OBSERVATIONAL AUDITING OF PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Differential privacy (DP) auditing is essential for evaluating privacy guarantees in machine learning systems. Existing auditing methods, however, pose a significant challenge for large-scale systems since they require modifying the training dataset—for instance, by injecting out-of-distribution canaries or removing samples from training. Such interventions on the training data pipeline are resource-intensive and involve considerable engineering overhead. We introduce a novel observational auditing framework that leverages the inherent randomness of data distributions, enabling privacy evaluation without altering the original dataset. Our approach extends privacy auditing beyond traditional membership inference to protected attributes, with labels as a special case, addressing a key gap in existing techniques. We provide theoretical foundations for our method and perform experiments on Criteo and CIFAR-10 datasets that demonstrate its effectiveness in auditing label privacy guarantees. This work opens new avenues for practical privacy auditing in large-scale production environments.

1 INTRODUCTION

Differential privacy (DP) auditing has become an important tool for evaluating privacy guarantees in machine learning systems. Recent advances in auditing methods that require only a single run have made it feasible to evaluate privacy for large-scale models without prohibitive computational costs (Steinke et al., 2024; Mahloujifar et al., 2025b). However, existing auditing approaches still require modifying the training dataset by injecting known entropy or canary data, which limits their applicability in industry-scale environments, where modifications to the training data pipeline require significant engineering overhead.

In this work, we propose a novel auditing methodology that eliminates the need for dataset modification. Our approach enables privacy evaluation using the natural nondeterminism present in the data distribution itself. We formalize and empirically validate this methodology in the setting of auditing Label DP, generalizable to privacy guarantees for any protected attribute. This capability addresses a significant gap in current auditing techniques, which have primarily focused on membership inference attacks (Shokri et al., 2017; Carlini et al., 2022) rather than attribute inference. In particular, existing methods for auditing Label DP either require adding out-of-distribution canaries to the training set (Malek et al., 2021) or are applicable only to a limited set of mechanisms (Busa-Fekete et al., 2024).

Observational privacy auditing cannot be done unconditionally, without making certain assumptions about the underlying distribution (Hernán & Robins, 2020, Chapter 3). Unlike anecdotal instances of privacy violations (Barbaro & Zeller Jr, 2006; Narayanan & Shmatikov, 2008; Carlini et al., 2021), auditing seeks to provide statistically valid measurements of memorization. In other words, the objective of privacy auditing is to establish *causality*—demonstrating that a model behaves in a certain way *because* it was trained on specific data. Traditionally, most rigorous membership inference attacks establish and measure causal effects through randomized control trials (RCTs), which require interference with the training data. By reframing auditing as a security game between two adversarial parties, our approach eliminates the need for training-time intervention.

Our key assumption is the availability of a distribution that approximates the ground truth. Concretely, the Label-DP auditing mechanism relies on access to a proxy label-generating distribution. The proxy does not need to match the ground truth distribution, provided the adversary cannot distinguish between them (with reasonable computational resources). Under this assumption, the counterfactual

examples generated by the challenger can be used to evaluate the attack on the model's claimed Label-DP guarantees *without training data intervention*. The attack can be made practical by using a model other than the target model as the proxy distribution. Further, in an incremental learning setting, earlier model checkpoints can be used as the proxy distribution, thus requiring no additional model training and minimal engineering overhead.

We demonstrate that our observational auditing framework provides accurate privacy bounds that match those obtained by interventional methods and provable privacy guarantees. Through a series of cryptography-inspired games we establish the theoretical foundations for auditing privacy without training data manipulation. By lowering the complexity of privacy auditing, our approach enables its application in a wider variety of contexts.

2 Background and Prior Work

Differential Privacy (DP) introduced by Dwork et al. (2006) is a leading framework for providing rigorous privacy guarantees in statistical data analysis and machine learning. In its standard formulation, DP bounds the impact that any single individual's record has on the outcome of a computation by constraining how much the distribution of outputs may differ between neighboring datasets. The definition's strong theoretical guarantees, resilience to arbitrary auxiliary data, and compositional properties have driven its adoption in academic research and industry deployments (Fioretto & Hentenryck, 2025).

The general DP framework can be adapted to settings where only certain parts of a dataset are considered private. This paper focuses on Label DP, which has emerged as an important objective for PPML, particularly in the domain of recommendation systems (Chaudhuri & Hsu, 2011; Ghazi et al., 2021; Malek et al., 2021; Wu et al., 2023). The following factors motivate Label DP as a uniquely valuable privacy concept:

- The label—representing the user's choice, expressed preference, or the outcome of an action—may be the only sensitive part of the record, with the rest being publicly available, static, non-sensitive data.
- In machine learning, labels are particularly vulnerable to memorization compared to other attributes since they most directly influence the loss function.
- In settings with mixed public/private features, instead of applying privacy-preserving techniques to sensitive features, one may exclude them from the model, potentially sacrificing some accuracy. In supervised learning, however, labels are indispensable—there is no analogous alternative to omitting sensitive features. In further separation, once training completes, labels can be safely discarded, whereas features must be available for inference.

Complementing the strong worst-case guarantees of DP that bound the privacy loss from above on all inputs, privacy auditing empirically measures the privacy loss on concrete instances, providing a lower bound on DP's numerical parameters. Privacy auditing can be used for finding bugs in claimed implementations of DP algorithms (Ding et al., 2018), advancing understanding of complex DP mechanisms (Malek et al., 2021; Nasr et al., 2023; 2025), or guardrailing models in a production environment (Agrawal & Book, 2025).

Privacy auditing consists of two components: a privacy game between the challenger and the attacker, and an auditing analysis that translates the attacker's success into lower bounds on the (ϵ, δ)-DP guarantee (or other forms of DP). The privacy game is characterized by the capabilities and resources of the parties, and the attacker's goals, such as reconstruction, membership or attribute inference. Auditing analyses, typically used for membership inference, can be applied to any stochastic privacy game (Swanberg et al., 2025). We show theoretical results for our label inference attack building on Steinke et al. (2024) and Mahloujifar et al. (2025b).

Membership inference attacks (MIAs)—where an adversary uses model access and knowledge of the data distribution to determine whether a sample was part of training—have received significant attention in the literature (Shokri et al., 2017; Yeom et al., 2018; Salem et al., 2019; Sablayrolles et al., 2019; Song & Shmatikov, 2019; Nasr et al., 2019; Leino & Fredrikson, 2020; Carlini et al., 2022; Ye et al., 2022; Zarifzadeh et al., 2024; Bertran et al., 2024). This attack category directly maps

to the differential privacy guarantee where two neighboring datasets differ in the presence of one training sample.

A less studied category is attribute inference attacks (Yeom et al., 2018) where the adversary reconstructs a protected attribute given access to a partial record, of which label inference attacks are a special case (Malek et al., 2021; Busa-Fekete et al., 2024). A difficulty and common pitfall with such attacks is to properly account for the adversary's baseline success, achieved by exploiting knowledge of the data distribution and correlations between the public and protected attributes (Jayaraman & Evans, 2022). The label inference attack of Malek et al. (2021) uses canaries with random binary labels, which sets the adversary's baseline accuracy to 0.5 and allows Label DP to be audited via standard MIA analyses. However, this attack interferes with the training dataset and can affect model performance if too many out-of-distribution samples are injected.

Most auditing methods in the literature are "interventional," as their privacy game involves manipulating the training dataset: MIAs require excluding a subset of the data from training (Steinke et al., 2024; Mahloujifar et al., 2025b), whereas Malek et al. (2021) modifies the training labels. These requirements severely restrict applications of these auditing methods. Instead, our label inference attack can run entirely post-training. MIAs can be stated as observational privacy games if the challenger is able to sample fresh samples from the distribution (Ye et al., 2022). However, obtaining new samples from the distribution (or from its close approximation), without affecting the training pipeline, remains an extremely challenging open problem (Meeus et al., 2025).

The closest observational auditing mechanism to ours is the recent work of Busa-Fekete et al. (2024), which measures the label reconstruction advantage of the adversary with and without access to the model. This metric is not translated into a lower bound on ϵ; in fact, such translation would be difficult because the adversary has a different baseline success (prior) for each sample. Additionally, this approach requires estimating probabilities of the mechanism's output given a particular label, limiting its applicability to simple mechanisms like randomized response and random label aggregation. In contrast, our label inference attack is observational and applicable to all mechanisms. It can audit label privacy in a statistically valid manner because it sets up a game where the baseline accuracy of the adversary is 0.5 for all samples.

3 PRELIMINARIES

Notation We use calligraphic letters such as $\mathcal{X}, \mathcal{Y}, \mathcal{D}$ to denote sets and distributions. Capital letters such as X, Y, D denote random variables and datasets, and lowercase letters denote their values. $\mathcal{X}^n$ is the set of all datasets of size n with elements from $\mathcal{X}$, whereas $\mathcal{X}^*$ is the set of all finite-size data sets with elements from $\mathcal{X}$. We use $\mathsf{Supp}(X)$ to denote the support of a random variable X.

In this work, we audit the simulation-based definition of differential privacy. It generalizes the traditional add/remove (or "leave-one-out", or "zero-out") notion of DP to support a privacy unit that is a subset of the sample's attributes. The definition compares the distribution of a mechanism M on dataset D with that of a simulator that emulates the output of M on D without seeing the protected attribute of a record. See Appendix A for further discussion of this definition.

Definition 1 (Simulation-based privacy for protected attributes (Mahloujifar et al., 2025a)). *Let records $(x, y) \in \mathcal{X} \times \mathcal{Y}$ be such that x is public or non-sensitive and thus need not be protected. We say that a randomized mechanism $M\colon (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Z}$ is (ϵ, δ)-Sim-DP with respect to a simulator $\mathrm{Sim}\colon (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \mathcal{Z}$ if for all datasets $D \in (\mathcal{X} \times \mathcal{Y})^*$, $(x, y) \in D$, and $D' = D \setminus \{(x, y)\}$ we have*

$$M(D) \approx_{\epsilon,\delta} \mathrm{Sim}(D', x) \tag{1}$$

For the more advanced notion of f-DP (Dong et al., 2020), a mechanism is f-Sim-DP if

$$M(D) \approx_{f} \mathrm{Sim}(D', x). \tag{2}$$

We also say M is (ϵ, δ)-SIM-DP (resp. f-SIM-DP) if there exists a simulator Sim for which (1) (resp. 2) holds.

Our auditing guarantees are stated for a family of generic simulators that treat M as a black-box. Such a simulator imputes the missing part y of the record based on the public part x and runs the original mechanism.

Definition 2 (Imputation-based simulator)**.** *For a mechanism $M\colon (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{Z}$, a data distribution $\mathcal{D}$ supported on $\mathcal{X}\times\mathcal{Y}$, a dataset $D \in (\mathcal{X}\times\mathcal{Y})^*$ and public part of a record $x \in \mathcal{X}$, the* imputation-based simulator $\mathrm{Sim}_{M,\mathcal{D}}$ *is defined as*

$$\mathrm{Sim}_{M,\mathcal{D}}(D, x) \triangleq M(D \cup \{(x, y')\}), \quad \textit{where } y' \sim \mathcal{D} \mid x.$$

Finally, we define empirical privacy auditing. A similar definition holds for (ϵ, δ)-Sim-DP.

Definition 3 (Auditing simulation-based DP)**.** *An audit procedure takes the description of a mechanism M, a trade-off function f, a simulator* Sim *and decides whether the mechanism satisfies f-*Sim*-DP. We define it as a two-step process.*

- $\mathrm{Game}\colon \mathcal{M}\times\mathcal{S}\to\mathcal{O}$. *The auditor runs a potentially randomized experiment/game using the description of mechanism $M \in \mathcal{M}$ and the simulator* Sim*. The auditor receives the game output $o \in \mathcal{O}$.*
- $\mathrm{Evaluate}\colon \mathcal{O}\times\mathcal{F}\to\{0,1\}$. *The output is 0 if the auditor rejects the hypothesis that M satisfies f-*Sim*-DP based on evidence o, and 1 otherwise.*

4 Observational versus Interventional Privacy Games

An observational privacy game considers the training dataset as a given. In contrast, an interventional privacy game interferes with the training data pipeline and the eventual dataset used for training. In this section, we formalize our observational privacy auditing framework. To that end, we introduce a generic attack game, generalizing Swanberg et al. (2025), as Algorithm 1.

The privacy game occurs between two parties: a challenger and an adversary. A key difference from Swanberg et al. (2025) is that we split the challenger algorithm into two stages: determining the training data ($G_{\text{intervention}}$) and determining the additional outputs provided to the adversary (G_{hint}). With this, we can separate observational games from interventional ones. Further we distinguish between the training dataset D and additional game artifacts S, which are used by the challenger to set up a stochastic game (e.g., sampling random bits).

Algorithm 1 Generic Attack Game (adapted from Swanberg et al. (2025))

Input Mechanism $M(\cdot)$, data distribution $\mathcal{D}$, distribution for game artifacts $\mathcal{D}_{\mathsf{prior}}$, adversary A

1: Sample training dataset $D = (x_1, \ldots, x_m)$ where $x_i \sim \mathcal{D}$.
2: Sample game artifacts $S \sim \mathcal{D}_{\mathsf{prior}}$.
3: Let $\overline{D} \leftarrow G_{\text{intervention}}(D, S)$. ▷ $G_{\text{intervention}}$ determines the training dataset for M
4: Let $o_1 \leftarrow M(\overline{D})$.
5: Let $o_2 \leftarrow G_{\text{hint}}(o_1, D, S)$. ▷ G_{hint} determines additional input provided to the attacker A
6: Run attack $A(o_1, o_2)$ with access to $\mathcal{D}$, $\mathcal{D}_{\texttt{prior}}$, M.
7: Measure adversary success with loss metric $\mathcal{L}(A(o_1, o_2), S)$.

The algorithm $G_{\text{intervention}}$ determines the training dataset for M, obtained from a potential modification of the fixed training set D. For instance, in the one-run MIA (Steinke et al., 2024) the artifacts are $S = (S_i \sim \mathrm{Uniform}(0,1)\colon i \in [m])$, where m is the number of canaries. The training dataset $\overline{D}$ is obtained from D by including all samples $x_i \in D$ where $S_i = 1$ and excluding all samples where $S_i = 0$.

The role of G_{hint} is to collect additional information the challenger provides to the adversary, based on game artifacts, training data, and the of the trained model $M(\overline{D})$). In the one-run MIA, o_2 is the vector of targets $x_1, \ldots, x_m$ from the training set D.

Definition 4 (Observational games)**.** *We call a privacy game, as outlined in Algorithm 1, observational if $\overline{D} = D$ (and as a result $o_1 = M(D)$). That is, M is trained on the original D, and the observations of the adversary consist of (1) output of $M(D)$ and (2) additional postprocessing of D, $M(D)$ according to the game artifacts S.*

An observational versus an interventional membership inference attack We have described how the (interventional) one-run MIA (Steinke et al. (2024)) can be framed as Algorithm 1. We now present an observational one-run MIA (following Ye et al. (2022)).

Given data distribution $\mathcal{D}$, sample a sequence S of game artifacts $S_i = (b_i, x_i') \sim \{0,1\} \times \mathcal{D}$ for $i \in [m]$. The bit b_i is sampled uniformly at random from $\{0,1\}$, whereas x_i is a fresh sample from the distribution (which is highly unlikely to be in D). Then, train model M on $\overline{D} = D$. Let $o_2[i] = x_i$ if $b_i = 0$ and $o_2[i] = x_i'$ if $b_i = 1$ for $i \in [m]$. That is, the adversary receives either a training sample x_i or a sample x_i' from the distribution with probability 0.5. The adversary has to guess b_i, i.e., which of the samples it is seeing. This game is observational because the training dataset for M remains unchanged.

From the adversary's perspective, the observational one-run MIA has the same distribution as the interventional MIA. For the challenger, the difference matters as the observational game does not alter the training pipeline. In this game, the source of counterfactual samples x_i' can be a distribution that approximates $\mathcal{D}$ sufficiently well.

5 Observational attribute inference

In this section, we describe our observational attribute inference attack. It allows privacy measurement with respect to any set of protected attributes, which can be the entire record (as in the observational MIA, Section 4) or just the label, for Label DP auditing. We provide theoretical results for obtaining empirical privacy lower bounds from our game. In particular, our analysis provides lower bounds on simulation-based DP in the add/remove privacy model (see Appendix A).

Algorithm 2 Observational attribute inference in one run

Input Oracle access to a mechanism $M(\cdot)$, data distribution $\mathcal{D}$ and approximate distribution $\mathcal{D}'$ supported on $\mathcal{X} \times \mathcal{Y}$, attacker A

1: Let $D^0 = \Big((x_1, y_1^0), \ldots, (x_m, y_m^0)\Big)$, where $(x_i, y_i^0) \sim \mathcal{D}$ for $i \in [m]$.
2: Run mechanism M on D^0 to get output o_1.
3: Sample game artifacts $\Big((b_1, y_1^1), \ldots, (b_m, y_m^1)\Big)$ such that $(b_i, y_i^1) \sim \mathsf{Bernouilli}(0.5) \times \mathcal{D}' \mid x_i$.
4: Construct a dataset $D^b = \Big((x_1, y_1^{b_1}), \ldots, (x_m, y_m^{b_m})\Big)$.
5: Run attack A with input $o_1 = M(D^0)$, $o_2 = D^b$, and access to $\mathcal{D}$, $\mathcal{D}'$.
6: Reconstruct a vector of predictions $b' = (b_1', \ldots, b_m')$ which is supported on $\{0, 1, \perp\}^m$.
7: Count c, the number of correct guesses where $b_i' = b_i$, and c', the total number of guesses where $b_i' \neq \perp$. $\triangleright$ $\perp$ indicates abstention from guessing
8: **return** (c, c').

Similar to prior auditing papers (Mahloujifar et al., 2025b; Steinke et al., 2024) the adversary can choose to abstain from guessing on samples where it is least confident, to boost its positive likelihood ratio. The observational game can use the entire dataset as canaries (i.e., $m = n$).

Obtaining approximate distributions. A key aspect in implementing the observational attribute inference attack is to produce the proxy distribution $\mathcal{D}'$ from which the counterfactual partial records y_i^1 are sampled. One option is to train an additional model M' to predict the missing attribute(s). For label inference attacks, which are a special case of Algorithm 2, we sample counterfactual label y_i^i from $\mathsf{Multinoulli}(M'(x_i))$. In an online machine learning system, where the model trains incrementally as more recent data becomes available, one can use a prior model checkpoint as the model M' (and run the attack on the newer data). This eliminates the need for training any additional models, making our proposed label inference attacks very lightweight in terms of computational overhead and implementation complexity.

Auditing guarantees when $\mathcal{D} = \mathcal{D}'$. We first establish auditing guarantees for the simpler case where $\mathcal{D} = \mathcal{D}'$, i.e., when we can sample counterfactual records from the ground truth distribution. Our bounds are stated for auditing f-DP, following an analogous argument in Mahloujifar et al. (2025b), which gives the tightest DP auditing analysis and can be translated to the language of (ϵ, δ). The auditing applies an evaluation procedure, Algorithm 3, described and analyzed in Appendix B.

Theorem 5 (Auditing f-DP with no distribution shift)**.** *Let $M\colon (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Z}$ be a mechanism, $\mathcal{D}$ the data distribution, and $\mathrm{Sim}_{M,\mathcal{D}}$ the imputation-based simulator (Definition 2). Let $C = \sum_{i \in [m]} \mathbf{1}[b_i' = b_i]$ be the total number of correct answers from the one-run observational attribute*

inference attack (Algorithm 2) for an adversary that makes c' guesses. If M is f-$\mathrm{Sim}_{M,\mathcal{D}}$-DP and Algorithm 3 returns True on $(c', c, M, f, \tau = 0.05)$, then $\Pr[C \geq c] \leq \tau$.

Auditing guarantees under distribution shift. Now we consider the more general case when $\mathcal{D} \neq \mathcal{D}'$. We state an information-theoretic bound, which depends on the total variation (TV) distance between $\mathcal{D}$ and $\mathcal{D}'$. Intuitively, the larger the distance between the two distributions, the weaker the lower bound we can obtain on the privacy guarantee. The value τ in Theorem 6 is an upper bound on the adversary's a priori (i.e., before having access to the target model) success probability in distinguishing whether a sample (x, y^b) is from $\mathcal{D}$ or $\mathcal{D}'$.

Theorem 6 can also be stated in terms of the adversary's ability to distinguish between $\mathcal{D}$ and $\mathcal{D}'$ given its resource constraints. Such a bound is particularly meaningful for Label-DP auditing, where we can assume that the adversary cannot distinguish $y \mid x$ from $\mathsf{Multinoulli}(M'(x))$ if M' is the best classifier on $\mathcal{D}$ available to the adversary.

We state here a bound akin to Steinke et al. (2024), that is easier to interpret in the case of distribution shift. We present a bound on auditing f-DP under distribution shift in Appendix B.

Theorem 6 (Auditing (ϵ, δ)-DP under distribution shift)**.** *Let $M\colon (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Z}$ be a mechanism, $\mathcal{D}$ the data distribution, $\mathcal{D}'$ the approximate distribution, and $\mathrm{Sim}_{M,\mathcal{D}'}$ the imputation-based simulator. Let $C = \sum_{i\in[m]} \mathbf{1}[b'_i = b_i]$ be the total number of correct answers from the one-run observational attribute inference attack (Algorithm 2). If $\mathcal{M}$ is (ϵ, δ)-$\mathrm{Sim}_{M,\mathcal{D}'}$-DP and $\mathrm{TV}(\mathcal{D}|x_i, \mathcal{D}'|x_i) \leq \tau$ for all $x_i \in D$, then*

$$\Pr_{S,D}[C \geq c \mid A(o_1, o_2) = b'] \leq \Pr_{\hat{C}_i \sim \mathsf{Bernoulli}(\beta)} \Big[\sum_{i\in[m]} \hat{C}_i \geq c \Big] + n\delta \tag{3}$$

where $\beta = \frac{e^\epsilon}{e^\epsilon + \frac{1-\tau}{1+\tau}}$.

A full theorem statement and its proof are deferred to Appendix C.

6 Experiments on auditing Label-DP

We validate our theoretical framework with experiments on two real representative datasets: CIFAR-10 (Krizhevsky, 2009) for image classification and Criteo (Criteo AI Lab, 2015) for tabular data with sensitive labels. For each dataset, we train classifiers using standard Label-DP mechanisms and empirically evaluate the privacy guarantees using our audit procedure. We also evaluate our Label DP auditing technique for Randomized Response (Warner, 1965) on synthetic data.

6.1 Label-DP Learning algorithms

The earliest approach to achieving Label DP is the Randomized Response (RR) mechanism (Warner, 1965). In RR, each training label is randomly replaced according to a fixed probability distribution before being shared with the learning algorithm. This randomization helps protect the privacy of individual labels. We briefly review several recent Label DP mechanisms that improve on RR.

Label Private One-Stage Training (LP-1ST, Ghazi et al. (2021)) Instead of using a fixed distribution as in RR, LP-1ST samples each training label y_i from a learned prior distribution $P(y|X_i)$. The prior can be estimated by observing the top-k predictions from a pretrained model (either in domain or out-of-domain), restricting RR to the most probable labels. Alternatively, the training can be split into multiple stages, where an earlier model provides the prior for the next (LP-MST).

Private Aggregation of Teacher Ensembles with FixMatch (PATE-FM, Malek et al. (2021)) PATE-FM combines the FixMatch semi-supervised learning algorithm (Sohn et al., 2020) with private aggregation. Multiple teacher models are trained, each using all features but only a unique, disjoint subset of the labels. The predictions from these teachers are then aggregated in a differentially private manner using the PATE framework (Papernot et al., 2017) to train a student model.

Additive Laplace with Iterative Bayesian Inference (ALIBI, Malek et al. (2021)) ALIBI achieves Label DP by adding Laplace noise to the one-hot encoded labels (Ghosh et al., 2012), making the released labels differentially private. Bayesian inference is applied to the noisy labels to obtain the most likely original (discrete) label, as differential privacy is preserved under post-processing.

6.2 Attack implementation and adversarial strategy

A key ingredient in implementing our label inference attack is generating the reconstructed label y^1 given features x (see Algorithm 2). We generate y^1 from the predictions of a reference model M', trained on separate data (but from approximately the same distribution) as the target model M. Specifically, $y^1 \sim \mathsf{Multinoulli}(M'(x))$, where $M'(x)$ are the predictions of M' on x for each class.

For the Criteo dataset, where data is collected over 28 consecutive days, we train M' on Day 0 data and the target model on Day 1 data. While there may be some distributional shift between Day 0 and Day 1 data, we assume this is small, and that the adversary cannot distinguish between the true labels y^0 and the reconstructed labels y^1 (before having access to the model).

For CIFAR-10 experiments we randomly split the training data ($n = 50$K samples) into two. We train M' on the first half, and the target model M on the second half.

We run the attack on $m = 200$K canaries for Criteo and $m = 10$K canaries for CIFAR-10. For the MIA experiments we use the same number of canaries for the non-members (taken from the test set).

The adversary obtains its guesses by computing per-example scores. Let x be the features and $y^b \in \{y^0, y^1\}$ be the label received by the adversary. The adversary computes a score that correlates with whether y^b is the reconstructed or the training label. The score consists of two components. The first component $s_1(x, y^b)$ is the difference in probabilities that y^b came from the training set versus the reconstructed distribution:

$$\begin{aligned} s_1(x, y^b) &= \Pr[y^0 = y^b \mid M(x)] - \Pr[y^1 = y^b \mid M'(x)] \\ &= M(x)[y^b] - M'(x)[y^b], \end{aligned}$$

where $M(x)[y^b]$ is the prediction of M on x for class y^b. Since the adversary's performance is measured at the tails of the score distribution, the adversary prefers to guess on samples where $\Pr[y^0 \neq y_1]$ is high. Thus the second component of the score is defined as

$$s_2(x, y^b) = \Pr[y^0 \neq y_1] = 1 - M'(x)[y^b].$$

The final score combines the two components as $s(x, y^b) = s_1(x, y^b) \cdot s_2(x, y^b)^t$ with a hyperparameter $t \geq 0$ that allows for weighting the two components separately. We use $t = 2$, as $t > 1$ gives tight lower bounds for RR. The adversary guesses on $c'\%$ of samples with highest absolute scores. We sweep $c' \in \{1, 2, \ldots, 100\}$ and report the highest ϵ achieved at 95% confidence, averaged over 100 repetitions of the game (resamplings of counterfactual labels).

6.3 CIFAR-10 Experiments

For CIFAR-10, we treat the image classes as sensitive labels. We train standard convolutional neural networks with varying privacy budgets $\epsilon \in \{1.0, 10.0, \infty\}$ (see model accuracy in Table 4, Appendix E). We then audit Label DP using our observational game and report results in Table 1.

Table 1: CIFAR-10. Auditing Label-DP algorithms under different ϵ with $\delta = 10^{-5}$.

Label-DP Algorithm	CIFAR-10		
	$\epsilon = \infty$	$\epsilon = 10.0$	$\epsilon = 1.0$
LP-1ST	$2.13 \pm .22$	$2.02 \pm .33$	$0.43 \pm .05$
LP-1ST (out-of-domain prior)	$2.26 \pm .22$	$1.86 \pm .25$	$0.90 \pm .07$
PATE-FM	$2.42 \pm .32$	$2.22 \pm .24$	$0.79 \pm .09$
ALIBI	$2.53 \pm .33$	$2.18 \pm .27$	$0.67 \pm .07$

6.4 Criteo Experiments

The Criteo dataset contains user click-through data with demographic information encoded as 13 numerical features and 26 categorical features. A binary label indicates whether the user clicked on the ad. The distribution of the labels is highly imbalanced, with only 3% of positives (clicks). The

Table 2: Criteo. Auditing Label-DP algorithms under different ϵ with $\delta = 10^{-5}$. Similarly Wu et al. (2023), for LP-1ST (domain prior) at $\epsilon \in \{0.1, 1, 2\}$ the training process did not produce meaningful outcomes.

Label-DP Algorithm	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 4$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.1$
LP-1ST	$1.37 \pm .14$	$1.29 \pm .12$	$1.22 \pm .17$	$0.59 \pm .05$	$0.34 \pm .02$	$0.06 \pm .01$
LP-1ST (domain prior)	$1.47 \pm .22$	$1.37 \pm .10$	$1.28 \pm .21$	—	—	—
LP-1ST (noise correction)	$1.31 \pm .12$	$1.26 \pm .11$	$1.04 \pm .16$	$0.52 \pm .07$	$0.40 \pm .08$	$0.06 \pm .01$
LP-2ST	$1.52 \pm .14$	$1.46 \pm .11$	$1.24 \pm .15$	$0.75 \pm .05$	$0.61 \pm .07$	$0.06 \pm .01$
PATE	$1.60 \pm .15$	$1.48 \pm .14$	$1.28 \pm .12$	$0.71 \pm .07$	$0.59 \pm .05$	$0.06 \pm .01$

overall dataset contains over 4 billion click log data points over a period of 24 days. We followed the same setup in Wu et al. (2023) where 1 million data points are selected for each day. We divide the data into 80% for training, 4% for validation, and 16% for testing. Model performance is evaluated using the log-loss metric on the test set.

We train gradient boosting decision trees with the CatBoost library (Prokhorenkova et al., 2018) with varying privacy budgets $\epsilon \in \{0.1, 1.0, 2, 0, 4.0, 8.0, \infty\}$ and evaluate label privacy using observational privacy auditing (Table 2). Table 5 (in Appendix E) shows model performance under different Label-DP algorithms and privacy budgets ϵ.

6.5 Comparison with Existing Methods

We compare our observational auditing approach against traditional canary-based methods to demonstrate the effectiveness and practicality of our framework. More specifically, we evaluate against the lightweight *difficulty calibration* MIA in Watson et al. (2022), where membership scores are adjusted to the difficulty of correctly classifying the target sample. For each canary datapoint, we set the calibrated membership scores as the difference in the loss between the target model and the reference model M'. Fig. 1 shows how our method achieves similar auditing results when compared to MIA on the CIFAR-10 and Criteo datasets.

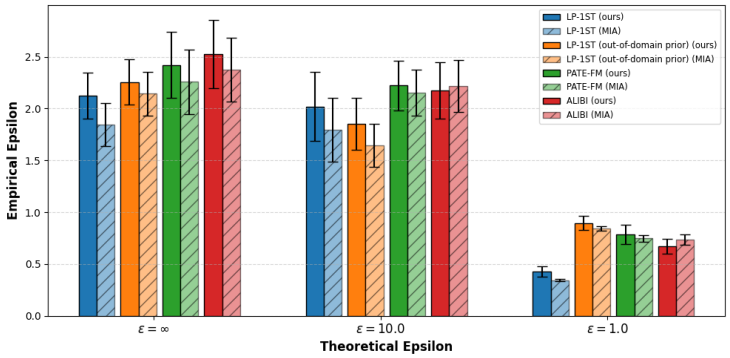
We leave for future work a comparison with more computationally intensive methods, such as Zarifzadeh et al. (2024), which require training multiple auxiliary (shadow) models to achieve state-of-the-art attack performances.

6.6 Synthetic data and auditing randomized response

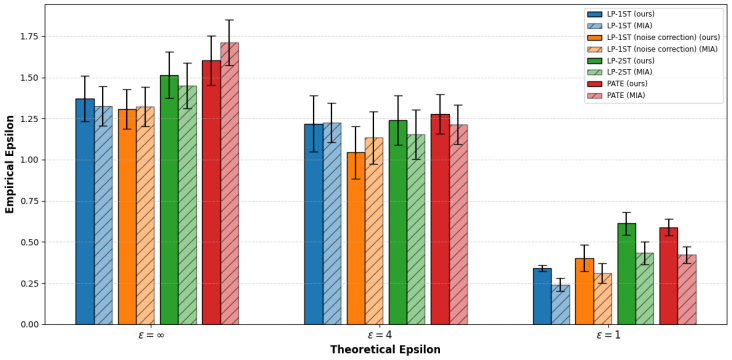
We empirically demonstrate the tightness of our Label-DP auditing algorithm for Randomized Response using synthetic data. The distribution consists of k balanced classes, each generated from a 5-dimensional Gaussian with the same covariance but a shifted mean. We generate $n = 10^6$ samples and experiment with $k \in \{2, 5, 10\}$. Counterfactual labels y^1 are generated using either the true distribution or the predictions of a logistic regression model. Figure 2 shows empirical epsilon lower bounds at 95% confidence when the adversary makes 0.1% non-abstaining guesses. We obtain tight lower bounds for $\epsilon \in [1, 4]$. At lower epsilons the audit overestimates privacy loss due to a higher variance induced by a small number of guesses. Using more guesses at lower epsilon fixes the issue (Appendix D). Obtaining tight lower bounds at very high epsilon is a limitation of current auditing methodology (Steinke et al., 2024; Mahloujifar et al., 2025b).

7 Discussion

In this paper, we establish a framework for auditing privacy without any intervention during the training process. This enables a principled privacy evaluation in settings where the training process is outside the control of the privacy auditor. This may sound counterintuitive as privacy auditing is a form of causal analysis. However, our method can provide provable guarantees on auditing performance under certain assumptions about the data distribution. We envision that our framework will broaden the scope of privacy auditing applications, as it does not require any supervision of the training process and can be conducted by third parties.



(a) CIFAR-10



(b) Criteo

Figure 1: Comparison with MIA for different Label-DP Algorithms on CIFAR-10 and Criteo datasets. The error bar represents the standard deviation across 100 different repetitions.

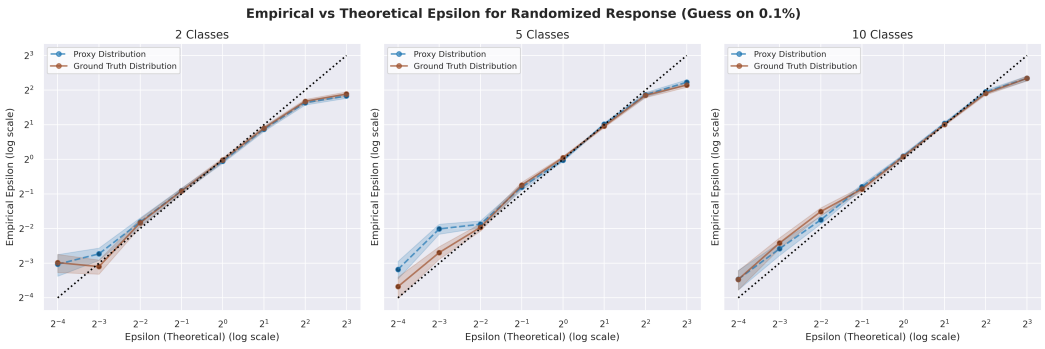


Figure 2: Auditing randomized response when the adversary guesses on 0.1% of samples. The counterfactual labels are generated either from the ground-truth distribution or a proxy distribution obtained from the predictions of logistic regression model.

REFERENCES

Nitin Agrawal and Laura Book. Quantifying reidentification risk for machine learning models. In *2025 USENIX Conference on Privacy Engineering Practice and Respect (PEPR '25)*, Santa Clara, CA, USA, June 2025. URL https://www.usenix.org/conference/pepr25/presentation/agrawal.

Michael Barbaro and Tom Zeller Jr. A face is exposed for AOL searcher No. 4417749. *The New York Times*, August 2006.

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

Róbert Busa-Fekete, Travis Dick, Claudio Gentile, Andrés Muñoz Medina, Adam Smith, and Marika Swanberg. Auditing privacy mechanisms via label inference attacks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 82823–82862, 2024.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Nicolas Papernot. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security '21)*, pp. 2633–2650, 2021.

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

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 155–186. JMLR Workshop and Conference Proceedings, 2011.

Criteo AI Lab. Criteo 1TB Click Logs Dataset. https://ailab.criteo.com/download-criteo-1tb-click-logs-dataset/, 2015.

Zeyu Ding, Yuxin Wang, Guanhong Wang, Danfeng Zhang, and Daniel Kifer. Detecting violations of differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, pp. 475–489, 2018.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5770–5781, 2020.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, pp. 265–284, 2006.

Ferdinando Fioretto and Pascal Van Hentenryck. *Differential Privacy in Artificial Intelligence: From Theory to Practice*. now publishers, Boston-Delft, 2025. doi: 10.1561/9781638284772. URL http://dx.doi.org/10.1561/9781638284772.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, 2021.

Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally optimal privacy mechanisms for minimax agents. In *Proceedings of the 2012 ACM Symposium on Theory of Computing*, pp. 351–360, 2012.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1569–1582, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, 2009.

Kalev Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1605–1622, 2020.

Saeed Mahloujifar, Chuan Guo, G. Edward Suh, and Kamalika Chaudhuri. Machine learning with privacy for protected attributes. In Marina Blanton, William Enck, and Cristina Nita-Rotaru (eds.), *IEEE Symposium on Security and Privacy, S&P 2025, San Francisco, CA, USA, May 12-15, 2025*, pp. 2640–2657. IEEE, 2025a.

Saeed Mahloujifar, Luca Melis, and Kamalika Chaudhuri. Auditing f-differential privacy in one run. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=OZSXYeqpI1.

Mani Malek, Ilya Mironov, Karthik Prasad, Igor Shilov, and Florian Tramèr. Antipodes of label differential privacy: PATE and ALIBI. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, 2021.

Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. SoK: Membership inference attacks on LLMs are rushing nowhere (and how to fix it). In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 385–401, 2025.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy (S&P)*, pp. 111–125. IEEE, 2008.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (S&P)*, pp. 739–753. IEEE, 2019.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *Proceedings of the 32nd USENIX Conference on Security Symposium*. USENIX Association, 2023.

Milad Nasr, Thomas Steinke, Borja Balle, Christopher A. Choquette-Choo, Arun Ganesh, Matthew Jagielski, Jamie Hayes, Abhradeep Guha Thakurta, Adam Smith, and Andreas Terzis. The last iterate advantage: Empirical auditing and principled heuristic analysis of differentially private SGD. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DwqoBkj2Mw.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *5th International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkwoSDPgg.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 6639–6649, 2018.

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.

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. In *Network and Distributed System Security Symposium (NDSS)*, 2019.

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

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Congzheng Song and Vitaly Shmatikov. Membership inference attacks against collaborative filtering algorithms. In *2019 IEEE Symposium on Security and Privacy (S&P)*, pp. 457–474. IEEE, 2019.

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

Marika Swanberg, Meenatchi Sundaram Muthu Selva Annamalai, Jamie Hayes, Borja Balle, and Adam D. Smith. Beyond the worst case: Extending differential privacy guarantees to realistic adversaries. *CoRR*, abs/2507.08158, 2025. doi: 10.48550/ARXIV.2507.08158. URL https://doi.org/10.48550/arXiv.2507.08158.

Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Jack Watson, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Difficulty calibration and membership inference attacks. In *International Conference on Learning Representations (ICLR)*, 2022.

Ruihan Wu, Jin Peng Zhou, Kilian Q. Weinberger, and Chuan Guo. Does label differential privacy prevent label inference attacks? In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTAT)*, volume 206, pp. 4336–4347. PMLR, April 2023.

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.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 58244–58282. PMLR, July 2024.

A SIMULATION-BASED DIFFERENTIAL PRIVACY

In this section, we motivate and formally define the simulation-based notion of differential privacy. We recall the standard definition of differential privacy, which we are going to extend and adapt to our setting.

Definition 7 (Differential privacy)**.** *A randomized mechanism $M\colon \mathcal{X}^* \to \mathcal{Z}$ satisfies (ϵ, δ)-differential privacy if for all pairs of neighboring $D, D' \subset \mathcal{X}^*$ and $E \subset \mathcal{Z}$ the following holds:*

$$Pr[M(D) \in E] \leq e^{\epsilon} Pr[M(D') \in E] + \delta \approx_{\epsilon,\delta} M(D'),$$

The definition depends on the notion of *neighboring datasets*, which is a symmetric binary relation on $\mathcal{X}^*$ denoted as $D \sim D'$. Choosing the neighboring relationship is an important part of mechanism design and has direct implications on the type of privacy guarantee. See Table 3 for some common definitions of neighboring datasets and the resulting DP notions.

Differentially private mechanisms are often applied to datasets containing records with both public and private components. Consider records of the form $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where x represents the public (non-sensitive) data and y the sensitive attributes that require protection. Differential privacy for this setting can be defined by letting $D \sim D'$ if they differ only in the sensitive attributes of a single record: that is, D and D' are identical except for one record being (x, y) in D and (x, y') in D'. This is a generalization of the original definition of Dwork et al. (2006), which modeled D as an indexed vector. In current terminology, this is the replacement model of differential privacy: the sensitive portion of a record is *replaced* with a different value. Semantically, this guarantees that an adversary observing M's output cannot distinguish between two possible private values y and y' of a user any better than without M, within an (ϵ, δ)-slack.

The alternative to the replacement model is add/remove, which stipulates that M's output on inputs with and without the user are (ϵ, δ)-indistinguishable. In addition to protecting the user's data, this model also hides the user's membership status and the size of the dataset. DP in the add/remove model implies DP in the replacement model (via the two-step hybrid, with looser parameters) but not vice versa.

We target the add/remove model of differential privacy. For datasets containing both public and private attributes, one way to define this model is by introducing a class of records where the sensitive parts are removed. In this formulation, two datasets D and D' are neighboring if they differ only in the pairs (x, y) and $(x, \perp)$. We propose an equivalent definition that is more explicit, as it introduces the notion of a *simulator*:

Definition (more formal version of Definition 1, Simulation-based privacy for protected attributes)**.** *Let records $(x, y) \in \mathcal{X} \times \mathcal{Y}$ be such that x is public or non-sensitive and thus need not be protected. We say that a randomized mechanism $M\colon (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Z}$ is (ϵ, δ)-Sim-DP with respect to a simulator $\mathrm{Sim}\colon (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \mathcal{Z}$ if for all datasets $D \in (\mathcal{X} \times \mathcal{Y})^*$, $(x, y) \in D$, $D' = D \setminus \{(x, y)\}$, and $E \in \mathsf{Supp}(M(\cdot))$ we have*

$$\Pr[M(D) \in E] \leq e^{\epsilon} \Pr[\mathrm{Sim}(D', x) \in E] + \delta.$$

We can also define the more advanced notion of f-differential privacy, where we call a mechanism f-Sim-DP if

$$\Pr[M(D) \in E] \leq f(\Pr[\mathrm{Sim}(D', x) \in E] + \delta.)$$

Semantically, (ϵ, δ)-SIM-DP means that anything that can be inferred about (x, y) from $M(D)$ could also be inferred without ever exposing y to M, within the standard (ϵ, δ)-bounds.

Note that f-DP generalizes (ϵ, δ)-DP by allowing a more complex relation between the probability distributions of $M(D)$ and $M(D')$. The following proposition shows how one can express approximate DP as an instantiation of f-DP.

Proposition 8 (Dong et al. (2020))**.** *A mechanism is (ϵ, δ)- Sim-DP if it is f-Sim-DP for a function f such that $1 - f(x) = e^{\epsilon} \cdot x + \delta$.*

Whenever we say that a mechanism satisfies f-(Sim)-DP, we implicitly imply that f is a valid trade-off function. That is, f is defined on the domain $[0, 1]$ and has a range of $[0, 1]$. Moreover, f is decreasing and convex with $f(x) \leq 1 - x$ for all $x \in [0, 1]$. This is without loss of generality. That

DP definition	Difference between $D \sim D'$
Replacement, DP	x, x'
Add/remove, DP	x
Replacement, Label DP	(x, y), (x, y')
Add/remove, Label DP	(x, y), $(x, \perp)$

Table 3: The difference between which two neighboring datasets D, D' under various DP definitions.

is, if a mechanism is f-DP for a an arbitrary function$f : [0, 1] \to [0, 1]$, then it is also f' -(Sim)-DP for valid trade-off function f' with $f'(x) \leq f(x)$ for all $x \in [0, 1]$ (See Proposition 2.2 in Dong et al. (2020)).

B Auditing Guarantees for f-differential privacy

In this section, we prove Theorem 5 which provides guarantees similar to Mahloujifar et al. (2025b) for auditing f-SIM-DP with our observational attribute inference attack.

Before stating the main theorem, we describe how the accuracy of the adversary from the observational attribute inference attack can be translated into a lower bound on f-SIM-DP.

Definition 9 (Obtaining empirical epsilon from f-SIM-DP auditing)**.** *Let* $(\mathrm{Game}, \mathrm{Evaluate})$ *be an audit procedure. The empirical privacy of a mechanism* M *for a family* F *of trade-off functions and a simulator* Sim *is the random variable distributed according to the output of the following process:*

1. *Obtain observation* $o \leftarrow \mathrm{Game}(M, \mathrm{Sim})$.
2. *Construct* $F_o = \mathrm{maximal}\{f \in F : \mathrm{Evaluate}(o, f) = 1\}$*, where the partial order on* F *is defined as* $f \prec g$ *iff* $f(x) \leq g(x)$ *for all* $x \in [0, 1]$.
3. *Compute*

$$\epsilon(\delta) = \min_{f \in F_o} \max_{x \in [0,1]} \log\left(\frac{1 - f(x) - \delta}{x}\right).$$

The empirical lower bound $\epsilon(\delta)$ is a random variable since it is a function of the output o of a randomized process Game. The point estimate of $\epsilon(\delta)$ is the lowest ϵ given δ guaranteed by an f-SIM-DP *not* rejected by the auditing procedure.

In Algorithm 3 we show how to audit a particular trade-off function f given number of non-abstaining guesses c' and number of correct guesses c. (This is the Evaluate function in Step 2 of Definition 9). The choice of a family of trade-off functions in Definition 9 should be based on the expectations of the true privacy curve. For example, if one expects the privacy curve of a mechanism to be similar to that of a Gaussian mechanism, then they would choose the set of all trade-off functions imposed by a Gaussian mechanism as the family. This is the choice we use in our experiments.

Finally, we state and prove a more general version of Theorem 5 that allows for the case when we sample counterfactual partial records from a proxy distribution $\mathcal{D}'$ different from $\mathcal{D}$.

Theorem 10 (Auditing f-DP with distribution shift)**.** *Let* $M : (\mathcal{X}, \mathcal{Y})^* \to \mathcal{Z}$ *be a mechanism,* $\mathcal{D}$ *the data distribution,* $\mathcal{D}'$ *an approximate distribution, and* $\mathrm{Sim}_{M,\mathcal{D}'}$ *the imputation-based simulator (Definition 2). Let* $C = \sum_{i \in [m]} \mathbf{1}[b'_i = b_i]$ *be the total number of correct answers from the one-run observational attribute inference attack (Algorithm 2) for an adversary that makes* c' *guesses. Let* $TV(\mathcal{D}|x, \mathcal{D}'|x) \leq \tau$ *for all* x *in the dataset* D *and define* $g : [0, 1] \to [0, 1]$ *such that* $g(s) = f(\min(1, s + \tau))$*. If* M *is* f-$\mathrm{Sim}_{M,\mathcal{D}'}$*-DP and Algorithm 3 returns True on* $(c', c, M, g, \gamma = 0.05)$*, then* $\Pr[C \geq c] \leq \gamma$.

Proof. The proof follows similarly to the proof of Theorem 3.2 in Mahloujifar et al. (2025b). We only need to prove a similar Lemma to that of their Lemma A.1 that is adapted to our setting of simulation based differential privacy with distribution shift.

Lemma 11. *Let* $M : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Z}$ *be a mechanism,* $\mathcal{D}$ *a distribution on* $\mathcal{X} \times \mathcal{Y}$ *and* $\mathrm{Sim}_{M,\mathcal{D}'}$ *the imputation-based simulator (Definition 2). Assume* $TV(\mathcal{D}, \mathcal{D}') \leq \tau$*. If* M *is* f-$\mathrm{Sim}_{M,\mathcal{D}'}$*-DP, then*

Algorithm 3 Iteratively deciding an upper bound probability of making more than c correct guesses (Mahloujifar et al., 2025b)

Input description of trade-off function f, number of guesses c', number of correct guesses c, number of samples m, probability threshold γ (default is $\gamma = 0.05$).

$\forall 0 \leq i \leq c$ set $h[i] = 0$, and $r[i] = 0$.
set $r[c] = \gamma \cdot \frac{c}{m}$.
set $h[c] = \gamma \cdot \frac{c'-c}{m}$.
for $i \in [c-1, \ldots, 0]$ **do**
  $h[i] = \bar{f}^{-1}\big(r[i+1]\big)$ ▷ $\overline{f}(x) = 1 - f(x)$
  $r[i] = r[i+1] + \frac{i}{c'-i} \cdot \big(h[i] - h[i+1]\big)$.
end for
if $r[0] + h[0] \geq \frac{c'}{m}$ **then**
  Return True; (Probability of c correct guesses (out of c') is less than γ)
else
  Return False; (Probability of having c correct guesses (out of c') could be more than γ).
end if

for any attack algorithm A and event E we have

$$f''_\tau(\Pr[M(D) \in E]) \leq \Pr[M(D) \in E \ \ \& \ \ b_1 = b'_1] \leq f'_\tau(\Pr[M(D) \in E]),$$

where

$$f'_\tau(x) = \sup\{t \in [0,s]; t + f(s - t + \tau) \leq 1\} \quad \textit{and} \quad f''_\tau(s) = \inf\{t \in [0,1]; f(t) + s - t \leq 1 - \tau\}$$

Proof. Fix a sample (x_1, y_1^0) with counterfactual label y_1^1. For simplicity, we drop the sample index 1 and use x, y^0, y^1. Let D denote the dataset containing (x, y^0) and let D' be a dataset obtained from D by replacing y^0 with y^1. We assume b'_1 is a deterministic function of $M(D)$ and y^{b_1}. Let $p = \Pr[M(D) \in E \ \ \& \ \ b_1 = b'_1]$ and $q = \Pr[M(D) \in E]$. We have

$$\begin{aligned}
p &= \Pr[M(D) \in E \ \ \& \ \ b_1 = b'_1] \\
&= \Pr[M(D) \in E \ \ \& \ \ b_1 = 1 \ \ \& \ \ b'_1 = 1] \\
&\quad + \Pr[M(D) \in E \ \ \& \ \ b_1 = 0 \ \ \& \ \ b'_1 = 0] \\
&= \mathop{\mathrm{E}}_{y^1, y^0, b_1, \theta \sim M(D)}[I(\theta \in E \ \ \& \ \ b_1 = 1 \ \ \& \ \ b'_1(\theta, y^1) = 1)] \\
&\quad + \mathop{\mathrm{E}}_{y^1, y^0, b_1, \theta \sim M(D)}[I(\theta \in E \ \ \& \ \ b_1 = 0 \ \ \& \ \ b'_1(\theta, y^0) = 0)] \\
&= 0.5 \cdot \mathop{\mathrm{E}}_{y^1, y^0, b_1, \theta \sim M(D)}[I(\theta \in E \ \ \& \ \ b'_1(\theta, y^1) = 1) \mid b_1 = 1] \\
&\quad + 0.5 \cdot \mathop{\mathrm{E}}_{y^1, y^0, b_1, \theta \sim M(D)}[I(\theta \in E \ \ \& \ \ b'_1(\theta, y^0) = 0) \mid b_1 = 0] \\
&= 0.5 \cdot \mathop{\mathrm{E}}_{y^1, y^0, b_1, \theta \sim M(D), \theta' \sim M(D')}[I(\theta \in E \ \ \& \ \ b'_1(\theta, y^1) = 1) \mid b_1 = 1] \\
&\quad + 0.5 \cdot \mathop{\mathrm{E}}_{y^1, y^0, b_1, \theta \sim M(D), \theta' \sim M(D')}[I(\theta \in E \ \ \& \ \ b'_1(\theta, y^0) = 0) \mid b_1 = 0] \\
&\leq 0.5 \cdot \Big(1 - f(\mathop{\mathrm{E}}_{y^1, y^0, b_1, \theta \sim M(D), \theta' \sim M(D')}[I(\theta' \in E \ \ \& \ \ b'_1(\theta', y^1) = 1) \mid b_1 = 1])\Big) \\
&\quad + 0.5 \cdot \Big(1 - f(\mathop{\mathrm{E}}_{y^1, y^0, b_1, \theta \sim M(D), \theta' \sim M(D')}[I(\theta' \in E \ \ \& \ \ b'_1(\theta', y^0) = 0) \mid b_1 = 0])\Big) \\
&\leq 1 - f(\mathop{\mathrm{E}}_{y^1, y^0, b_1, \theta \sim M(D), \theta' \sim M(D')}[I(\theta' \in E \ \ \& \ \ b_1 \neq b'_1(\theta', y_1^{1-b_1})]) \quad \text{(By convexity of } f\text{.)} \\
&\leq 1 - f(q - p + \tau). \quad \text{(By the fact that } f \text{ is decreasing.)}
\end{aligned}$$

This implies $p + f(q - p + \tau) \leq 1$ which in turn implies $p \leq f'(q)$. Similarly, for the other side we repeat the argument up until the last 3 steps. That is

$$
\begin{aligned}
q &= 0.5 \cdot \mathop{\mathrm{E}}_{y^1,y^0,b_1,\theta\sim M(D),\theta'\sim M(D')}[I(\theta \in E \ \ \& \ \ b_1'(\theta, y^1) = 1) \mid b_1 = 1] \\
&\quad + 0.5 \cdot \mathop{\mathrm{E}}_{y^1,y^0,b_1,\theta\sim M(D),\theta'\sim M(D')}[I(\theta \in E \ \ \& \ \ b_1'(\theta, y^0) = 0) \mid b_1 = 0] \\
&\geq 0.5 \cdot \Big(f^{-1}\Big(1 - \mathop{\mathrm{E}}_{y^1,y^0,b_1,\theta\sim M(D),\theta'\sim M(D')}[I(\theta' \in E \ \ \& \ \ b_1'(\theta', y^1) = 1) \mid b_1 = 1]\Big)\Big) \\
&\quad + 0.5 \cdot \Big(f^{-1}\Big(1 - \mathop{\mathrm{E}}_{y^1,y^0,b_1,\theta\sim M(D),\theta'\sim M(D')}[I(\theta' \in E \ \ \& \ \ b_1'(\theta', y^0) = 0) \mid b_1 = 0]\Big)\Big) \\
&\geq f^{-1}(1 - \mathop{\mathrm{E}}_{y^1,y^0,b_1,\theta\sim M(D),\theta'\sim M(D')}[I(\theta' \in E \ \ \& \ \ b_1 \neq b_1'])) \quad \text{(By convexity of } f\text{.)} \\
&= f^{-1}(1 - q + p + \tau) \quad \text{(By the fact that } f^{-1} \text{ is decreasing.).}
\end{aligned}
$$

This implies $p \geq f^{-1}(1 - q + p)$ which in turn implies $f(p) + q - p + \tau \leq 1$. So we have $p \geq f''(q)$. □

We can now plug Lemma 11 in the proof of Theorem 3.2 in Mahloujifar et al. (2025b). Note that g is a valid trade-off function and hence f'_τ has all the desired properties proved in Proposition A.2. Therefore, we can adopt the same proof to finish the proof of our Theorem. □

C AUDITING GUARANTEES FOR (ϵ, δ)-DIFFERENTIAL PRIVACY

In this section, we prove similar guarantees as in Steinke et al. (2024); Swanberg et al. (2025), which can be translated into empirical lower bounds on (ϵ, δ)-SIM-DP. The proofs are very similar to the two prior works, thus to avoid repetition, we provide a proof only for the more simple $(\epsilon, 0)$-differential privacy case.

We first state Theorem 12, which an analytical version of Theorem 6, and yields better lower bounds than Theorem 6.

Theorem 12 (Auditing (ϵ, δ)-SIM-DP)**.** *Let $M : (\mathcal{X}, \mathcal{Y})^* \to \mathcal{Z}$ be a mechanism, $\mathcal{D}$ the data distribution, $\mathcal{D}'$ the approximate distribution, and $\mathrm{Sim}_{M,\mathcal{D}'}$ the imputation-based simulator (Definition 2). Let $C = \sum_{i\in[m]} \mathbf{1}[b'_i = b_i]$ be the total number of correct answers from the one-run observational attribute inference attack (Algorithm 2). If M is (ϵ, δ)-$\mathrm{Sim}_{M,\mathcal{D}'}$-DP and $TV(\mathcal{D}|x_i, \mathcal{D}'|x_i) \leq \tau$ for all $x_i \in D$, then*

$$
\Pr_{S,D}[C \geq c \mid A(o_1, o_2) = b'] \leq \Pr_{\hat{C}_i \sim \mathsf{Bernoulli}(\beta)} \Big[\sum_{i\in[m]} \hat{C}_i \geq c\Big] + \alpha \cdot m \cdot \delta, \tag{4}
$$

where $\beta = \frac{e^\epsilon}{e^\epsilon + \frac{1-\tau}{1+\tau}}$ and

$$
\alpha = \max\Big\{\frac{1}{j}\Big(\Big[\sum_{i\in[m]} \hat{C}_i \geq c - j\Big] - \Big[\sum_{i\in[m]} \hat{C}_i \geq c\Big]\Big) : j \in [m]\Big\}.
$$

Then, we state and prove the version of Theorem 6 for pure-SIM-DP.

Theorem 13 (Auditing $(\epsilon, 0)$-SIM-DP)**.** *Let $M : (\mathcal{X}, \mathcal{Y})^* \to \mathcal{Z}$ be a mechanism, $\mathcal{D}$ the data distribution, $\mathcal{D}'$ the approximate distribution, and $\mathrm{Sim}_{M,\mathcal{D}'}$ the imputation-based simulator (Definition 2). Let $C = \sum_{i\in[m]} \mathbf{1}[b'_i = b_i]$ be the total number of correct answers from the one-run observational attribute inference attack (Algorithm 2). If M is $(\epsilon, 0)$-$\mathrm{Sim}_{M,\mathcal{D}'}$-DP and $TV(\mathcal{D}|x_i, \mathcal{D}'|x_i) \leq \tau$ for all $x_i \in D$, then*

$$
\Pr_{S,D}[C \geq c \mid A(o_1, o_2) = b'] \leq \Pr_{\hat{C}_i \sim \mathsf{Bernoulli}(\beta)} \Big[\sum_{i\in[m]} \hat{C}_i \geq c\Big], \tag{5}
$$

where $\beta = \frac{e^\epsilon}{e^\epsilon + \frac{1-\tau}{1+\tau}}$.

Proof. The proof is Bayesian and proceeds iteratively. Conditioned on the output of the game (both challenger and adversarial output), we determine the probability of a successful guess at step i. Let g be a function which determines the adversary's answer $b'_i \in \{0,1\}$ given a sample $(x_i, y_i^{b_i})$, where $b_i \in \{0,1\}$, $y_i^0 \sim \mathcal{D}|x$, and $y_i^1 \sim \mathcal{D}'|x$. The adversary has knowledge of $\mathcal{D}$ and $\mathcal{D}'$. We can assume w.l.o.g. that g is deterministic. Note that if $D \equiv D'$ then $\Pr[g(x_i, y_i^{b_i}) = b_i] = \frac{1}{2}$, where randomness is over the sampling of y_i^0, y_i^1, and b. In general,

$$\Pr[g(x_i, y_i^b) = b] = \frac{1}{2} + \frac{1}{2} TV(\mathcal{D}|x_i, \mathcal{D}'|x_i) \leq \frac{\tau + 1}{2}, \tag{6}$$

based on the success of the Bayes optimal classifier.

Let $D_{<i}$ be the dataset consisting of the samples $\{(x_j, y_i^{b_j})\}$ for $j < i$. We condition on the value of $D_{<i}$ and abuse notation by using $D_{<i}$ as a random variable and as a fixed value.

Applying Bayes' theorem and the law of total probability we obtain

$$\Pr_{y_i^0, y_i^1}[g(x_i, y_i^0) = 0 \mid A(o_1, o_2) = b', D_{<i}] = \frac{1}{1 + \frac{\Pr[A(o_1,o_2)=b' | g(x_i,y_i^1)=0, D_{<i}] \cdot \Pr[g(x_i,y_i^1)=0 | D_{<i}]}{\Pr[A(o_1,o_2)=b' | g(x_i,y_i^0)=0, D_{<i}] \cdot \Pr[g(x_i,y_i^0)=0 | D_{<i}]}}.$$

From Equation 6,

$$\frac{\Pr[g(x_i, y_i^1) = 0 \mid D_{<i}]}{\Pr[g(x_i, y_i^0) = 0 \mid D_{<i}]} = \frac{\Pr[g(x_i, y_i^1) = 0]}{\Pr[g(x_i, y_i^0) = 1]} \geq \frac{1 - \tau}{1 + \tau}.$$

Finally, since M is $(\epsilon, 0)$-$\mathrm{Sim}_{M,\mathcal{D}'}$-DP then A is also $(\epsilon, 0)$-$\mathrm{Sim}_{M,\mathcal{D}'}$-DP as it post-processes M. From the definition of simulation-based DP, we obtain

$$\Pr_{y_i^0, y_i^1}[g(x_i, y_i^0) = 0 \mid A(o_1, o_2) = b', D_{<i}] \leq \frac{e^\epsilon}{e^\epsilon + \frac{1-\tau}{1+\tau}}.$$

As a result,

$$\Pr_{y_i^0, y_i^1, b_i}[b_i = b'_i \mid A(o_1, o_2) = b', D_{<i}] \leq \frac{e^\epsilon}{e^\epsilon + \frac{1-\tau}{1+\tau}}.$$

We now prove the result by induction, using the concept of stochastic dominance (Definition 4.8 of Steinke et al. (2024)). We assume inductively that $C_{m-1} = \sum_{i=1}^{m-1} \mathbf{1}[b'_i = b_i]$ is stochastically dominated by $\sum_{i=1}^{m-1} \hat{C}_i$, where $\hat{C}_i \sim \mathsf{Bernoulli}(\beta)$. We have obtained that, conditioned on C_{m-1}, the variable $\mathbf{1}[b'_i = b_i]$ is stochastically dominated by $\mathsf{Bernoulli}(\beta)$. Applying Lemma 4.9 from Steinke et al. (2024) concludes the proof. □

To obtain an empirical epsilon lower bound from Theorems 12 and 13, we fix δ and choose a desired confidence $\gamma < 1$ (e.g., $\gamma = 0.05$ and $\delta = 10^{-5}$ for our experiments). Then, we find an $\epsilon \geq 0$ so that the left hand side of the inequality in Theorems 12 and 13 is equal to γ. This value of ϵ is our lower bound.

D Auditing Randomized Response

We empirically demonstrate the tightness of our Label-DP auditing algorithm for Randomized Response on synthetic data sampled from a mixture of Gaussian distributions.

We experiment with binary and multi-class labels. For binary labels, y is sampled from $\mathsf{Bernoulli}(p)$. For multi-class labels y are sampled from a generalized Bernoulli distribution with equal probabilities for each class. Given a label y, the features x are sampled as:

$$x \mid y \sim \mathcal{N}(\mu_y, I_d),$$

where I_d is the d-dimensional identity matrix, and $\mu_y = e_y$, where e_y is a d-dimensional index vector that is 1 at index y.

The output of the randomized response mechanism are the noisy labels. Recall that y_i^0 is the training set label and y_i^1 is the reconstructed label. We generate y_i^1 in two ways:

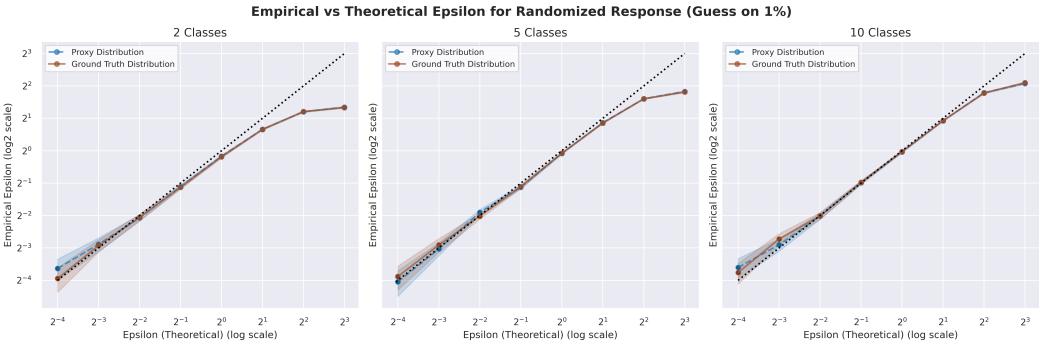


Figure 3: Auditing randomized response when the adversary guesses on 1% of samples. The counterfactual labels are generated either from the ground-truth distribution or a proxy distribution obtained from the predictions of logistic regression model.

- Ground truth: By using the posterior $\mathcal{D} \mid x$ where $\mathcal{D}$ is the data-generating distribution.
- Proxy: By using the predictions of a Logistic Regression model trained on fresh data from the distribution (with default hyper-parameters from sciki-learn).

This way, we compare the effect of using a proxy distribution instead of the ground truth for generating the labels y_i^1. In Figures 2 and 3, we see that the two methods for generating y^1 give similar results, as Logistic Regression can approximate a mixture of Gaussians quite well.

We use $n = 10^6$, $d = 5$, and the number of classes in $\{2, 5, 10\}$. We run our attack on the output of Randomized Response and with a fraction of non-abstaining guesses in $\{0.1\%, 1\%\}$. We use Theorem 13 to obtain the empirical epsilon achieved at 95% confidence. For each dataset and RR output, we repeat the game 10^2 times (with a fresh vector of reconstructed labels) and compute the average and standard deviation of the empirical epsilon.

While Figure 2 shows that with 0.1% adversarial guesses we overestimate the privacy loss given by the theoretical epsilon, the lower bound is tight with 1% guesses in the low-epsilon regime. However, the lower bounds are not as tight in the medium epsilon regime $[1, 4]$. These experiments show that choosing the number of adversarial guesses should take into consideration the privacy regime we are targeting.

E Accuracy of Models trained on Criteo and CIFAR-10

Tables 4 and 5 show models performance on the test set respectively on CIFAR 10 and Criteo.

Table 4: CIFAR-10. Model test accuracy of label-DP models under different ε.

Label-DP Algorithm	CIFAR-10		
	$\varepsilon = \infty$	$\varepsilon = 10.0$	$\varepsilon = 1.0$
LP-1ST	91.3	91.07	60.4
LP-1ST (out-of-domain prior)	92.1	91.5	87.9
PATE-FM	92.5	92.3	91.3
ALIBI	90.1	87.1	66.9

Table 5: Criteo. Log-loss of Label-DP algorithms on the test set under different ε.

Label-DP Algorithm	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 4$	$\varepsilon = 2$	$\varepsilon = 1$	$\varepsilon = 0.1$
LP-1ST	0.130	0.130	0.136	0.206	0.362	0.653
LP-1ST (domain prior)	0.130	0.130	0.136	-	-	-
LP-1ST (noise correction)	0.130	0.130	0.131	0.156	0.171	0.645
LP-2ST	0.130	0.130	0.123	0.207	0.342	0.527
PATE	0.130	0.151	0.156	0.188	0.255	0.680