=1}^{|X|}\hat{p}_i$.

E.1 PROOF OF OVERESTIMATION $\mathbb{E}[s^2] \ge \frac{\sum_{i=1}^{|X|}\sigma_i^2}{|X|}$

Lemma E.3. *Given a set of independent random variables $\hat{p}_1, \cdots, \hat{p}_{|X|}$ with variances $\sigma_1^2, \cdots, \sigma_{|X|}^2$ (which may not be identically distributed), the sample variance s^2 defined as*

$$s^2 = \frac{1}{|X|-1}\sum_{i=1}^{|X|}(\hat{p}_i - \hat{p})^2 \quad \textit{where} \quad \hat{p} = \frac{1}{|X|}\sum_{i=1}^{|X|}\hat{p}_i$$

is at least as large as the average variance $\frac{\sum_{i=1}^{|X|}\sigma_i^2}{|X|}$.

Proof. We start from the expectation of the squared differences:

$$\mathbb{E}[(\hat{p}_i - \hat{p})^2] = \mathbb{E}[\hat{p}_i^2] - 2\mathbb{E}[\hat{p}_i\hat{p}] + \mathbb{E}[\hat{p}^2] \tag{23}$$

We next analyze each term.

For the first term, we have

$$\mathbb{E}[\hat{p}_i^2] = \sigma_i^2 + \mu_i^2 \tag{24}$$

where $\mu_i = \mathbb{E}[\hat{p}_i]$ for any i.

For the second term, we expand it as

$$\begin{aligned}
\mathbb{E}[\hat{p}_i\hat{p}] =& \mathbb{E}\left[\hat{p}_i \cdot \frac{1}{|X|}\sum_{j=1}^{|X|}\hat{p}_j\right] = \frac{1}{|X|}\sum_{j=1}^{|X|}\mathbb{E}[\hat{p}_i\hat{p}_j] &(25)\\
=& \frac{1}{|X|}\left(\mathbb{E}[\hat{p}_i^2] + \sum_{j\neq i}\mathbb{E}[\hat{p}_i]\mathbb{E}[\hat{p}_j]\right) &(26)\\
=& \frac{1}{|X|}\left(\sigma_i^2 + \mu_i^2 + \mu_i\sum_{j\neq i}\mu_j\right) &(27)\\
=& \frac{1}{|X|}\left(\sigma_i^2 + \mu_i^2 + \mu_i\left(\sum_{j=1}^{|X|}\mu_j - \mu_i\right)\right) &(28)\\
=& \frac{1}{|X|}\left(\sigma_i^2 + \mu_i\sum_{j=1}^{|X|}\mu_j\right) &(29)
\end{aligned}$$

where equation 26 is by computing $\mathbb{E}[\hat{p}_i\hat{p}_j]$ as follows.

- If $i = j$, $\mathbb{E}[\hat{p}_i\hat{p}_i] = \mathbb{E}[\hat{p}_i^2] = \sigma_i^2 + \mu_i^2$.
- If $i \neq j$, $\mathbb{E}[\hat{p}_i\hat{p}_j] = \mathbb{E}[\hat{p}_i]\mathbb{E}[\hat{p}_j] = \mu_i\mu_j$.

For the third term, by definition, we have

$$\mathbb{E}[\hat{p}^2] = \mathrm{Var}(\hat{p}) + (\mathbb{E}[\hat{p}])^2 = \frac{1}{|X|^2}\sum_{j=1}^{|X|}\sigma_j^2 + \left(\frac{1}{|X|}\sum_{j=1}^{|X|}\mu_j\right)^2 \tag{30}$$

By plugging equation 24, equation 29 and equation 30 into equation 23, we prove that

$$\mathbb{E}[(\hat{p}_i - \hat{p})^2] = \sigma_i^2 + \mu_i^2 - \frac{2}{|X|}\left(\sigma_i^2 + \mu_i\sum_{j=1}^{|X|}\mu_j\right) + \frac{1}{|X|^2}\sum_{j=1}^{|X|}\sigma_j^2 + \left(\frac{1}{|X|}\sum_{j=1}^{|X|}\mu_j\right)^2 \tag{31}$$

Thus

$$\begin{aligned}
\sum_{i=1}^{|X|}\mathbb{E}[(\hat{p}_i - \hat{p})^2] =& \sum_{i=1}^{|X|}\sigma_i^2 + \sum_{i=1}^{|X|}\mu_i^2 - \sum_{i=1}^{|X|}\frac{2}{|X|}\left(\sigma_i^2 + \mu_i\sum_{j=1}^{|X|}\mu_j\right)\\
&+ \sum_{i=1}^{|X|}\left(\frac{1}{|X|^2}\sum_{j=1}^{|X|}\sigma_j^2 + \left(\frac{1}{|X|}\sum_{j=1}^{|X|}\mu_j\right)^2\right)\\
=& \left(1 - \frac{1}{|X|}\right)\sum_{i=1}^{|X|}\sigma_i^2 + \sum_{i=1}^{|X|}\mu_i^2 - \frac{1}{|X|}\left(\sum_{j=1}^{|X|}\mu_j\right)^2 &(32)
\end{aligned}$$

Recognize that by Cauchy–Schwarz inequality, the terms involving μ_i can be lower bounded as follows.

$$\sum_{i=1}^{|X|} \mu_i^2 - \frac{1}{|X|}\left(\sum_{i=1}^{|X|} \mu_i\right)^2 \geq \sum_{i=1}^{|X|} \mu_i^2 - \frac{1}{|X|} \cdot |X| \cdot \sum_{i=1}^{|X|} \mu_i^2 = 0$$

Thus Equation (32) can be immediately written as

$$\sum_{i=1}^{|X|} \mathbb{E}[(\hat{p}_i - \hat{p})^2] \geq \left(1 - \frac{1}{|X|}\right) \sum_{i=1}^{|X|} \sigma_i^2$$

Thus by the definition of s^2 and by the additivity of expectation, we compute expected value of the sample variance as follows.

$$\begin{aligned}\mathbb{E}[s^2] =& \frac{1}{|X|-1} \sum_{i=1}^{|X|} \mathbb{E}[(\hat{p}_i - \hat{p})^2] \\ \geq& \frac{1}{|X|-1}\left(\left(1 - \frac{1}{|X|}\right) \sum_{i=1}^{|X|} \sigma_i^2\right) = \frac{\sum_{i=1}^{|X|} \sigma_i^2}{|X|}\end{aligned}$$

Therefore, s^2 is at least as larger as the average variance $\frac{\sum_{i=1}^{|X|} \sigma_i^2}{|X|}$.

□

E.2 PROOF OF THE LYAPUNOV CONDITION

We next prove that $\hat{p}_1, \cdots, \hat{p}_{|X|}$ satisfy Lyapunov condition (Definition E.1).

Lemma E.4 (Boundedness)**.** *The random variables $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{|X|}$ have bounded third moments.*

Proof. Observe that the estimated global TPR $= \frac{1}{|X|}\sum_{i=1}^{n}$ TPR$_i$ and FPR $= \frac{1}{|X|}\sum_{i=1}^{|X|}$ FPR$_i$ are constants over the random trials with respect to the membership identification algorithm and x_i, and by the assumption that TPR $\neq$ FPR (as stipulated in Equation (6)), we prove that $\hat{p}_i$ is upper-bounded and lower-bounded by constant. Specifically,

$$\hat{p}_i = \begin{cases} a = \frac{-\text{FPR}}{\text{TPR}-\text{FPR}} & \text{when } \hat{m}_i = 0, \\ b = \frac{1-\text{FPR}}{\text{TPR}-\text{FPR}} & \text{when } \hat{m}_i = 1. \end{cases}$$

Since $\hat{p}_i$ is bounded, i.e., $|\hat{p}_i - \mu_i| \leq C$ for some constant C, all moments are automatically bounded. Specifically, the third-order moments $\mathbb{E}[|\hat{p}_i - \mu_i|^3] \leq C^3$. □

Lemma E.5 (Limit Property)**.** *The sequence $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{|X|}$ satisfies the following:*

$$\lim_{|X|\to\infty} \frac{1}{s_{|X|}^{2+\delta}} \sum_{i=1}^{|X|} \mathbb{E}\left[|\hat{p}_i - \mu_i|^{2+\delta}\right] = 0 \, for \, \delta = 1$$

where $\mu_i = \mathbb{E}[\hat{p}_i] = p$ is the mean of $\hat{p}_i$, $\sigma_i^2 = Var(\hat{p}_i)$ is the variance of $\hat{p}_i$, and $s_{|X|}^2 = \sum_{i=1}^{|X|} \sigma_i^2$ is the sum of the variances.

Proof. Assume $\delta = 1$ for simplicity (a common choice). Denote M_i as the third moments of $\hat{p}_i$, i.e., $M_i = \mathbb{E}[|\hat{p}_i - \mu_i|^3]$. From Lemma E.4, we have that there exists C_i such that $M_i \leq C_i$ for all i. Without loss of generality, we can assume $C_1 = C_2 = \ldots = C_{|X|} = C$. (A simple choice is to let $C = \max(C_1, C_2, \ldots, C_{|X|})$.) Then, we have

$$\frac{1}{s_{|X|}^3} \sum_{i=1}^{|X|} M_i \leq \frac{1}{s_{|X|}^3} \sum_{i=1}^{|X|} C = \frac{C|X|}{s_{|X|}^3}.$$

Since $s^2_{|X|}$ grows linearly with $|X|$, we have $s^2_{|X|} = O(|X|)$. Therefore, from the algebraic properties of big-O notation, we have $s_{|X|} = O(\sqrt{|X|})$. Thus,

$$\frac{C|X|}{s^3_{|X|}} = \frac{C|X|}{(O(\sqrt{|X|}))^3} = \frac{C|X|}{O((\sqrt{|X|})^3)} = \frac{C|X|}{O(|X|^{3/2})} = \frac{C}{O(\sqrt{|X|})} \to 0 \text{ as } |X| \to \infty.$$

$\square$

Since the independent random variables $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{|X|}$ satisfy $\mathbb{E}[\hat{p}_i] = p_i < \infty$ and $\mathbb{E}[\hat{p}_i^2] - p_i^2 < \infty$ for any $i = 1, 2, \ldots, |X|$ (Lemma E.4), and have the limit property when $\delta = 1$ (Equation (20)) (Lemma E.5), we finish the proof that $\{X_i\}_{i=1}^n$ satisfies Lyapunov condition. Therefore, by applying the Lyapunov CLT Theorem E.2, the sum $\frac{\sum_{i=1}^{|X|} \hat{p}_i}{s_n}$ converges in distribution to a normal distribution as $|X|$ increases. This ensures that the confidence interval derived remains valid even when the individual $\hat{p}_i$ are not identically distributed but are independent.

F PROOF OF THE SAMPLING ERROR OF DATASET-LEVEL DEBIASING

Under the mild assumption in Appendix E that membership predictions on different target data records are independent (experimentally validated in Figure 4 that $MIA(\theta_j, x_i)$ and $MIA(\theta_j, x_{i'})$ are weakly correlated random variables for any $i \neq i'$, where the randomness is over θ_j and the randomness of the MIA algorithm). We next provide the formal proof:

Observe that Equation (7) and Equation (8) are empirical average of the membership prediction $MIA(\theta_j, x_i)$ over reference models $\theta_j, j = 1, \cdots, N$ and target data records $x_i, i = 1, \cdots, |X|$, and that the reference models are independently trained on random subsets of the target dataset. Then under the independent assumption and by the additive property of variance for independent random variables, we compute the variance of Equation (7) and Equation (8) as follows.

$$\text{Var}[Equation\ (7)] = \frac{1}{N \cdot |X|} \text{Var}[\mathbf{1}_{MIA(\theta\_j, x\_i)=1 \text{ and } x\_i \notin \theta\_j}]$$

$$\text{Var}[Equation\ (8)] = \frac{1}{N \cdot |X|} \text{Var}[\mathbf{1}_{MIA(\theta\_j, x\_i)=1 \text{ and } x\_i \in \theta\_j}]$$

where $\theta_j \xleftarrow{uniform} \{\theta_1, \cdots, \theta_N\}$ and $x_i \xleftarrow{uniform} \{x_1, \cdots, x_{|X|}\}$. Thus the sampling error of estimates in Equation (7) and Equation (8) is asymptotically upper bounded by $O\left(\frac{1}{\sqrt{N \cdot |X|}}\right)$.

G EXPERIMENT DETAILS AND ADDITIONAL RESULTS

G.1 EXPERIMENT SETTINGS

Evaluation setting: We consider an evaluation granularity of 5%, which is small enough for relatively accurate and meaningful dataset usage inference (DUI). To measure the DUI performance, we compute mean absolute error$\mathbb{E}[|p - \hat{p}|]$ over 30 random trials for each proportion $p \in \{0.00, 0.05, \ldots, 1.00\}$, and then evaluate the maximum error among all proportions, denoted as $\max_p \mathbb{E}[|p - \hat{p}|]$. In each trial, we train the target model θ on a freshly sampled p proportion of the protected dataset X (using uniformly random sampling or special sampling), combined with randomly sampled remaining training data from a data population pool (that does not overlap with X). The size of X in the experiments presented in Section 5.3 is 500, to maintain a challenging and practical relative size. (In practice, the model is usually trained on a mixture of datasets, and the overall size of these additional datasets is typically much larger than the target dataset.) We then test the performance under size 5,000 and 25,000 and show in Section 5.4 that dataset usage estimation error decreases as the dataset size increases.

Model training details: We consider three model architectures: a standard 5-layer fully connected network (FC-5), a wide ResNet with width 2 (WRN28-2) (Zagoruyko & Komodakis, 2016), and ResNet-34 (He et al., 2016). We use three benchmark datasets: CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and a scaled-down version of ImageNet (Tiny-ImageNet) (Le & Yang, 2015). We sample the protected dataset X from the training set of each benchmark dataset and use the remaining training set as the population pool.

For target models, we train them on a freshly sampled random p proportion of the protected dataset X, combined with randomly sampled remaining training data from the population pool (that does not overlap with X), such that the complete training dataset size for the target model is always half of the training set of the benchmark dataset. For example, the CIFAR-10 training set consists of 50,000 images. Therefore, the training set size for the target model is 25,000.

For the reference models used for our method and MIA baselines, we train each one on half of X combined with remaining data sampled from the population pool, ensuring the complete training set of the reference model has the same size as the target models.

For all target models and reference models, we trained them to achieve the following test accuracies: for CIFAR-10, 53% for FC-5, and 92% for WRN28-2; for CIFAR-100, 68% for WRN28-2, and 65% for ResNet-34; for Tiny-ImageNet, 58% for WRN28-2 and 54% for ResNet-34.

G.2 IMPLEMENTATION DETAILS OF BASELINES

MLE baselines The MLE baselines require to estimate the distribution for the observation of models trained on each proportion of X. To approximate the distribution of $o|p$, empirically, we train N reference models on freshly sampled random p-fraction of the target dataset X combined with remaining data sampled from the population pool for each of the 21 possible proportion values (i.e., $p = 0.00, 0.05, \ldots, 1.00$), employing the same algorithm $\mathcal{T}$ that was used for training the target models (in total $21N$ reference models). The complete training set size of the reference model is the same as the reference models used for our method. Regarding the two choices of observation o in the MLE baselines:

- **Joint logits of all records in target dataset (high-dimensional):** Building on prior research (Carlini et al., 2022) which demonstrated that scaled logits for individual data points can serve as effective observations and empirically follow a Gaussian distribution, we naturally extend this approach to a dataset-wide scale. We concatenate these point-wise logits into high-dimensional joint logits and model their distribution using a multivariate Gaussian.

$$P(o|p = q) = P(\text{logit}(X; \theta)|\mathcal{N}(\boldsymbol{\mu_q}, \boldsymbol{\Sigma_q})),$$

  where logit$(X; \theta)$ is the the logits of model θ on all records in X, and μ_q and $\boldsymbol{\Sigma_q}$ are its empirical mean and covariance matrix across N reference models trained with q proportion of X.
- **Averaged logits of all records in the target dataset (one-dimensional):** To reduce the dimensionality of the observation (and thus enable more accurate empirical mean and covariance estimation), we additionally consider the average of the logits over the dataset X, following the approach

in Maini et al. (2021). The likelihood is then modelled as the following one-dimensional Gaussian distribution.

$$P(o|p=q) = P\left(\frac{1}{|X|}\sum_{x\in X}\text{logit}(x;\theta)\middle|\mathcal{N}(\mu_q,\sigma_q^2)\right),$$

where μ_q and σ_q are the empirical mean and variance of the averaged logits of all records in X across N reference models trained with q proportion of X.

The **step-by-step procedure** is as follows:

1. For each $q \in \{0.00, 0.05, \ldots, 1.00\}$:
    (a) Train N reference models on a proportion q of X combined with data randomly sampled from the population pool.
    (b) Query the reference models using X to obtain observations o(e.g., the average logits) for all data in X and estimate the Gaussian distribution on these observations.
    (c) Query the target model using X to obtain the target observation o_{tar}.
    (d) Calculate the likelihood of observing o_{tar} given the distribution for q.
2. Select the q with the highest likelihood of observing o_{tar}.

MIA baselines To demonstrate the importance of debiasing, we also compare our methods against two intuitive baselines: the direct aggregation of MIA guesses and MIA scores. (Some might think that MIA scores indicate the probability or confidence that a data point is a member, which could be used to estimate dataset usage.) We employ the same $2N$ reference models as our method, with each model trained on half of dataset X:

- **MIA Guess:** This involves directly counting MIA membership predictions, where the proportion is calculated as the fraction of member predictions: $\hat{p} = \frac{1}{|X|}\sum_{i=1}^{|X|}\hat{m}_i$. For a fair comparison, we use the same MIA algorithms that were employed in our methods.
- **MIA Score:** This takes the average of MIA confidence scores across all data points: $\hat{p} = \frac{1}{|X|}\sum_{i=1}^{|X|}\hat{c}_i$, where $\hat{c}_i$ is the confidence score of RMIA (i.e., the probability of observing that i-th record's likelihood to be a member is greater than randomly sampled population data points).

The step-by-step implementation of the MIA algorithms is provided in Appendix C.2 and Appendix C.3. These implementations follow the standard procedures described in the original papers.

G.3 FORMING SUBGROUPS BY $TPR_i - FPR_i$

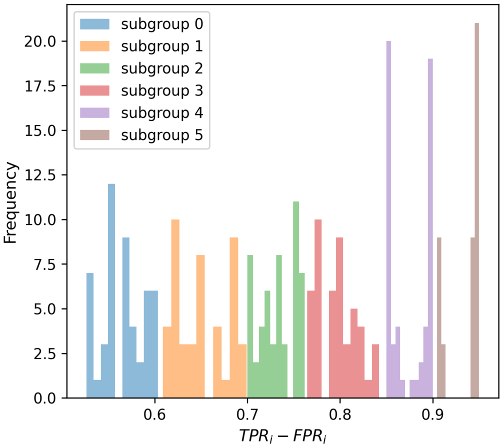

Figure 6: Samples of $TPR_i - FPR_i$ value for 6 subgroups.

G.4 RESULTS USING DIFFERENT MIA

In the main paper, we present the results of our method using RMIA (Zarifzadeh et al., 2024) due to its superior performance at a limited number of reference models. Additionally, we have implemented another state-of-the-art MIA algorithm, LiRA (Carlini et al., 2022). (Comparisons between RMIA and LiRA can be found in Appendix C.) All results in the main paper, when including the performance of LiRA, are provided in this section.

Comparison under fixed computation budget Figure 7 presents the complete version of Figure 2, including the performance of the MIA baseline using LiRA and our method using LiRA. Our methods (red and pink) are clearly more accurate in proportion prediction than all baselines. The comparisons between our method using LiRA and direct predictions from LiRA further confirm the effectiveness of our debiasing method. While our method using RMIA is more robust with a smaller number of reference models compared to using LiRA (analyzed in Appendix C.5), the MIA Baseline (RMIA) and MIA Score (RMIA) show less sensitivity to the number of reference models, possibly due to their inclusion of population data. Overall, they perform better with fewer reference models than the MIA Baseline (LiRA), which is consistent with the results reported in Zarifzadeh et al. (2024).

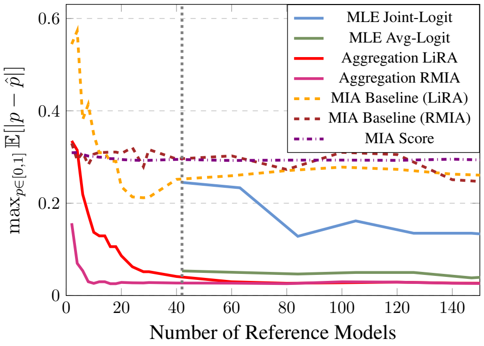

Figure 7: Maximum mean absolute error (MAE) of all methods under an increasing number of reference models. The grey dotted line represents the minimum computational budget required for MLE methods (2 for each proportion). Notably, our method can yield meaningful predictions even under limited computational budget (e.g., a single pair of reference models).

Confidence interval Figure 8 adds the confidence interval of our method using LiRA (corresponding to Figure 3 in the main paper). The top figure demonstrates the effectiveness of our methods using both membership identification methods, while the lower figure illustrates that our 95% confidence interval predictions are consistently tight.

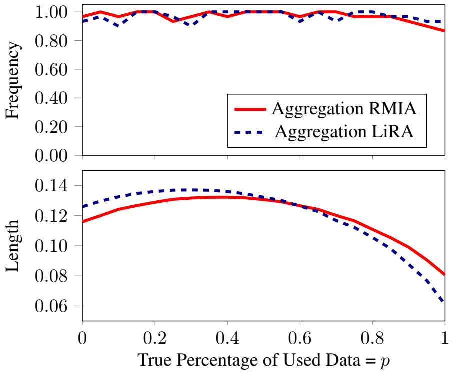

Figure 8: *(Confidence Interval)* **Top:** Frequency of the true proportion p falling within the predicted 95% confidence interval for different p. The frequency was computed across 30 interval predictions for each p. **Bottom:** Average width of the predicted interval over 30 tests for each p.

G.5 PERFORMANCE OF OUR METHOD IN DETERMINING BINARY DATASET USAGE

While methods restricted to binary predictions under an all-or-none dataset usage scenario cannot ensure consistent predictions for partial utilization (Figure 1), a method providing fine-grained estimates can naturally be reduced to solve the binary problem.

To illustrate this, consider the null hypothesis ($H_0 : s = s' + \tau$) used in prior binary dataset inference literature. For different contexts, s and s' can take the following forms:

1. **Dataset Inference Maini et al. (2021):** s and s' represent the distances to the decision boundary measured on a private dataset and a population dataset, respectively. This hypothesis assumes that if the model was trained on the private dataset, the distance measured on the private dataset would exceed that on the public dataset.
2. **LLM Dataset Inference Maini et al. (2024):** s and s' are the weighted aggregations of 52 MIA scores over the private and population datasets. This hypothesis assumes that merged MIA scores would be significantly higher for members than non-members over enough samples.
3. **Backdoor Watermarks Guo et al. (2023); Li et al. (2023); Tang et al. (2023); Hu et al. (2022); Li et al. (2022b); Wenger et al. (2022); Sablayrolles et al. (2020):** s and s' are the confidence score on the target label given backdoored inputs and given clean inputs. This hypothesis assumes that a model trained on a poisoned dataset (if successfully backdoored) will assign higher confidence to the target class when triggered, but not for clean inputs.

For DUCI, a straightforward simplification to the dataset inference problem can be made by set $s = \hat{p}$, $s' = 0$, and τ serves as a threshold, which may vary depending on the data type. Table 9 report the performance of our method adapted for solving the binary dataset inference problem.

Table 5: Comparison of p-values between DUCI and binary dataset usage algorithms for determining whether a dataset X (size 500) has been used. The complete training dataset of the target model has a size of 25,000. For p-values, a smaller value for **Dataset Used** is better, while a larger value for **Dataset Not Used** is better.

Methods	**p-value**	
	Dataset Used ↓	**Dataset Not Used** ↑
Backdoor Watermark (poison 30% of X)	7.10×10^{-5}	0.334
Backdoor Watermark (poison 100% of X)	$\mathbf{6.18 \times 10^{-54}}$	**1.000**
Dataset Inference	7.27×10^{-10}	0.937
Ours	$\mathbf{3.15 \times 10^{-51}}$	**1.000**

Consistent with the performance shown in Figure 1, all methods can perfectly solve the binary dataset usage problem when the significance level is set to common thresholds such as 0.05 or 0.01. This challenge becomes especially critical and significantly impacts performance when the dataset is not fully sampled or when the protected dataset's relative size is small. Our method performs exceptionally well, achieving comparable results to backdoor watermarking when the entire dataset is poisoned. In principle, the performance of Dataset Inference should be close to our method; however, the slight drop in performance may be attributed to the choice of signal, as the loss-based score is less distinguishable in distribution than the likelihood ratio-based score. We did not compare with (Maini et al., 2024) as combining multiple MIAs is orthogonal to our approach. Our method can debias any number of MIAs using the same reference models without retraining, with combination possible after debiasing if needed.

Finally, it is important to note that directly comparing the reported error of DUCI at $p = 0$ and $p = 1$ to that of dataset inference is inherently unfair. DUCI predicts a continuous value, whereas dataset inference is a simple binary classification task.

H ON (THE POSSIBILITY OF) IMPROVING DATASET USAGE INFERENCE VIA PAIR-WISE "CORRELATION"

Our method in Section 4 treat individual records in the target dataset separately, and thus implicitly require "independence" among records to perform well. **Will the ignorance of "correlations" between records predictions make our method sub-optimal for the problem of dataset usage inference? If the "correlations" between records predictions exists, how can we debias membership predictions?** In this section, we explore the room for improving our aggregation method via exploiting correlations among different records in the target dataset. We first show that under natural datasets (without manually crafted data), the room for improvement is small as the correlation between records prediction is almost zero. However, in Appendix H.2, we demonstrate that with a dataset containing specially designed canaries (i.e., correlation exists), our method can be extended to second-order debiasing given pairwise membership predictions, enabling more accurate dataset usage inference.

H.1 ON THE POSSIBILITY OF COVARIANCE REDUCTION

For the sake of understanding the role of correlation, let us now focus on the simplest scenario of inferring the usage of a dataset with two records. Under such settings, the mean squared error (MSE) for our proportion estimator in Equation (4) is is given by:

$$\begin{aligned}\mathrm{MSE}(\hat{p}) =& \mathbb{E}\left[\left(\frac{1}{2}(\hat{p}_1+\hat{p}_2)-p\right)^2\right] && (33)\\ =& \frac{1}{4}\mathrm{Var}(\hat{p}_1)+\frac{1}{4}\mathrm{Var}(\hat{p}_2)+\frac{1}{2}\mathrm{Cov}(\hat{p}_1,\hat{p}_2) && (34)\end{aligned}$$

where the last equality is by the unbiasedness property of $\hat{p}_1$ and $\hat{p}_2$. This naturally motivates us to ask: *Can we reduce the correlation (i.e., covariance term defined in Equation (34)) between data pairs in the target dataset, thereby reducing the error of dataset usage inference estimator?* To answer this question, we evaluate the magnitude of correlations among data records in the target dataset, and plot the histogram in Figure 4. We observe that a majority of data pairings incur near-zero covariance values. This is in line with prior works Pillutla et al. (2023) that observe small correlation between membership guesses on different data records. Only a small number of outliers (less than 0.01%) show positive covariance values larger than 0.05 – these pairs sometimes consist of records very similar to each other (examples are provided in Figure 10). These trends imply that the room for improving our aggregation methods via reducing covariance between data pairs is *generally very small*. It is an intriguing open problem as to whether our method could be further improved if the data owner is allowed to enforce high correlations among different data records, e.g., by modifying the target dataset.

H.2 ON THE POSSIBILITY OF PAIR-WISE BIAS REDUCTION (SECOND-ORDER DEBIASING)

If correlations between membership predictions exist (i.e., $\mathbb{E}[\hat{p}_i\hat{p}_j]-\mathbb{E}[\hat{p}_i]\mathbb{E}[\hat{p}_j]\neq 0$*), is it still possible to aggregate group-level statistics unbiasedly to predict dataset usage?* In this section, we provide an example of second-order debiasing (when all records in the target dataset are uniformly randomly sampled, i.e, $p_i = p$ for any record i when the ground-truth proportion is p).

Debiasing pair-wise membership guesses to estimate dataset usage proportion Suppose a membership identification method is able to predict pair-wise memberships $\hat{m}_i\hat{m}_j$ for any pair of points (x_i, x_j) where $i, j \in 1, 2, \ldots, |X|$. We now apply it to randomly sampled data record pairs from the target dataset. For each data pair (x_i, x_j), denote m_i and m_j as their ground-truth memberships, and denote $\hat{m}_i$ and $\hat{m}_j$ as their pair-wise membership guesses (by definition $m_i, m_j, \hat{m}_i, \hat{m}_j \in \{0, 1\}$). Let $P(\hat{m}_i\hat{m}_j = 00)$, $P(\hat{m}_i\hat{m}_j = 01)$, $P(\hat{m}_i\hat{m}_j = 10)$ and $P(\hat{m}_i\hat{m}_j = 00)$ be the frequency of each pair-wise guesses.

To estimate the dataset usage proportion, we need to perform debiasing operations (similar to Section 4 [6]) on the observed pair-wise membership guesses. For this, we first estimate the following matrix M (i.e., all off-diagonal values in this matrix represent error types, analogous to TPR and FPR) of conditional distribution of the pair-wise membership guesses $\hat{m}_i\hat{m}_j$ given different ground-truth pair-wise membership values m_im_j, on empirically trained *reference models* (of which we know the training datasets).

$$M = \begin{bmatrix} P(00|00) & P(00|01) & P(00|10) & P(00|11) \\ P(01|00) & P(01|01) & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ P(11|00) & \cdots & \cdots & P(11|11) \end{bmatrix},$$

where for brevity we denoted $P(a|b)$ to be $P(\hat{m}_i\hat{m}_j = a|m_im_j = b)$, and the randomness is over the training algorithm and the random data pairs.

Therefore, by the total probability, we have

$$M \cdot \begin{bmatrix} P(m_im_j = 00) \\ P(m_im_j = 01) \\ P(m_im_j = 10) \\ P(m_im_j = 11) \end{bmatrix} = \begin{bmatrix} P(\hat{m}_i\hat{m}_j = 00) \\ P(\hat{m}_i\hat{m}_j = 01) \\ P(\hat{m}_i\hat{m}_j = 10) \\ P(\hat{m}_i\hat{m}_j = 11) \end{bmatrix} \tag{35}$$

To estimate the dataset usage proportion, one only need to observe that when a randomly sampled q-proportion of the target dataset is used for training, the ground-truth pair-wise membership values follow the following distribution: $P(m_im_j = 11) = q^2$, $P(m_im_j = 01) = P(m_im_j = 10) = (1-q)q$, and $P(m_im_j = 00) = (1-q)^2$. By plugging them to Equation (35), we can estimate the dataset usage proportion by finding a q that enables the closest observations to the empirically observed pair-wise membership guesses, as follows.

$$\hat{p} = \arg\min_q \mathbb{E}\left[\left\| M \begin{bmatrix} (1-q)^2 \\ (1-q)q \\ q(1-q) \\ q^2 \end{bmatrix} - \begin{bmatrix} P(\hat{m}_i\hat{m}_j = 00) \\ P(\hat{m}_i\hat{m}_j = 01) \\ P(\hat{m}_i\hat{m}_j = 10) \\ P(\hat{m}_i\hat{m}_j = 11) \end{bmatrix} \right\|_2^2 \right]$$

H.2.1 EXAMPLE OF PAIR-WISE DEBIASING GIVEN LIKELIHOOD-BASED MEMBERSHIP PREDICTION

We begin by selecting the most basic pair-wise statistic—pair-wise likelihood (analogous to loss and considered the weakest MIA signal)—as the basis for generating pair-wise membership predictions. As baselines, we include likelihood-ratio-based methods (a stronger likelihood-based signal), per-point MIA (LiRA Carlini et al. (2022)), and the best signal (RMIA). Our results demonstrate that, under second-order debiasing, even the weakest signal performs comparably to the aggregation of membership predictions using the best first-order signal.

Pair-wise membership inference The idea of pair-wise membership inference is to design a test that can simultaneously guess the membership status of a pair of points (x_i, x_j) for any $i, j \in 1, 2, \ldots, |X|$. We can use the general maximum likelihood estimation to design a pair-wise test, as follows.

$$\hat{m}_i\hat{m}_j = \arg\max_{s \in \{00,01,10,11\}} P(o|m_im_j = s) \tag{36}$$

where we consider the observation o as averaged logits of the data pair (x_i, x_j) and model the distribution of this observation o when $m_im_j = s$ for each $s \in \{00, 01, 10, 11\}$ as a multivariate Gaussian, similarly as Equation (11).

---

[6] The second-order debiasing is a natural extension of straightforward debiasing since under uniformly random sampling, method in Section 4 is solving $\hat{p} = \arg\min_q \mathbb{E}\left[\left\| M \begin{bmatrix} (1-q) \\ q \end{bmatrix} - \begin{bmatrix} P(\hat{m}_i = 0) \\ P(\hat{m}_i = 1) \end{bmatrix} \right\|_2^2 \right]$ in principle.

Performance of the second-order debiasing We refer to our method presented in the main paper, which is based on the aggregation of debiased individual membership guesses, as the aggregating individual statistics method (Agg Individual) to distinguish it from the pair-wise test. In Figure 9, we observe that the pair-wise test enables better performance (smaller bias) for proportion estimation compared to the Agg Individual method using LiRA. This suggests (1) the second-order debiasing is accurate, and (2) that pair-wise statistics (naive likelihood value) may capture more information than the corresponding individual statistics (stronger likelihood ratio).

However, we observe that the pair-wise test does not outperform the RMIA-based Agg Individual method (which is stronger than the LiRA-based method). We hypothesize that this is because the score (likelihood based rank score) used in RMIA is much stronger than the score used in LiRA and our pair-wise test. The RMIA score captures more data information, bridging the gap between first-order and second-order statistics. It is an intriguing open question whether we can design more powerful pair-wise tests using stronger MIA scores like those used in RMIA.[7]

Additionally, it is important to note that the current pair-wise test requires much higher computational cost than the aggregation of individual statistics due to the computation of MLE for each sampled data pair. The computational cost for inference of the pair-wise test is $O(|X|^2)$, while our Agg Individual method is $O(|X|)$. Therefore, identifying ways to reduce this cost could be another direction for future research.

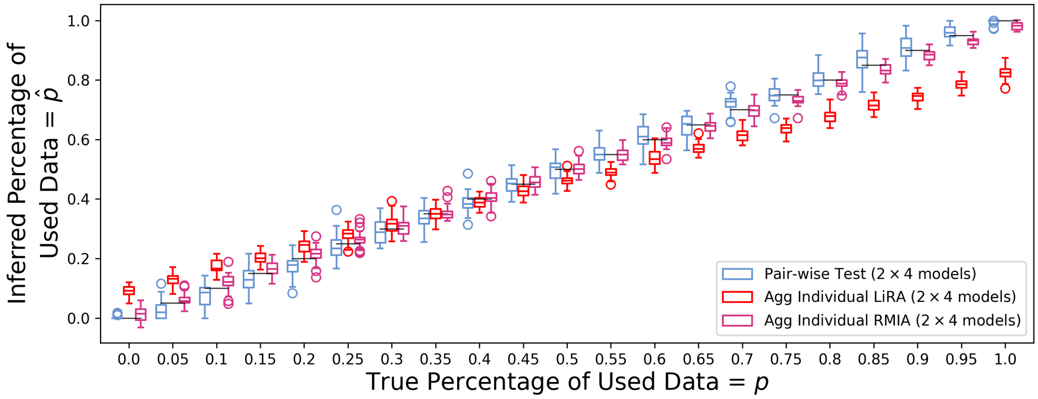

Figure 9: Comparison between the dataset usage inference using pair-wise statistics and individual statistics (red and pink) when only 4 pairs of reference models are used. For the pair-wise test, we follow Appendix H.2 and aggregate the debiased pair-wise membership guesses over 20000 randomly sampled data pairs in the target dataset. All evaluation settings are the same as **??**.

---

[7] There is no straightforward method for adapting RMIA scores for pair-wise testing.

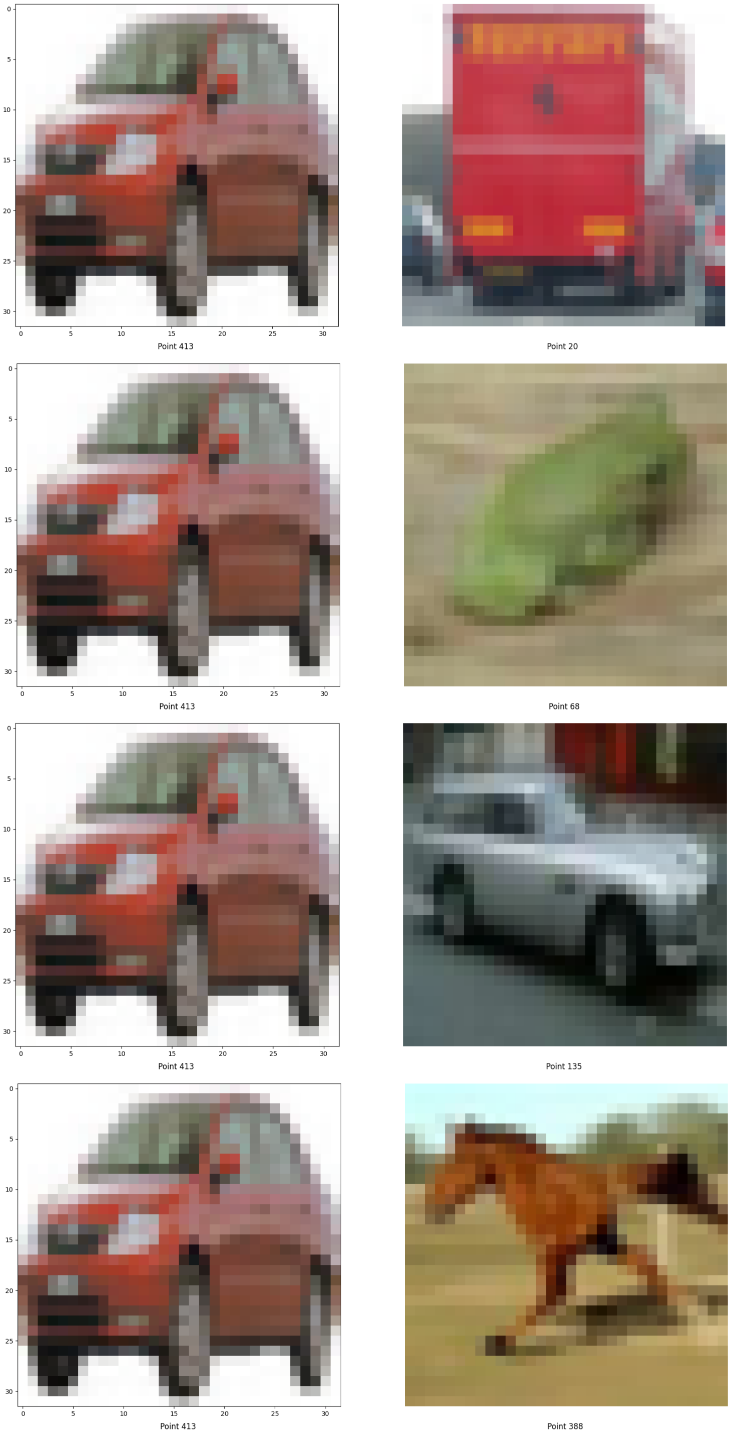

Figure 10: Examples of pairs with large covariance.

I CASE STUDY: QUANTIFYING BOOK COPYRIGHT INFRINGEMENT

The issue of copyright infringement in books can be considered a specific case of data usage inference: given a model that likely infringes on book copyrights, the author or publishing house would want to determine the extent to which their content is used. We conducted our experiment using pre-trained GPT-2 (Radford et al., 2019). For a rigorous setting, we selected books from the BookMIA dataset constructed by Shi et al. (2023). This dataset contains 50 new books with first editions in 2023, which could not have been included in the pre-training data of GPT-2, and 50 old books—known to be memorized by ChatGPT—whose memorization status by GPT-2 is uncertain. We split each book into chunks of 512 tokens.

Target Model: We partition the 50 new books in the BookMIA dataset Shi et al. (2023) into two subsets: a protected pool of 30 books and a population pool of 20 books. We fine-tune a total of $m = 4$ target models. For fine-tuning each of the m target models, we first randomly partition the protected pool into six disjoint sets of books $S_0, S_{0.2}, S_{0.4}, S_{0.6}, S_{0.8}$, and $S_{1.0}$, where each S_p contains $n = 5$ books as the target books for testing the case of infringement proportion p. The fine-tuning dataset for each target model is then constructed as follows:

1. **Protected pool chunks:** For each protected book $X \in S_p$ for $p \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$: we select the *beginning* p proportion of the chunks from the protected book X.
2. **Population pool chunks:**We add another $n = 5$ randomly sampled books from the population pool (as the remaining dataset).

Therefore, for each of these $m = 4$ target models, we have $n = 5$ target books with p proportion of content included in the training for each possible proportion value $p \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. This setup results in a total of 20 trials for each p.

Note that in this experiment, the used chunks are sequentially chosen rather than uniformly randomly chosen as in Section 5.1, to maintain the continuity of the text in a book.

Reference model: We train two reference models, each with a fine-tuning set that includes two parts:

1. **Protected pool chunks:** Each model uses half of the chunks from each of the 30 books in the protected pool. More specifically, both models are trained on all 30 books, but each model only uses half amount of the chunks from each book without overlap.
2. **Population pool chunks:** Each model also includes half of the chunks from $2n = 10$ randomly selected books from the population pool of 20 books as the remaining data.

This ensures that 1) for each chunk used in training the target model, one reference model is fine-tuned on it while the other is not, and 2) among the remaining data (exclusive the target books), we can select chunks that are present in only one of the reference models as the *population data* (used for launching the MIA in Zarifzadeh et al. (2024)).

All target and references models incur training loss of around 2.9 and test loss of around 3.2, which is even lower than the loss reported by Prat (2023), thus indicating the high performance of our fine-tuned model. The time for fine-tuning a single model (both target and reference models) for 6 epochs on a single NVIDIA GeForce RTX 3090 GPU is approximately 17 minutes.

Experiment setting We use the same MIA algorithm (Zarifzadeh et al., 2024) for all methods and select the optimal MIA threshold as described in Appendix C.4 on reference models. The above-mentioned population data are used for launching the MIA (Zarifzadeh et al., 2024). We consider MIA Guess and MIA Score as our baselines:

1. **MIA Guess**: For each book X with chunks $\{c_i\}_{i=1}^{|X|}$, we perform MIA on each chunk c_i to obtain an individual membership prediction $\hat{m}_i$. We then compute the ratio $\hat{p} = \frac{1}{|X|}\sum_{i=1}^{|X|} \hat{m}_i$.
2. **MIA Score**: Similar to MIA Guess, we conduct the same process but, instead of obtaining binary predictions, we directly aggregate the MIA scores as described in Section 5.2.

For our methods, we repeat the debiasing process described in Algorithm 1 given by $\mathbb{E}(\hat{m}|m = 0)$ and $\mathbb{E}(\hat{m}|m = 1)$ estimated on top of the membership guesses or scores provided by the MIA. The inference time for our methods on 30 target books, including the estimation of $P(\hat{m}|m=0)$ and $P(\hat{m}|m=1)$, on a single NVIDIA GeForce RTX 3090 GPU takes 408 seconds.

Results and analysis Although MIA on text generation models has been explored since (Song & Shmatikov, 2019), previous research typically conducts MIA by querying individual texts and gathering a single statistic to infer whether the entire collection is part of the training data. A few discussions on book copyright infringement issue in prior works (Radford et al., 2019) also focus on this Yes or NO question, which is just the binary version of our dataset usage inference problem.

Our experiment shows that membership inference methods struggle to distinguish varying extents of infringement, particularly when the infringement is less significant. As observed in Table 6, for both MIA baselines, the error in proportion estimation is large when the actual fraction of the book used in the target model's training is small. This is expected because when less content from a book is used, the target model learns a poor and incomplete representation of this book's data distribution, making it harder to distinguish membership. However, our method proves effective in more reliably estimating the extent of copyright infringement, despite the significant bias from the accumulation of individual membership prediction errors.[8] For example, our method achieves a maximum mean absolute error of 0.168 across all tested proportions, compared to 0.5 for the MIA Score and 0.335 for the MIA Guess.

Table 6: Mean Absolute Error $\mathbb{E}[|\hat{p}_i - p|]$ for different values of p. The lowest error under each proportion p

proportion p	MIA Guess	MIA Score	Our Method
0.0	0.3350	0.5019	**0.1356**
0.2	0.2200	0.4041	**0.1379**
0.4	0.2875	0.3270	**0.1290**
0.6	0.2125	0.2245	**0.1684**
0.8	0.1200	0.1228	**0.0874**
1.0	0.1750	0.0091	0.0188
$\max_{p\in[0,1]} \mathbb{E}[|\hat{p}_i - p|]$	0.3350	0.5019	**0.1684**

---

[8] We observed that the MIA score is more stable than MIA Guess due to the small number of chunks per book. For our method, we chose to debias the MIA score (replacing the binary score $\hat{m}_i$ in Equation (5), Equation (8), and Equation (7) with the continuous, non-binary MIA score).