Causal Evaluation of Membership Inference Attacks

Mathieu Even
Inria PreMeDICaL, Inserm, Montpellier, France
mathieu.even@inria.fr

Clément Berenfeld
Inria PreMeDICaL, Inserm, Montpellier, France

Linus Bleistein
School of Computer and Communication Science (EPFL)
School of Life Sciences (EPFL)
Lausanne, Switzerland

Tudor Cebere
Inria PreMeDICaL, Inserm, Montpellier, France

Julie Josse
Inria PreMeDICaL, Inserm, Montpellier, France

Aurélien Bellet
Inria PreMeDICaL, Inserm, Montpellier, France
aurelien.bellet@inria.fr

Abstract

Membership Inference Attacks (MIAs) aim to distinguish training points (members) from unseen data (non-members), and are widely used to quantify memorization and assess privacy risks. Standard MIA evaluation requires repeated retraining, which is computationally costly for large models. One-run (single training with randomized data inclusion) and zero-run (post hoc evaluation) methods are often used instead, but their statistical validity remains unclear. We address this gap by framing MIA evaluation as a causal inference problem, defining *memorization as the causal effect of including a data point in the training set*. This novel formulation reveals and formalizes key sources of bias in existing protocols: one-run methods suffer from interference between jointly included points, while zero-run evaluations are additionally confounded by distribution shift between member and non-member evaluation data. We derive causal analogues of standard MIA metrics and propose practical estimators for multi-run, one-run, and zero-run regimes with non-asymptotic consistency guarantees. We validate our approach in several settings, including pre-trained and fine-tuned LLMs, showing that it enables reliable measurement of MIA performance without retraining and under distribution shift. Overall, our framework provides a principled foundation for privacy evaluation in modern AI systems.

1 Introduction

The rapid deployment of large-scale machine learning models has intensified concerns regarding data privacy and intellectual property, as such models are susceptible to privacy attacks that can infer sensitive information about individual training data points [68, 9, 10, 52, 2, 24]. Consequently, data owners and regulators increasingly question whether unauthorized information has been added to training sets, potentially exposing personal or copyrighted content. Recent guidance from data protection authorities emphasizes the use of attack-based evaluations to assess privacy leakage and to determine whether a trained model may itself constitute personal data under existing regulatory frameworks [19], while recent high-profile legal challenges concerning data provenance and usage have raised similar questions regarding whether specific data was used in model training [54, 14].

In this context, *Membership Inference Attacks* (MIAs) offer a compelling conceptual and practical tool, as they probe the minimal form of information a model can retain about a data point: whether it was part of the training set [11]. MIAs provide a principled measure of memorization and unintended privacy leakage, revealing when a model's internal representations encode training-specific artifacts that distinguish members from non-members, even when direct data leakage is rare or difficult to

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

observe. MIAs also play a key role in auditing differential privacy [31, 53] and enabling other privacy attacks, including data reconstruction [10, 9, 52, 2] and dataset inference [43, 44].

The standard framework for evaluating MIAs, the *multi-run* regime, involves retraining models hundreds or thousands of times with and without a given point [68, 81, 11, 83]. While statistically rigorous, this approach is computationally prohibitive for modern large-scale models, prompting a shift toward more efficient alternatives. In *one-run* approaches, a single model is trained on a randomized subset of data [72, 42, 41], whereas *zero-run* methods evaluate a deployed model post hoc [80, 66, 4, 49]. The latter scenario is particularly relevant when an external evaluator cannot directly intervene in model training, which requires provider cooperation and trust, and cannot independently retrain the model due to withheld information about the training data or algorithm, as typically the case when using MIAs to assess the privacy risks of LLMs deployed by dominant tech companies.

However, when moving from controlled retraining to more practical settings, the statistical validity of MIA evaluation becomes unclear. In particular, recent work shows that zero-run MIAs on LLMs suffer from *systematic bias caused by distribution shift* in the non-member data used for evaluation [17, 84, 50]. For example, in typical MIA evaluations, *members* are documents published before the LLM training cutoff date—e.g., arXiv papers or Wikipedia articles that were almost certainly included in training—while *non-members* are documents published afterward. These sets often differ in topic, style, and temporal context, causing MIA success in membership prediction to reflect distributional differences rather than true memorization, thereby overestimating privacy leakage [50].

Contributions. In this work, we reframe the evaluation of MIAs as a causal inference problem, *defining memorization as the causal effect of including a data point in the training set*. This shift from a predictive to a causal perspective allows us to explicitly identify the assumptions required to interpret MIA performance as true memorization rather than an artifact of confounding factors, and enables the design of robust estimators that correct systematic biases in existing approaches. Specifically, our contributions are as follows:

1. **Causal formalization of MIA evaluation**: We introduce a causal evaluation framework for MIAs that isolates the effect of a training point on the model, accounting for non-causal signals that can distort memorization estimates. Accordingly, we define causal counterparts to standard MIA metrics, shifting evaluation from correlation to causation.
2. **A causal taxonomy**: We analyze multi-run, one-run, and zero-run MIA evaluations from a causal lens, formalizing their error sources: *interference* [29] between jointly inserted points in one-run settings, and *confounding* [61] from non-random assignments in zero-run scenarios.
3. **Principled estimators:** We propose practical estimators for causal MIA metrics that correct for distribution shift bias and come with non-asymptotic consistency guarantees. As a byproduct, we obtain a new approach to causal effect estimation under random interference, a challenging setting [12], leveraging learning-theoretic tools such as algorithmic stability [7] in a novel way.
4. **Empirical validation:** Through experiments on synthetic data, ResNet on CIFAR-10, and both pretrained and fine-tuned LLMs, we show that our causal estimators provide practical and reliable solutions for post-hoc MIA evaluation, effectively correcting evaluation biases observed in prior work on large language models.

2 Background

Here and throughout, a training algorithm $\mathcal{A}$ is a (potentially randomized) function that maps a dataset $\mathcal{D} \subset \mathcal{X}$ to model parameters $\theta = \mathcal{A}(\mathcal{D})$ in a parameter space $\Theta \subset \mathbb{R}^D$.

2.1 Membership Inference Attacks (MIAs)

An MIA aims to determine whether a specific record $x \in \mathcal{X}$ was included in a model's training set. Typical MIAs use a scoring function $s_{\mathrm{MIA}}(x, \theta) \in \mathbb{R}$ (e.g., negative loss or log-likelihood of θ on x) to compute a *membership confidence score*, which is thresholded to predict membership. **In this work, our focus is not on designing MIAs** (i.e., constructing a suitable s_{MIA}), **but on the validity of the three evaluation regimes described in Figure 1**. These regimes differ in how they collect the evidence $(X_i, A_i, Y_i)_{i\in[n]}$, where $X_i \in \mathcal{X}$ is a data point, $A_i \in \{0, 1\}$ is its membership in the training set, and Y_i is its MIA score. Evaluating MIAs requires specifying the distribution over which memorization is measured; throughout the paper, we denote this target distribution on $\mathcal{X}$ by $\mathcal{P}_T$.

Algorithm 1 Multi-run

Input: algorithm $\mathcal{A}$, dataset $\mathcal{D}$, target distribution $\mathcal{P}_T$, number of samples n
for $i \in [n]$ **do**
  Sample $X_i \sim \mathcal{P}_T$
  $A_i \sim \text{Bernoulli}(1/2)$
  $\mathcal{D}_{\text{train,i}} \leftarrow \begin{cases} \mathcal{D} \cup \{X_i\} & \text{if } A_i = 1, \\ \mathcal{D} & \text{otherwise.} \end{cases}$
  $\theta_i \leftarrow \mathcal{A}(\mathcal{D}_{\text{train,i}})$
  $Y_i \leftarrow s_{\text{MIA}}(X_i, \theta_i)$
Output: $(X_i, A_i, Y_i)_{i \in [n]}$

Algorithm 2 One-run

Input: algorithm $\mathcal{A}$, dataset $\mathcal{D}$, target distribution $\mathcal{P}_T$, number of samples n
for $i \in [n]$ **do**
  Sample $X_i \sim \mathcal{P}_T$
  $A_i \sim \text{Bernoulli}(1/2)$
$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D} \cup \{X_i : A_i = 1\}$
$\theta \leftarrow \mathcal{A}(\mathcal{D}_{\text{train}})$
for $i \in [n]$ **do**
  $Y_i \leftarrow s_{\text{MIA}}(X_i, \theta)$
Output: $(X_i, A_i, Y_i)_{i \in [n]}$

Algorithm 3 Zero-run

Input: trained model $\theta \leftarrow \mathcal{A}(\mathcal{D}_{\text{train}})$, datasets $\mathcal{D}_T \subset \mathcal{D}_{\text{train}}$ (of underlying distribution $\mathcal{P}_T$) and $\mathcal{D}_0 \subset \mathcal{X} \backslash \mathcal{D}_{\text{train}}$
$(X_i)_{i \in [n]} \leftarrow \mathcal{D}_T \cup \mathcal{D}_0$
for $i \in [n]$ **do**
  $A_i \leftarrow \mathbb{1}_{\{X_i \in \mathcal{D}_T\}}$
  $Y_i \leftarrow s_{\text{MIA}}(X_i, \theta)$
Output: $(X_i, A_i, Y_i)_{i \in [n]}$

Figure 1: MIA evaluation regimes studied in this work, with the MIA abstracted by its score function s_{MIA}. Green highlights the key steps of the multi-run setting; red indicates the successive modifications introduced when moving to the one-run and zero-run settings.

In the **Multi-run (Algorithm 1)** regime, each sample (X_i, A_i, Y_i) is generated from an independently trained model θ_i, with $X_i \sim \mathcal{P}_T$ and randomized membership A_i. In the **One-run (Algorithm 2)** regime introduced by Steinke et al. [72], a single model θ is trained, with n points drawn from $\mathcal{P}_T$ randomly included in the training set $\mathcal{D}_{\text{train}}$. Finally, the **Zero-run (Algorithm 3)** regime, widely used in recent studies on LLMs [66, 48, 49], evaluates MIAs post-hoc on a fixed model, without retraining or randomized membership. The training set $\mathcal{D}_{\text{train}}$ is fixed and unknown to the evaluator. Evaluation uses proxy datasets: $\mathcal{D}_T \subset \mathcal{D}_{\text{train}}$ (known members, e.g., pre-2023 Wikipedia) and $\mathcal{D}_0 \subset \mathcal{X} \backslash \mathcal{D}_{\text{train}}$ (known non-members, e.g., post-release articles).

Evaluation metrics. The evidence $(X_i, A_i, Y_i)_{i \in [n]}$ is used to compute standard MIA performance metrics, measuring memorization by how well the scores Y_i separate members ($A_i = 1$) from non-members ($A_i = 0$): **(i)** *Membership advantage*: $\mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0]$; **(ii)** *AUC*: $\mathbb{P}(Y_i > Y_j \mid A_i = 1, A_j = 0)$; **(iii)** *TPR at fixed FPR*: $\mathbb{P}(Y_i \geqslant t_\alpha \mid A_i = 1)$ with t_α the $(1 - \alpha)$-quantile of $Y_i \mid A_i = 0$.

2.2 Causal Inference

In causal inference [26, 62, 75], observations are tuples (X_i, A_i, Y_i), where X_i denotes covariates, $A_i \in \{0, 1\}$ is a treatment indicator, and Y_i is here a real-valued outcome. Under the potential outcomes framework [71, 30, 75], each unit has two counterfactual outcomes $Y_i(0)$ and $Y_i(1)$, but only $Y_i = Y_i(A_i)$ is observed. Individual treatment effects $Y_i(1) - Y_i(0)$ are thus not identifiable, meaning they cannot be uniquely determined from the distribution of the observed data. Causal inference instead focuses on population-level quantities such as the average treatment effect (ATE) $\mathbb{E}[Y_i(1) - Y_i(0)]$, which are identifiable under the following standard assumptions.

Assumption 2.1 (Standard causal assumptions)**.** For all $i \in [n]$, *(1) (SUTVA)* $Y_i = Y_i(A_i)$; the assignment mechanism satisfies either *(2a) (Randomized)* $A_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$ or *(2b) (Unconfoundedness)* $A_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i$; and *(3) (Overlap)* $0 < \mathbb{P}(A_i = 1 | X_i) < 1$ almost surely.

SUTVA and overlap imply a unit's outcome does not depend on other units' assignments, and that each unit has a nonzero probability of receiving either treatment. The assumption on the assignment mechanism depends on the setting.

RCT vs observational study. In randomized control trials (RCTs), treatment is assigned randomly to each unit, so the *Randomized* assumption holds. In observational studies, treatment may correlate with covariates that also affect outcomes, violating this assumption. *Unconfoundedness* addresses this by requiring that, conditional on observed covariates, treatment assignment is as good as random.

Causal inference with interference. When the SUTVA assumption is violated, *interference* occurs: a unit's outcome may depend on other units' treatment assignments, for instance through shared resources, competition effects, or herd immunity in vaccine trials [70]. Causal inference with interference thus requires additional formalism [29]. Specifically, interference is modeled by generalized potential outcomes $Y_i(\bar{a})$ indexed by the full assignment vector $\bar{a} \in \{0, 1\}^n$, with

observed outcome $Y_i = Y_i(\bar{A})$ for $\bar{A} = (A_j)_{j\in[n]}$. We define the marginal potential outcome as:

$$Y_i(a) = Y_i(A_1, \ldots, A_{i-1}, a, A_{i+1}, \ldots, A_n)\,. \tag{1}$$

In the absence of interference, $Y_i(\bar{a}) = Y_i(a_i)$, recovering the standard setting. Standard causal assumptions under interference [55] that generalize Assumption 2.1 are detailed in Section B.1.

3 MIA Evaluation Under a Causal Lens

Our first key contribution is to frame membership inference as a causal question: *what is the effect of including a given data point in the training set on the learned model?* In this formulation, the *treatment* A_i is the inclusion of a data point X_i in the training set, and the *outcome* Y_i is the membership confidence score (which depends on both the model and the data point X_i). Potential outcomes $Y_i(1)$ and $Y_i(0)$ correspond to the model's output on X_i under two counterfactual worlds: one in which X_i was used for training and one in which it was not. Since only one of these worlds is observed, individual effects are unidentifiable, but population-level causal effects can be estimated. Casting membership inference in this causal framework clarifies what is being measured and enables principled control of interference and confounding that would otherwise bias the evaluation.

3.1 Causal Interpretation of MIA Evaluation Regimes

We now show that we can explicitly map the three scenarios corresponding to Algorithms 1 to 3 to standard causal inference settings, with their corresponding causal graphs shown in Figure 2.

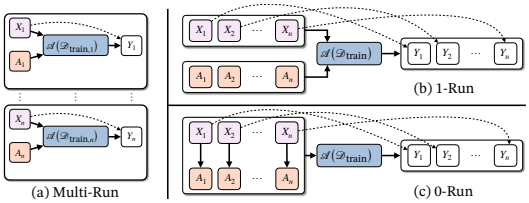


Figure 2: Causal graphs.

Multi-run ↔ RCT. The treatment is randomized, (X_i, A_i, Y_i) are i.i.d., corresponding to a RCT setting in causal inference where Assumption 2.1 with randomized assignments holds.

One-run ↔ RCT with interference. While the randomized assignment property is preserved, the one-run setting trades computational efficiency (training a single model instead of n) for the presence of *interference* between units. Indeed, the assignment A_j of point $j \in [n]$ can affect Y_i for $i \neq j$, since including or excluding X_j changes the learned model used to compute all outcomes. The one-run regime thus corresponds to an RCT with interference. In the causal graph shown in Figure 2b), interference arises through the shared mediator $\mathcal{A}(\mathcal{D}_{\text{train}})$. Without this mediator, the graph would contain direct edges $A_i \to Y_j$ for all $i, j \in [n]$, corresponding to a complete interference graph. While causal quantities are identifiable under standard assumptions (Assumption B.1) that here hold by design, additional assumptions will be required to ensure the consistency of our estimators, as we detail in Section 5.

Zero-run ↔ observational study with interference. This setting inherits interference from the one-run scenario and adds *confounding*. As shown by the edge from X_i to A_i in Figure 2c), membership is not assigned independently of the data points due to the distribution shift between members and non-members. This confounding factor, which is at the heart of misleading conclusions observed in memorization studies for LLMs [50], turns the problem into an *observational study*. Framing the problem causally makes this shift explicit and will enable principled debiasing of evaluation metrics via confounder adjustment, as we describe in Section 6.

3.2 Causal MIA Evaluation Metrics

When members ($A = 1$) and non-members ($A = 0$) follow different distributions, the evaluation metrics defined in Section 2.1 may conflate true memorization effects (i.e., differences in Y_i versus Y_j for $A_i \neq A_j$ even when $X_i \approx X_j$) with distribution shift effects, where differences in Y arise simply because X_i and X_j differ. To address this, we introduce causal metrics that enable meaningful MIA evaluation under both interference and distribution shift. Table 1 summarizes these metrics alongside their corresponding classical counterparts, where $\mathbb{E}_{\mathcal{P}_T}$ and $\mathbb{P}_{\mathcal{P}_T}$ denote expectations and probabilities over $X_i \sim \mathcal{P}_T$, where $\mathcal{P}_T$ is the target distribution. Under interference, all causal quantities implicitly depend on the dataset size n, as can be seen from (1).

Table 1: Classical MIA metrics and their causal counterparts.

Classical metrics	Causal metrics
Membership Advantage $\mathbb{E}[Y_i \mid A_i{=}1] - \mathbb{E}[Y_i \mid A_i{=}0]$	Average Treatment Effect $\tau_{\mathrm{ATE}} = \mathbb{E}_{\mathcal{P}_T}[Y_i(1) - Y_i(0)]$
$\mathrm{TPR}(t){=}\mathbb{P}(Y_i > t \mid A_i{=}1)$ $\mathrm{FPR}(t){=}\mathbb{P}(Y_i > t \mid A_i{=}0)$	$\tau_{\mathrm{TPR}}(t){=}\mathbb{P}_{\mathcal{P}_T}(Y_i(1) > t)$ $\tau_{\mathrm{FPR}}(t){=}\mathbb{P}_{\mathcal{P}_T}(Y_i(0) > t)$
AUC Area under $\{\mathrm{FPR}(t), \mathrm{TPR}(t)\}$	Causal AUC Area under $\{\tau_{\mathrm{FPR}}(t), \tau_{\mathrm{TPR}}(t)\}$
TPR at fixed FPR α $\mathbb{P}(Y_i \geqslant t_\alpha \mid A_i{=}1)$ t_α: $1{-}\alpha$ quantile of $Y_i \mid A_i{=}0$	Causal TPR at fixed FPR α $\tau_{\mathrm{TPR@FPR}}(\alpha) = \mathbb{P}_{\mathcal{P}_T}(Y_i(1) \geqslant q_\alpha)$ q_α: $1{-}\alpha$ quantile of $Y_i(0) \mid X_i \sim \mathcal{P}_T$

In the next sections, for simplicity of presentation, we will focus on estimating the **Average Treatment Effect (ATE)** $\tau_{\mathrm{ATE}} = \mathbb{E}_{\mathcal{P}_T}[Y_i(1) - Y_i(0)]$, which admits a direct interpretation as a *causal membership advantage*. Extensions of our results to the other causal metrics are provided in Section D.

Insight # 1

Our causal framework provides a principled way to characterize biases in MIA evaluation (interference and confounding) and to define performance metrics that disentangle distribution shift between member and non-member data from genuine memorization.

4 Consistency in the Multi-Run Regime

In the multi-run setting, identification follows from Assumption 2.1, which is satisfied by design.

Proposition 4.1. *If $(X_i, A_i, Y_i)_{i \in [n]}$ are obtained using Algorithm 1, causal evaluation metrics (Table 1, right column) are identifiable and equal to classical evaluation metrics (Table 1, left).*

These properties then enable estimation of the causal evaluation metrics. Focusing on the ATE (the causal membership advantage) τ_{ATE}, letting $n_a = \#\{i : A_i = a\}$, we define the empirical estimator:

$$\hat{\tau}_{\mathrm{ATE}} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i \,. \tag{2}$$

Proposition 4.2. *If $(X_i, A_i, Y_i)_{i \in [n]}$ are obtained using Algorithm 1, $\hat{\tau}_{\mathrm{ATE}}$ is a consistent estimator of τ_{ATE} and if outcomes Y_i lie in $[0,1]$,[1] with probability $1{-}4e^{-t}$: $|\hat{\tau}_{\mathrm{ATE}} {-} \tau_{\mathrm{ATE}}| \leqslant \sqrt{2t/n_1} {+} \sqrt{2t/n_0}$.*

5 Handling Interference in One-Run

We now consider the one-run setting, in which SUTVA in Assumption 2.1 is violated due to *interference*, so identification is no longer straightforward. We remind that we work with generalized counterfactuals: for $\bar{a} \in \{0,1\}^n$, $Y_i(\bar{a})$ is the outcome of unit $i \in [n]$ we would have observed had treatment assignments been $A_j = a_j$ for all $j \in [n]$, thereby explicitly modeling interference between data points. A generalization of Assumption 2.1 to settings with interference is provided in Section B.1 and holds by design, allowing us to generalize Proposition 4.1 to the one-run setting.

Proposition 5.1. *If $(X_i, A_i, Y_i)_{i \in [n]}$ are obtained using Algorithm 2, causal evaluation metrics (Table 1, right) are identifiable and equal to classic evaluation metrics (Table 1, left).*

We consider the same estimators as in the multi-run setting, such as $\hat{\tau}_{\mathrm{ATE}}$ defined in Equation (2) for the ATE. Consistency, however, is no longer guaranteed, since interference induces dependence between samples. As aptly noted by Cinelli et al. [12, Section 3.2], "all no-interference applications are alike; each SUTVA violation violates SUTVA in its own way", underscoring the absence of a one-size-fits-all solution for handling interference. In our setting, as illustrated in Figure 2b, interference arises because a single model is trained jointly on all members, inducing a complete interference graph over data points. We thus propose to impose assumptions directly on the training algorithm $\mathcal{A}$, since

[1] Our results readily extend to outcomes in a bounded space of diameter B, via a rescaling of the results by B.

the interference structure is intrinsic and unavoidable. To that end, we draw on algorithmic stability notions from learning theory [7], which formalize the limited sensitivity of a model's behavior to any single training point. In particular, error stability (Def. 5.2) bounds changes in expected loss, while uniform training stability (Def. 5.3) controls the worst-case change in loss at any *training* point.

Definition 5.2 (Error stability [7])**.** A learning algorithm $\mathcal{A}$ is α-error stable with respect to a loss $\mathcal{L} : \Theta \times \mathcal{X} \to \mathbb{R}_{\geqslant 0}$ if for all datasets $\mathcal{D} = \{x_1, \ldots, x_n\}$, all $i \in [n]$ and $\mathcal{D}'$ such that $\mathcal{D}' = \{x'_1, \ldots, x'_n\}$ with $x_j = x'_j$ if $j \neq i$, for $X \sim \mathcal{P}_T$ independent of the rest, we have:

$$|\mathbb{E}_{\mathcal{A}}[\mathbb{E}_X[\mathcal{L}(\mathcal{A}(\mathcal{D}), X)]] - \mathbb{E}_{\mathcal{A}}[\mathbb{E}_X[\mathcal{L}(\mathcal{A}(\mathcal{D}'), X)]]| \leqslant \alpha\,.$$

Definition 5.3 (β-uniform training stability)**.** A learning algorithm $\mathcal{A}$ is β-uniform training stable with respect to a loss $\mathcal{L}$ if for all datasets $\mathcal{D} = \{x_1, \ldots, x_n\}$, all $i \in [n]$ and $\mathcal{D}'$ such that $\mathcal{D}' = \{x'_1, \ldots, x'_n\}$ with $x_j = x'_j$ if $j \neq i$, for any $k \in [n] \backslash \{i\}$, we have a.s. over the randomness of $\mathcal{A}$:

$$|\mathcal{L}(\mathcal{A}(\mathcal{D}), x_k) - \mathcal{L}(\mathcal{A}(\mathcal{D}'), x_k)| \leqslant \beta\,.$$

Definition 5.3 holds for uniformly stable algorithms with β equal to the stability parameter. Interpolating models have $\beta = 0$, and near-interpolating models (with small training loss) have small β.

Theorem 5.4. *Assume that $(X_i, A_i, Y_i)_{i\in[n]}$ are obtained using Algorithm 2, with $Y_i = -\mathcal{L}(\theta, X_i)$ corresponding to the standard loss-based attack. Assume that $\mathcal{A}$ is α-error stable (Definition 5.2), β-uniform training stable (Definition 5.3), and that outcomes Y_i are in $[0, 1]$ almost surely. Then, with probability $1 - Ce^{-t}$ for some numerical constant $C > 0$:*

$$|\hat{\tau}_{\mathrm{ATE}} - \tau_{\mathrm{ATE}}| = \mathcal{O}\big(\sqrt{t/n} + \sqrt{nt\alpha^2} + \sqrt{nt\beta^2}\big)\,.$$

In the perfect interpolation regime ($\beta = 0$), the first term $\sqrt{t/n}$ is expected and tends to 0. For consistency, we thus need the second term $\sqrt{nt\alpha^2}$ to also be negligible, which is the case as long as the stability parameter $\alpha = o(1/\sqrt{n})$. For regularized convex or strongly convex problems, this will always hold since then α is of order $1/N$ where N is the total number of training points in $\mathcal{D}_{\mathrm{train}}$, leading to $\sqrt{nt\alpha^2} = \mathcal{O}(1/\sqrt{N})$ where $N \geqslant n$. In most applications, n will be negligible in front of N, so the second term becomes negligible in front of the first. For nonconvex problems, error stability no longer holds in general, but for SGD training with smooth, Lipschitz losses and early-stopping, error stability holds with α of order $N^{-1+o(1)}$ [23]. The term $\alpha\sqrt{n}$ will thus be of order $1/\sqrt{N}$ if $n = \mathcal{O}(N^{1-o(1)})$. In the general case, β is also required to be small enough. This is the case for large overparametrized and near-interpolating networks, or for uniformly stable algorithms. We also provide a negative result for the underparametrized regime (i.e., $\frac{D}{n} \to 0$) in Section C.

Remark 5.5 (Extension to other MIAs)**.** Consistency results in this section are specific to loss-based MIAs, as they rely on standard learning-theoretic assumptions (Definitions 5.2 and 5.3). Theorem 5.4 extends to other MIAs by assuming stability of the confidence score function s_{MIA}.

Insight # 2

In the one-run regime, consistency of MIA metrics can be retained under stability assumptions on the training algorithm, which serves as the mediator of interference between data points.

6 Adjusting for Confounders in Zero-Run

In the zero-run regime, the presence of confounders introduces an additional challenge for identifying and estimating causal evaluation metrics, due to the distributional shift between member data ($\mathcal{D}_T$) and non-member data ($\mathcal{D}_0$). Indeed, the problem departs from a RCT and becomes an observational study, since $\mathcal{P}_{X_i|A_i=1} \neq \mathcal{P}_{X_i|A_i=0}$. As is standard in causal inference, observational studies cannot be conducted without assumptions: the identifiability of causal estimands (such as τ_{ATE}) depends on them. We thus assume that any member would have had a non-zero probability of being non-member.

Assumption 6.1 (Domain overlap)**.** $\mathcal{D}_T$ and $\mathcal{D}_0$ consist of independent samples from distributions $\mathcal{P}_T$ and $\mathcal{P}_0$, respectively, with $\frac{\mathrm{d}\mathcal{P}_T}{\mathrm{d}\mathcal{P}_0}(x) < \infty$ for all $x \in \mathrm{Supp}(\mathcal{P}_T)$.

Under this assumption, *overlap* holds (Assumption 2.1) together with identifiability. Note that under Assumption 6.1 we have $\mathcal{P}_T = \mathcal{P}_{X_i|A_i=1}$ and $\mathcal{P}_0 = \mathcal{P}_{X_i|A_i=0}$.

Proposition 6.2. *Under Assumption 6.1, if $(X_i, A_i, Y_i)_{i\in[n]}$ are obtained according to Algorithm 3, the causal evaluation metrics are identifiable. Furthermore, $\tau_{\mathrm{ATE}} = \frac{\mathbb{E}[A_i Y_i]}{\mathbb{P}(A_i=1)} - \mathbb{E}\big[\frac{\pi(X_i)(1-A_i)}{1-\pi(X_i)} Y_i\big]\mathbb{P}(A_i=1)^{-1}$, where $\pi(x) = \mathbb{P}(A_i = 1 | X_i = x)$, $\forall x \in \mathcal{X}$.*

Proposition 6.2 shows that under distribution shift, the left and right columns of Table 1 no longer coincide. Consequently, estimators of standard evaluation metrics (left column) yield biased assessments of MIA performance. In particular, $\hat{\tau}_{\mathrm{ATE}}$ in (2) estimates $\mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0]$, which differs from the *causal* membership advantage τ_{ATE} because $\mathcal{P}_{Y_i|A_i=0} \neq \mathcal{P}_{Y_i(0)|A_i=1}$, due to the distribution shift $\mathcal{P}_0 = \mathcal{P}_{X_i|A_i=0} \neq \mathcal{P}_{X_i|A_i=1} = \mathcal{P}_T$. To correct this confounding bias and recover τ_{ATE}, Proposition 4.1 naturally suggests an *inverse probability weighting* (IPW) estimator:

$$\hat{\tau}^{(\mathrm{IPW})}_{\mathrm{ATE}} = \frac{1}{n_1}\sum_{A_i=1} Y_i - \frac{1}{n_1}\sum_{A_i=0} \frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)} Y_i \,, \tag{3}$$

where $\hat{\pi}$ is an estimate of the *propensity score* π defined in Proposition 6.2. The estimator (3) belongs to the family of IPW debiasing methods widely used in causal inference with observational data [61].[2] The propensity score estimator $\hat{\pi}$ can be obtained by training any probabilistic binary classifier $\mathcal{X} \to [0, 1]$ to distinguish members from non-members, yielding an estimate of the probability that any $x \in \mathcal{X}$ is a member. Member and non-member data can be split into two independent parts, one used in Algorithm 3 and one used to learn the classifier $\hat{\pi}$. Cross-fitting, which alternates the roles of these splits and averages the results, can be used to leverage all the data efficiently. Section 7 shows how $\hat{\pi}$ is estimated in practice: the overlap assumption is made on the underlying data representations (e.g., text or image embeddings), and propensity scores are functions of these representations.

A well-trained propensity score model captures the distributional differences between the conditional feature distributions $X_i \mid A_i = 1$ and $X_i \mid A_i = 0$, assigning higher scores to regions of the feature space that are more characteristic of members than non-members. Consistency of $\hat{\pi}$ is measured as $\Delta_{\hat{\pi}} = \mathbb{E}_{\mathcal{P}_0}\left[\left|\frac{\pi(X_i)}{1-\pi(X_i)} - \frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\right|\right]$: $\hat{\pi}$ is ℓ^1-consistent if $\Delta_{\hat{\pi}} \to 0$. Next assumption is necessary to ensure that each data point $x \in \mathcal{D}_T$ has a positive (and lower-bounded, as opposed to Assumption 6.1) probability of being observed in the non-member set. Theorem 6.4 shows that if propensity scores are learned well-enough, the ATE for the negative loss-based attack is well estimated by $\hat{\tau}^{(\mathrm{IPW})}_{\mathrm{ATE}}$.

Assumption 6.3. There exists $\eta > 0$ such that $\pi(x) \leqslant 1 - \eta$ and $\hat{\pi}(x) \leqslant 1 - \eta$ for all $x \in \mathrm{supp}(\mathcal{P}_T)$.

Theorem 6.4. *Assume that $(X_i, A_i, Y_i)_{i\in[n]}$ are obtained using Algorithm 3 with $Y_i = -\mathcal{L}(\theta, X_i)$, that Assumptions 6.1 and 6.3 hold, and that $\mathcal{A}$ is α-error stable and β-uniform training stable. Assume outcomes Y_i are in $[0, 1]$ almost surely. Furthermore, assume that $\hat{\pi}$ is independent from $(X_i, A_i, Y_i)_{i\in[n]}$. Then, with probability $1 - Ce^{-t}$:*

$$|\hat{\tau}^{(\mathrm{IPW})}_{\mathrm{ATE}} - \tau_{\mathrm{ATE}}| = \mathcal{O}\big(\Delta_{\hat{\pi}} + \tfrac{1}{\eta}\sqrt{t/n} + \sqrt{nt\alpha^2} + \sqrt{nt\beta^2}\big)\,.$$

Theorem 6.4 has the same overall structure as the one-run result (Theorem 5.4), up to two additional terms: the additive bias $\Delta_{\hat{\pi}}$ of the propensity score estimator, and the factor $1/\eta$. The presence of $1/\eta$ is expected: when η is close to zero, IPW estimators are known to become unstable.

Robustness of propensity score estimation. Propensity score estimation is a key challenge in the application of IPW correction. Extreme or near-zero propensity scores can lead to instability and high variance when overlap is poor, while misspecification of the parametric model class can introduce bias. In practice, it is therefore advisable to consider multiple model classes, including non-parametric approaches, both to balance bias and variance and to provide robustness checks. Bootstrap-based confidence intervals and calibration methods, such as post-hoc data-driven calibration, can further help assess the reliability of the estimated propensity scores and the robustness of the resulting estimates. Additionally, sensitivity analyses can examine how much the estimated propensity scores would need to deviate from their fitted values for the estimated ATE, $\hat{\tau}^{(\mathrm{IPW})}_{\mathrm{ATE}}$, to lose statistical significance [28, 85]. For further discussion and practical guidance, we refer readers to the extensive literature on effective propensity score estimation in causal inference [67, 3, 39], as well as to the broader framework of *Veridical Data Science* [82].

[2] Note that $\mathcal{P}_T = \mathcal{P}_{X_i|A_i=1}$, so that the target ATE on $\mathcal{P}_T$ is in fact the ATT (ATE on the treated). This explains why our IPW estimator in Equation (3) differs from the classical IPW estimator.

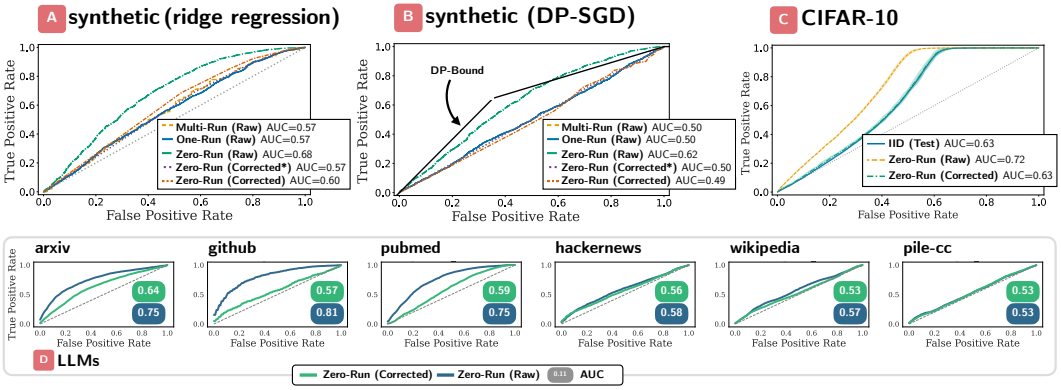


Figure 3: ROC curves and AUCs for linear regression on synthetic data (**A** and **B**), Resnet-16 on CIFAR-10 data (**C**), and `Pythia-12b` on six different text corpora (**D**).

Other observational estimators. Alternative causal inference estimators can be used to mitigate the instability of IPW in the presence of propensity scores. Letting $\hat{\mu}_0$ be an outcome regression model that estimates conditional MIA response $\mu_0 : x \in \mathcal{X} \mapsto \mathbb{E}\left[Y_i(0)|X_i = x\right]$, the *G-Formula* [77] and *Augmented Inverse Probability Weighting* (AIPW) estimators write as [69]:

$$\hat{\tau}^{(\mathrm{G})}_{\mathrm{ATE}} = \frac{1}{n_1}\sum_{A_i=1} Y_i - \hat{\mu}_0(X_i)\,, \qquad \hat{\tau}^{(\mathrm{AIPW})}_{\mathrm{ATE}} = \hat{\tau}^{(\mathrm{G})}_{\mathrm{ATE}} - \frac{1}{n_1}\sum_{i=1}^{n} \frac{(1-A_i)\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\left(Y_i - \hat{\mu}_0(X_i)\right). \quad (4)$$

The outcome regression model $\hat{\mu}_0$ is typically learned by fitting a regression of Y_i on X_i using non-member data points (i.e., those with $A_i = 0$).

The AIPW estimator is *doubly robust*, meaning it is consistent if either the propensity score model or the outcome regression is correctly specified. Although they are more robust to extreme propensity scores, G-Formula and AIPW still require Assumptions 6.1 and 6.3 to hold. Section E generalizes these two estimators to other causal metrics.

Insight # 3

Propensity-score based correction enables unbiased evaluation of MIAs in the zero-run regime via simple binary classifiers. This yields a practical, scalable way to correct distribution-shift bias without requiring model retraining or randomized data inclusion.

7 Experiments

We illustrate how our causal framework debiases MIA evaluation across four settings. In Figures 3 and 4, Multi-Run, One-Run, and Zero-Run correspond to Algorithms 1 to 3, respectively. *Raw* denotes uncorrected estimates, while *Corrected* applies propensity-score adjustment in the Zero-Run regime. Finally, *Zero-Run (IID)* serves as an idealized baseline used for comparison and sanity checking. In this setting, non-members are obtained as a randomized test split from the same dataset as the members, so no correction is needed. Additional experimental details, results and visualizations are provided in Section H.

Synthetic data. We consider linear regression on Gaussian samples, using overparameterized ridge regression and underparameterized DP-SGD, both satisfying error stability and uniform training stability. We evaluate a loss-based MIA under three regimes: *multi-run*, *one-run*, and *zero-run*, with non-members drawn from a shifted Gaussian. We consider both oracle propensity scores (denoted as *Corrected**), and estimated ones with logistic regression. ROC curves and AUCs are reported in Figure 3 (A-B). Multi-run and one-run evaluations nearly coincide as interference is negligible, whereas naive zero-run overestimates MIA performance under distribution shift. Propensity-score correction restores agreement with non-shifted baselines. For DP-SGD, the naive zero-run evaluation violates the ROC upper bound implied by (ε, δ)-DP [16], while the corrected versions respects it, highlighting the importance of debiasing.

ResNet on CIFAR-10. We perform zero-run evaluation on CIFAR-10 [37] using a ResNet-16. Members are sampled from a mixture of clean and noisy images (90%/10%) and non-members from

the reverse (10%/90%), inducing strong distribution shift while satisfying overlap. Propensity scores are estimated via a fine-tuned ImageNet-pretrained ResNet classifier. Figure 3 (C) reports ROC curves and AUCs for a loss-based MIA. Naive zero-run evaluation inflates attack performance (AUC ≈ 0.72) due to distribution shift, misleadingly suggesting high memorization. Propensity-score correction recovers the true causal signal (AUC ≈ 0.63), closely matching an IID baseline where non-members are drawn from the same mixture as members (AUC ≈ 0.65), confirming effective debiasing.

Pretrained LLM. We perform a zero-run evaluation on `Pythia-12b` [6], an LLM for which membership information is available for a subset of data points. We consider 6 datasets where MIA evaluation biases has been recently highlighted [50] due to distribution shift induced by deduplication [17]. For each corpus, we split data into train/test sets stratified by membership, vectorize training data using a `BagOfWords` encoder [50], and fit a `RandomForest` classifier to estimate propensity scores. We consider the standard MIA based on perplexity. Results are shown in Figure 3 (D). When distribution shift is large (e.g., `arxiv`, `github`, `pubmed_central`), members become easily distinguishable from non-members and naive evaluation substantially inflates ROC curves. In contrast, causal correction largely removes the signal, indicating little true memorization, consistent with prior work on MIAs for pretrained LLMs [17]. For `hackernews`, `wikipedia`, and `pile_cc`, shifts are weaker and the correction is correspondingly modest.

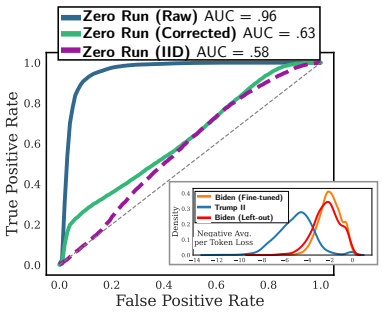


Figure 4: Fine-tuned `Pythia-70m` model: raw, corrected and pseudo-ground truth (IID) ROC curves.

Fine-tuned LLM. To construct an idealized IID reference setting for MIA evaluation on an LLM, we fine-tune `Pythia-70m` on a scraped dataset of White House press statement titles under the Biden (2021-2025) and Trump II (from 2025) administrations, denoted by B and T respectively, between which strong linguistic distribution shifts have been documented [36]).[3] We further split B into two sets (B_1, B_2) and fine-tune the model on B_1 using `LoRA` [27]. This controlled construction enables an IID evaluation baseline: since B_2 is drawn from the same distribution as B_1, MIA performance on (B_1, B_2) serves as a pseudo-ground truth reference, in contrast to (B_1, T) which exhibits distribution shift. We evaluate a loss-based MIA using the negative average per-token loss on both (B_1, B_2) and (B_1, T), with and without propensity-score correction (zero-run). Results are shown in Figure 4. We find that naive MIA on (B_1, T) is severely inflated by distribution shift (AUC > 0.95), while propensity-score correction closely matches the IID reference (B_1, B_2) (AUC ≈ 0.6), supporting the effectiveness of our debiased estimator.

Insight # 4

Our approach offers a practical solution to a key problem in MIA for LLMs: correcting evaluation biases induced by distribution shift, which systematically inflate perceived memorization [50, 17]. Our framework yields faithful estimates of true memorization in language models.

8 Related Work

An extended discussion of related work is provided in Appendix A.

Limits of the one-run regime. Recent work has studied the fundamental limits of the one-run regime for auditing differential privacy (DP) guarantees [34, 79]. Their objectives and conceptual frameworks differ from ours: they derive DP lower bounds, whereas we aim to obtain unbiased MIA evaluation.

Formalizing and addressing MIA bias. MIAs are traditionally framed as randomized privacy games [80, 65]. However, unlike our causal framework, these games fail to capture the interference and distribution shifts characteristic of the one-run and zero-run regimes [50]. Focusing on distribution shift in MIA evaluation for LLMs, prior work [50, 18] suggest potential solutions, but these require intervention prior or during training (e.g., randomized train/test splits or data injection) or rely on heuristics. Kazmi et al. [33] propose using synthetic data points, but this approach is costly,

[3] We manually verified that none of these texts appear in the pretraining data of `Pythia-70m`.

data-hungry, and difficult to ground theoretically. In constract, we provide principled, causally valid estimators that rely only on simple binary classifiers or regressors.

Causality and MIAs. We mention two prior works at the intersection of MIAs and causality, though orthogonal to our focus. Tople et al. [74] show that causal learning can effectively reduce MIA success—a topic orthogonal to our work. Baluta et al. [1] use causal reasoning to investigate whether MIA performance arises solely from imperfect generalization, as suggested by prior work [81, 63].

Causal inference under interference. Contrarily to previous work which has focused on known and sparse interference patterns [70, 29, 46, 56, 5, 13, 40], interference in our setting is dense, unobserved, and arises as a random function of the covariates X_i through the training algorithm $\mathcal{A}$, which acts as a mediator (Figure 2). Consequently, we impose assumptions on the mediator itself, leveraging learning-theoretic bounds on the influence of individual training points on the model's behavior. Our estimands are thus not conditioned on a interference graph. Together, our mediator-induced interference model, the unconditional estimands, and the mild assumptions we impose distinguish our setting from existing work and make Theorem 5.4 independently interesting within the interference literature.

9 Conclusion

Prior work [49, 17] has raised concerns that conclusions drawn by earlier MIA studies regarding memorization and privacy may have been overstated, notably due to methodological shortcomings in their evaluation protocols. In this paper, we introduce a formal causal framework for MIA evaluation that rigorously characterizes these methodological issues and provides principled solutions to address them. Specifically, we develop statistically valid estimators that account for interference and distribution shifts in both one-run and zero-run regimes. Our framework unifies the evaluation of MIAs, provides a theoretical understanding of the sources of bias, and offers a principled, low-cost, and readily implementable correction strategy based on standard causal inference methodologies. Importantly, it also enables practical post-hoc assessment without requiring access to, or trust in, model providers.

Our contribution addresses a significant gap in the literature: in the absence of practical solutions to the aforementioned methodological issues, recent studies have continued to evaluate MIAs under distributionally shifted conditions without adequately addressing the resulting biases, see e.g. [60, 59]. More broadly, we argue for a causal perspective on MIA evaluation, moving beyond correlations between membership and algorithmic outputs to estimating their causal relationship. We argue that this perspective is essential to obtaining reliable assessments of memorization and privacy risks and avoiding potentially misleading conclusions.

While we focus on population-level metrics—evaluating on-average memorization over a target population—it is also crucial to understand subgroup disparities in memorization [11, 38]. In causal terms, this amounts to estimating heterogeneous treatment effects [76]. Our framework provides a foundation for principled subgroup-level memorization estimates, even in challenging one-run and zero-run regimes. Moreover, this causal perspective may extend beyond MIAs to other privacy attacks, such as attribute inference [22], where existing evaluation biases [32] could be mitigated by counterfactual reasoning.

Acknowledgements

Parts of this work were conducted while LB was a member of PreMeDICaL team, Inria, Idesp, Inserm, Université de Montpellier. LB's work at EPFL is supported by an EPFL AI Center Postdoctoral Fellowship. Parts of the experiments of this paper were run on the RCP Cluster at EPFL.

This work is partially supported by grant ANR-20-CE23-0015 (Project PRIDE), ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR, and ANR-22-PESN-0014 project of the Digital Health PEPR under the France 2030 program. This work was performed using HPC resources from GENCI–IDRIS (Grant 2023-AD011014018R3).

References

[1] Teodora Baluta, Shiqi Shen, S. Hitarth, Shruti Tople, and Prateek Saxena. Membership inference attacks and generalization: A causal perspective. In *CCS*, 2022.

[2] Federico Barbero, Xiangming Gu, Christopher A Choquette-Choo, Chawin Sitawarin, Matthew Jagielski, Itay Yona, Petar Veličković, Ilia Shumailov, and Jamie Hayes. Extracting alignment data in open models. *arXiv preprint arXiv:2510.18554*, 2025.

[3] Umberto Benedetto, Stuart J Head, Gianni D Angelini, and Eugene H Blackstone. Statistical primer: propensity score matching and its alternatives. *European Journal of Cardio-Thoracic Surgery*, 53(6):1112–1117, 2018.

[4] 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*, 36:314–330, 2023.

[5] Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*, pages 1028–1038. PMLR, 2020.

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

[7] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[8] Marc Buyse. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, 29(30):3245–3257, dec 2010.

[9] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284, 2019.

[10] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[11] 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)*, pages 1897–1914. IEEE, 2022.

[12] Carlos Cinelli, Avi Feller, Guido Imbens, Edward Kennedy, Sara Magliacane, and Jose Zubizarreta. Challenges in statistics: A dozen challenges in causality and causal inference. *arXiv preprint arXiv:2508.17099*, 2025.

[13] Duncan A Clark and Mark S Handcock. An approach to causal inference over stochastic networks. *arXiv preprint arXiv:2106.14145*, 2021.

[14] Concord Music Group v. Anthropic. Concord Music Group, Inc. v. Anthropic PBC, 5:24-cv-03811 - CourtListener.com, 2023. URL `https://www.courtlistener.com/docket/68889092/concord-music-group-inc-v-anthropic-pbc/`.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[16] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

[17] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

[18] Cédric Eichler, Nathan Champeil, Nicolas Anciaux, Alexandra Bensamoun, Héber H. Arcolezi, and José Maria De Fuentes. Nob-mias: Non-biased membership inference attacks assessment on large language models with ex-post dataset construction. In Mahmoud Barhamgi, Hua Wang, and Xin Wang, editors, *Web Information Systems Engineering – WISE 2024*, pages 441–456, Singapore, 2025. Springer Nature Singapore. ISBN 978-981-96-0570-5.

[19] European Data Protection Board. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of ai models, 2024. URL `https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en`.

[20] Mathieu Even and Julie Josse. Rethinking the win ratio: A causal framework for hierarchical outcome analysis, 2025. URL `https://arxiv.org/abs/2501.16933`.

[21] Laura Forastiere, Davide Del Prete, and Valerio Leone Sciabolazza. Causal inference on networks under continuous treatment interference. *Social Networks*, 76:88–111, 2024.

[22] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *CCS*, 2015.

[23] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[24] Jamie Hayes, Ilia Shumailov, Christopher A Choquette-Choo, Matthew Jagielski, George Kaissis, Katherine Lee, Milad Nasr, Sahra Ghalebikesabi, Niloofar Mireshghallah, Meenatchi Sundaram Mutu Selva Annamalai, et al. Strong membership inference attacks on massive datasets and (moderately) large language models. *arXiv e-prints*, pages arXiv–2505, 2025.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] Miguel A Hernán and James M Robins. Causal inference, 2010.

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

[28] Melody Huang and Samuel D Pimentel. Variance-based sensitivity analysis for weighting estimators results in more informative bounds. *Biometrika*, 112(1):asae040, 2025.

[29] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the american statistical association*, 103(482):832–842, 2008.

[30] Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge, UK, 2015. ISBN 978-0521885881.

[31] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33: 22205–22216, 2020.

[32] Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? In *PCCS*, 2022.

[33] Mishaal Kazmi, Hadrien Lautraite, Alireza Akbari, Qiaoyue Tang, Mauricio Soroco, Tao Wang, Sébastien Gambs, and Mathias Lécuyer. Panoramia: Privacy auditing of machine learning models without retraining. *Advances in Neural Information Processing Systems*, 37:57262–57300, 2024.

[34] Amit Keinan, Moshe Shenfeld, and Katrina Ligett. How well can differential privacy be audited in one run? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=004uTlSufe.

[35] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A Watermark for Large Language Models, May 2024. URL http://arxiv.org/abs/2301.10226. arXiv:2301.10226 [cs].

[36] Robert Körner, Jennifer R Overbeck, Erik Körner, and Astrid Schütz. How the linguistic styles of donald trump and joe biden reflect different forms of power. *Journal of Language and Social Psychology*, 41(6):631–658, 2022.

[37] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[38] Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. Disparate vulnerability to membership inference attacks. *Proc. Priv. Enhancing Technol.*, 2022(1):460–480, 2022.

[39] Mingxiang Li. Using the propensity score method to estimate causal effects: A review and practical guide. *Organizational Research Methods*, 16(2):188–226, 2013.

[40] Shuangning Li and Stefan Wager. Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, 50(4):2334–2358, 2022.

[41] Terrance Liu, Matteo Boglioni, Yiwei Fu, Shengyuan Hu, Pratiksha Thaker, and Zhiwei Steven Wu. Enhancing One-run Privacy Auditing with Quantile Regression-Based Membership Inference. *arXiv preprint arXiv:2506:15349*, 2025.

[42] Saeed Mahloujifar, Luca Melis, and Kamalika Chaudhuri. Auditing f-differential privacy in one run. *arXiv preprint arXiv:2410.22235*, 2024.

[43] Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset Inference: Ownership Resolution in Machine Learning, April 2021. URL http://arxiv.org/abs/2104.10706. arXiv:2104.10706 [stat].

[44] Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: did you train on my dataset? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[45] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, mar 1947. ISSN 0003-4851.

[46] Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

[47] Lu Mao. On causal estimation using u-statistics. *Biometrika*, 105(1):215–220, dec 2017. ISSN 1464-3510.

[48] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.719. URL https://aclanthology.org/2023.findings-acl.719/.

[49] Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2369–2385, 2024.

[50] 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). *arXiv preprint arXiv:2406.17975*, 2024.

[51] Caleb H Miles, Maya Petersen, and Mark J van der Laan. Causal inference when counterfactuals depend on the proportion of all subjects exposed. *Biometrics*, 75(3):768–777, 2019.

[52] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

[53] 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 *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1631–1648, 2023.

[54] New York Times v. OpenAI. NYT v. OpenAI: The Times's About-Face, 2023. URL `https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/`.

[55] Elizabeth L Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J Van der Laan. Causal inference for social network data. *Journal of the American Statistical Association*, 119(545):597–611, 2024.

[56] Georgia Papadogeorgou, Fabrizia Mealli, and Corwin M Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3): 778–787, 2019.

[57] Yannik Pitcan. A note on concentration inequalities for u-statistics. *arXiv preprint arXiv:1712.06160*, 2017.

[58] S. J. Pocock, C. A. Ariti, T. J. Collier, and D. Wang. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*, 33(2):176–182, sep 2011. ISSN 1522-9645.

[59] Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. Scaling up membership inference: When and how attacks succeed on large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4165–4182, 2025.

[60] Tao Qi, Jinhua Yin, Dongqi Cai, Yueqi Xie, Huili Wang, Zhiyang Hu, Peiru Yang, Guoshun Nan, Zhili Zhou, Chuhan Wu, et al. Auditing unauthorized training data from ai generated content using information isotopes. *Nature Communications*, 17(1):3007, 2026.

[61] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[62] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[63] Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *ICML*, 2019.

[64] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: tracing through training, February 2020. URL `http://arxiv.org/abs/2002.00937`. arXiv:2002.00937 [stat].

[65] Ahmed Salem, Giovanni Cherubin, David Evans, Boris Kopf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Beguelin. SoK: Let the Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning . In *IEEE Symposium on Security and Privacy (S&P)*, 2023.

[66] 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.

[67] Koichiro Shiba and Takuya Kawahara. Using propensity scores for causal inference: pitfalls and tips. *Journal of epidemiology*, 31(8):457–463, 2021.

[68] 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)*, pages 3–18. IEEE, 2017.

[69] Heng Shu and Zhiqiang Tan. Improved estimation of average treatment effects on the treated: Local efficiency, double robustness, and beyond. *arXiv preprint arXiv:1808.01408*, 2018.

[70] Michael E Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476): 1398–1407, 2006.

[71] Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

[72] Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 36:49268–49280, 2023.

[73] Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

[74] Shruti Tople, Amit Sharma, and Aditya Nori. Alleviating privacy attacks via causal learning. In *International Conference on Machine Learning*, pages 9537–9547. PMLR, 2020.

[75] Stefan Wager. Causal inference: A statistical learning approach, 2024.

[76] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[77] Aolin Wang, Roch A Nianogo, and Onyebuchi A Arah. G-computation of average treatment effects on the treated and the untreated. *BMC medical research methodology*, 17(1):3, 2017.

[78] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80, dec 1945. ISSN 0099-4987.

[79] Zihang Xiang, Tianhao Wang, and Di Wang. Privacy audit as bits transmission: (im)possibilities for audit by one run. In *Proceedings of the 34th USENIX Conference on Security Symposium*, 2025.

[80] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *CCS*, 2022.

[81] 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)*, pages 268–282. IEEE, 2018.

[82] Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020.

[83] Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=sT7UJh5CTc`.

[84] Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. Membership inference attacks cannot prove that a model was trained on your data. *arXiv preprint arXiv:2409.19798*, 2024.

[85] Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761, 2019.

A Extended Related Work

Limits of the one-run regime. Recent work has studied the fundamental limits of the one-run regime for auditing differential privacy (DP) guarantees [34, 79]. These studies, like ours, trace the key limitation to interference between data points. Beyond this shared root cause, the objectives and conceptual frameworks differ: the prior work focuses on establishing DP lower bounds using information-theoretic perspectives such as bit transmission or hypothesis testing, whereas we focus on evaluating MIAs, adopt a causal framing that explicitly captures how interference can invalidate counterfactual reasoning, and propose solutions to mitigate these limitations.

Formalizing and addressing MIA bias. MIAs are traditionally framed as randomized privacy games [80, 65]. However, unlike our causal framework, these games fail to capture the interference and distribution shifts characteristic of the one-run and zero-run regimes [50]. Focusing on distribution shift in MIA evaluation for LLMs, Meeus et al. [50] suggest potential solutions, but these require intervention priors or during training (e.g., randomized train/test splits or data injection) or rely on heuristics. Kazmi et al. [33] propose using synthetic data points from a generative model trained on the member data as non-members, but this approach is costly, data-hungry, and difficult to evaluate. While they provide theoretical conditions for obtaining meaningful DP lower bounds, their method offers no guarantees for MIA performance estimates. Eichler et al. [18] develops techniques to mitigate distribution shift in post-hoc text data, yet offer no formal guarantees for the resulting MIA estimates. In contrast, we provide principled, causally valid estimators that rely only on simple binary classifiers or regressors.

Zhang et al. [84] argue that several alternative techniques can be used to demonstrate to a third party that a model was trained on their data when the training set is unknown and retraining is impractical, including data watermarking [64, 35], random canary insertion [9], and data reconstruction attacks [10, 52]. The first two approaches require proactive intervention and trust in the model provider, while data reconstruction is fundamentally more difficult than membership inference. Developing statistically valid post-hoc MIA evaluations that do not rely on model provider cooperation or trust—the focus of our work—thus remains an important and practically relevant problem.

Causality and MIAs. We mention two prior works at the intersection of MIAs and causality, though orthogonal to our focus. Tople et al. [74] show that causal learning can effectively reduce MIA success: unlike predictive models that may exploit spurious correlations, causal models rely on stable, invariant relationships between input and features outcomes. Baluta et al. [1] use causal reasoning to investigate whether MIA performance arises solely from imperfect generalization, as suggested by prior work [81, 63]. They find that while overfitting contributes to MIA success, it is in fact driven by a complex interplay of multiple factors.

Causal inference under interference. Seminal work formalized causal inference with interference via exposure maps [70, 29, 46], which describe how each unit's outcome depends on other units' treatment assignments. These mappings are often represented by a graph, whose edges encode potential interference. Subsequent work has largely focused on fixed and sparse exposure maps [5, 56, 55], with extensions to random [13, 40], weighted [21], and dense [51] settings. In these frameworks, estimands are typically direct or indirect effects defined conditional on an observed exposure map. In contrast, interference in our setting is dense, unobserved, and arises as a random function of the covariates X_i through the training algorithm $\mathcal{A}$, which acts as a mediator (Figure 2). Rather than observing or modeling the exposure map directly, we impose assumptions on the mediator itself, leveraging learning-theoretic bounds on the influence of individual training points on the model's behavior at other points. As a result, our estimands are not conditioned on a particular exposure map: τ_{ATE} averages over realizations of the mediator $\mathcal{A}$, and thus over the induced interference patterns. Together, our mediator-induced interference model, the unconditional estimands, and the mild assumptions we impose distinguish our setting from existing work and make Theorem 5.4 independently interesting within the interference literature.

B Additional Causal Inference Details

B.1 Causal Inference Assumptions

We here explain the different regimes used in causal inference to estimate treatment effects. In particular, we explicit the underlying assumptions that are made in these regimes.

Randomized Control Trial (RCT). In RCTs, treatment is randomized and there are no interferences. Assumption 2.1 is thus assumed to hold with *2. (a)* for assignment mechanism. The overlap assumption holds, since $\mathbb{P}(A_i = 1|X_i) = \mathbb{P}(A_i = 1)$ does not depend on the covariates. The SUTVA assumption means that there are no interferences, and is usually made for RCTs.

Observational Studies. RCTs are usually expensive and have restrictive eligibility criteria, hindering both their statistical power (smaller samples) and their population. As a consequence, the promise of observational data through the use of hospital electronic health records has emerged [61]. However this comes at the cost of non-randomized treatment assignments, and estimating treatment effects on observational data is infeasible without additional structural assumption. As a consequence, if there are no interferences or if they are neglected, Assumption 2.1 is assumed to hold with *2. (b)* for the assignment mechanism: the observed covariates X_i are rich enough, so that there are no unobserved confounders. The overlap assumption is assumed to hold here, and does not hold by design as in RCTs: given covariates X_i, the probability of patient i of being treated is bounded away from 0 and 1.

RCT with Interferences. For vaccine trials for instance, the SUTVA assumption can no longer be made, due to herd immunity. As a consequence, it is replaced using the interference formalism introduced in Section 2.2 as follows.

Assumption B.1 (Assumptions under interference)**.**

1. *(Interference)* $Y_i = Y_i(\bar{A})$;
2. *(Generalized Unconfoundedness)* $\bar{A} \perp\!\!\!\perp \{\bar{Y}(\bar{a})\}_{\bar{a}}$;
3. *(Generalized Overlap)* For all $i \neq j$,
$$0 < \mathbb{P}\left(A_i = a \mid \bar{A}_{-i}\right) < 1.$$

Observational Studies with Interferences. Combining both difficulties (confounders and interferences), such as vaccine observational data, requires making the following assumption.

Assumption B.2 (Interference and confounders)**.**

1. *(Interference)* $Y_i = Y_i(\bar{A})$;
2. *(Conditional Generalized Unconfoundedness)* $\bar{A} \perp\!\!\!\perp \{\bar{Y}(\bar{a})\}_{\bar{a}} \mid \bar{X}$;
3. *(Conditional Generalized Overlap)* For all $i \neq j$,
$$0 < \mathbb{P}\left(A_i = a \mid \bar{X}, \bar{A}_{-i}\right) < 1 \text{ a.s.}$$

B.2 Discussion on the Causal AUC

The classical AUC compares the scores of outcomes for treated ($A_i = 1$) and control ($A_i = 0$) datapoints, and can be expressed as:
$$\mathbb{P}(Y_i \geqslant Y_j | A_i = 1, A_j = 0) \ .$$
It is however not straightforward to write the causal AUC, defined as the area under the causal ROC curve, as a comparison estimand. In causal inference, the most common quantity for comparing potential outcomes is the probability of necessity and sufficiency (PNS) [73], defined as $\mathrm{PNS} = \mathbb{P}(Y_i(1) > Y_i(0))$. This represents the probability that a given datapoint has a higher score if treated compared to not being treated. While PNS is useful for assessing treatment effectiveness, it is unidentifiable in practice. Instead, the Mann-Whitney-Wilcoxon test statistics [78, 45], along with other pairwise comparison methods like the Win Ratio or Net Benefit [58, 8, 47, 20], estimate an identifiable quantity known as the *Win Proportion* (WP), given by:
$$\mathrm{WP} \stackrel{\text{def}}{=} \mathbb{P}(Y_i(1) > Y_j(0) \mid A_i = A_j = 1), .$$
This estimand is a causal interpretation of the AUC, replacing the unknown counterfactual with that of another datapoint, making it identifiable under standard causal assumptions. However, the WP is not well-defined in the presence of interference, which makes it unidentifiable. In the one-run and zero-run settings, the joint distribution of $(Y_i(1), Y_j(0))$ becomes unidentifiable because changes in A_i or A_j (due to datapoint insertions) affect the potential outcomes $Y_i(1)$ and $Y_j(0)$. Therefore, a Causal Comparison AUC can be defined as follows using causal inference with interference notations.

Definition B.3 (Causal Comparison AUC)**.** For $i \neq j$, and $a_i, a_j \in \{0,1\}$, define $Y_{i,j}(a_i, a_j)$ as $Y_{i,j}(a_i, a_j) =$

$$Y_i(A_1, \ldots, A_{i-1}, a_i, A_{i+1}, \ldots, A_{j-1}, a_j, A_{j+1}, \ldots, A_n).$$

We then define the causal AUC as:

$$\tilde{\tau}_{\mathrm{AUC}} \stackrel{\text{def}}{=} \mathbb{P}\left(Y_{i,j}(1,0) > Y_{j,i}(0,1) \,|\, A_i = A_j = 1\right) .$$

Importantly, note that if SUTVA holds (no interferences, $Y_i(\bar{a}) = Y_i(a_i)$ for any $\bar{a} \in \{0,1\}^n$) we have $Y_{i,j}(a_i, a_j) = Y_i(a_i)$ and $\tau_{\mathrm{AUC}} = \mathrm{WP}$. Definition B.3 is thus a strict generalization of the WP estimand to the interference setting. Under interferences, $\tilde{\tau}_{\mathrm{AUC}}$ is however in general not identifiable, even under Assumption B.1 or Assumption B.2, that would instead need to be replaced by the following (interference observational setting).

Assumption B.4 (Alternative assumption)**.**

1. *(Interference)* $Y_i = Y_i(\bar{A})$;
2. *(Conditional Generalized Unconfoundedness)* $\bar{A} \perp\!\!\!\perp \{\bar{Y}(\bar{a})\}_{\bar{a}} \mid \bar{X}$;
3. *(Conditional Generalized Overlap)* For all $i \neq j$,

$$0 < \mathbb{P}\left(A_i = a, A_j = a' \mid \bar{X}, \bar{A}_{-(i,j)}\right) < 1 \text{ a.s.}$$

B.3 Identifiability Proofs

The proofs of the ATE parts of Propositions 4.1 to 6.2 are all implied by the following.

Proposition B.5. *Assume that Assumption B.2 holds. Then:*

$$\tau_{\mathrm{ATE}} = \frac{\mathbb{E}[A_i Y_i]}{\mathbb{P}(A_i=1)} - \frac{\mathbb{E}\left[\frac{\pi(X_i)(1-A_i)}{1-\pi(X_i)} Y_i\right]}{\mathbb{P}(A_i=0)} . \tag{5}$$

Proposition B.5 implies Proposition 6.2 since under Assumption 6.1 we have that Assumption B.2 holds in the Zero-run setting. Then, in the One-run and Multi-run settings, Assumptions B.1 and 2.1 respectively hold, which implies Assumption B.2 and thus Proposition B.5. Then specifying Proposition B.5 in the One-run setting for which π is constant equal to 0.5, we get the ATE parts of Propositions 5.1 and 4.1. Identification for the other causal metrics follow from similar arguments.

Proof of Proposition B.5. Under the assumptions of either of Propositions 4.1 to 6.2, we have that Assumption B.2 holds. First,

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}_T}[Y_i(1)] &= \mathbb{E}\left[Y_i(1) | A_i = 1\right] \\
&= \mathbb{E}\left[Y_i | A_i = 1\right] \quad \text{(Interference assumption in Assumption B.2)} \\
&= \mathbb{E}\left[A_i Y_i | A_i = 1\right] .
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}\left[A_i Y_i\right] &= \mathbb{E}\left[A_i Y_i | A_i = 1\right] \mathbb{P}\left(A_i = 1\right) + \mathbb{E}\left[A_i Y_i | A_i = 0\right] \mathbb{P}\left(A_i = 0\right) \\
&\mathbb{E}\left[A_i Y_i | A_i = 1\right] \mathbb{P}\left(A_i = 1\right) + 0 ,
\end{aligned}$$

leading to $\mathbb{E}_{\mathcal{P}_T}[Y_i(1)] = \frac{\mathbb{E}[A_i Y_i]}{\mathbb{P}(A_i=1)}$. For the other term,

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}_T}[Y_i(0)] &= \mathbb{E}\left[Y_i(0) | X_i \sim \mathcal{P}_T\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i(0) | \bar{X}\right] | X_i \sim \mathcal{P}_T\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i(0) | \bar{X}, \bar{A}_{-i}, A_i = 0\right] | X_i \sim \mathcal{P}_T\right] \quad \text{(Conditional Generalized Unconfoundedness)} \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i | \bar{X}, \bar{A}_{-i}, A_i = 0\right] | X_i \sim \mathcal{P}_T\right] \quad \text{(Interference assumption in Assumption B.2)} \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i | X_i, A_i = 0\right] | X_i \sim \mathcal{P}_T\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(1 - A_i) Y_i | X_i, A_i = 0\right] | X_i \sim \mathcal{P}_T\right] .
\end{aligned}$$

Now,

$$\begin{aligned}\mathbb{E}\left[(1-A_i)Y_i|X_i\right] &= \mathbb{E}\left[(1-A_i)Y_i|X_i, A_i=0\right]\mathbb{P}\left((1-A_i)A_i=0|X_i\right) + \mathbb{E}\left[(1-A_i)Y_i|X_i, A_i=1\right]\mathbb{P}\left(A_i=1|X_i\right)\\ &= \mathbb{E}\left[(1-A_i)Y_i|X_i, A_i=0\right](1-\pi(X_i)) + 0\,.\end{aligned}$$

This leads to:

$$\begin{aligned}\mathbb{E}_{\mathcal{P}_T}[Y_i(0)] &= \mathbb{E}\left[\mathbb{E}\left[(1-A_i)Y_i|X_i\right](1-\pi(X_i))^{-1}|X_i\sim\mathcal{P}_T\right]\\ &= \mathbb{E}\left[g(X_i)|X_i\sim\mathcal{P}_T\right]\,.\end{aligned}$$

with $g(x) = \mathbb{E}\left[(1-A_i)Y_i|X_i=x\right](1-\pi(x))^{-1}$. Now, notice that the density of $\mathcal{P}_T$ with respect to $\mathcal{P}_{X_i}$ is proportional to π: $\frac{\mathrm{d}\mathcal{P}_T(x)}{\mathrm{d}\mathcal{P}_{X_i}}(x) = \frac{\pi((x)}{\mathbb{P}(A_i=1)}$. Thus, $\mathbb{E}\left[g(X_i)|X_i\sim\mathcal{P}_T\right] = \mathbb{E}\left[\frac{\pi(X_i)}{\mathbb{P}(A_i=1)}g(X_i)|X_i\sim\mathcal{P}_{X_i}\right]$, concluding the proof. □

C Results in the underparametrized regime

The next theorem considers the underparameterized setting and shows that MIAs based on the loss cannot recover any signal ($\tau_{\mathrm{ATE}}\approx 0$) when the number of samples becomes large in front of the dimension of the model (i.e., $\frac{D}{n}\to 0$).

Theorem C.1 (Underparametrized Regime)**.** *Consider the assumptions of Theorem 5.4 but without uniform training stability. Further assume that $\theta=\mathcal{A}(\mathcal{D})\in\mathbb{R}^D$ has norm almost surely bounded by a constant B and that $\theta\mapsto\mathcal{L}(\theta,x)$ is L-Lipschitz for all $x\in\mathcal{X}$. Then:*

$$|\tau_{\mathrm{ATE}}| = \mathcal{O}\Big(\sqrt{n\log(n)\alpha^2} + \sqrt{\tfrac{D\log(BLn)}{n}}\Big)\,.$$

D Estimators for and generalization to all Causal MIA Evaluation Metrics

D.1 Multi-run setting

Estimators of the other causal metrics beyond the ATE in the multi-run setting write as:

$$\begin{aligned}\hat{\tau}_{\mathrm{ATE}} &= \frac{1}{n_1}\sum_{i:A_i=1}Y_i - \frac{1}{n_0}\sum_{i:A_i=0}Y_i\,,\\ \hat{\tau}_{\mathrm{TPR}}(t) &= \frac{1}{n_1}\sum_{i:A_i=1}\mathbb{1}_{\{Y_i\geqslant t\}}\\ \hat{\tau}_{\mathrm{FPR}}(t) &= \frac{1}{n_0}\sum_{i:A_i=0}\mathbb{1}_{\{Y_i\geqslant t\}}\\ \hat{\tau}_{\mathrm{AUC}} &= \frac{1}{n_1 n_0}\sum_{(i,j):(A_i,A_j)=(1,0)}\mathbb{1}_{\{Y_i>Y_j\}}\,,\\ \hat{\tau}_{\mathrm{TPR@FPR}}(\alpha) &= \frac{1}{n_1}\sum_{i:A_i=1}\mathbb{1}_{\{Y_i\geqslant\hat{t}_\alpha\}}\,,\end{aligned} \tag{6}$$

where $\hat{t}_\alpha$ is an estimator of the $1-\alpha$ quantile of $\{Y_i : A_i=0\}$, of respectively the ATE (causal membership advantage), the causal TPR and FPR at threshold $t\in\mathbb{R}$, the causal AUC and the causal TPR at fixed FPR $\alpha\in(0,1)$.

Proposition D.1. *Assume that $(X_i,A_i,Y_i)_{i\in[n]}$ is obtained in the Multi Run setting (Algorithm 1). Then, $\hat{\tau}_{\mathrm{ATE}}$, $\hat{\tau}_{\mathrm{ATE}}(t)$ for $t\in\mathbb{R}$ and $\hat{\tau}_{\mathrm{AUC}}$ are consistent estimators of τ_{ATE}, $\tau_{\mathrm{ATE}}(t)$ and τ_{AUC}. If $t\mapsto\mathrm{FPR}(t)$ is continuous, $\hat{\tau}_{\mathrm{TPR@FPR}}(\alpha)$ is a consistent estimator of $\tau_{\mathrm{TPR@FPR}}(\alpha)$ for $\alpha\in[0,1]$.*

Moving beyond consistency, non-asymptotic rates can be obtained directly from the difference in means estimator properties [75, e.g.,]. For $\hat{\tau}_{\mathrm{TPR}}(t)$, with probability $1-4e^{-\lambda}$,

$$|\hat{\tau}_{\mathrm{TPR}}(t) - \tau_{\mathrm{TPR}}(t)| \leqslant \sqrt{\frac{\lambda}{2n_1}} + \sqrt{\frac{\lambda}{2n_0}}\,,$$

and similarly for τ_{TPR}, while for the causal membership advantage (ATE) if outcomes Y_i are bounded by a constant B:

$$|\hat{\tau}_{\mathrm{ATE}} - \tau_{\mathrm{ATE}}| \leqslant B\sqrt{\frac{2\lambda}{n_1}} + B\sqrt{\frac{2\lambda}{n_0}}\,.$$

Non-asymptotic results for the causal AUC can be obtained via U-statistics [57].

D.2 One-run setting

In the one-run setting, the estimators of the different causal evaluation metrics are the same as in the multi-run (see Equation (6)), at the exception of the causal AUC estimator, which is now the area under the estimated causal ROC curve $\{(\hat{\tau}_{\mathrm{FPR}}(t), \hat{\tau}_{\mathrm{TPR}}(t))\}$.

D.3 Zero-run setting

In the zero-run setting, the estimators of the different causal evaluation metrics are now weighted with propensity scores as follows:

$$\begin{aligned}
\hat{\tau}^{(\mathrm{IPW})}_{\mathrm{ATE}} &= \frac{1}{n_1}\sum_{A_i=1} Y_i - \frac{1}{n_0}\sum_{A_i=0}\frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}Y_i\,,\\
\hat{\tau}^{(\mathrm{IPW})}_{\mathrm{TPR}}(t) &= \frac{1}{n_1}\sum_{A_i=1} \mathbb{1}_{\{Y_i \geqslant t\}}\\
\hat{\tau}^{(\mathrm{IPW})}_{\mathrm{FPR}}(t) &= \frac{1}{n_1}\sum_{A_i=0}\frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\mathbb{1}_{\{Y_i \geqslant t\}}\\
\hat{\tau}^{(\mathrm{IPW})}_{\mathrm{TPR@FPR}}(\alpha) &= \frac{1}{n_1}\sum_{A_i=1} \mathbb{1}_{\{Y_i \geqslant \hat{t}_\alpha\}}
\end{aligned} \tag{7}$$

with

$$\hat{t}_\alpha \in \operatorname{argmax}\left\{ t \,:\, \frac{1}{n_0}\sum_{A_i=1}\frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\mathbb{1}_{\{Y_i \geqslant t\}} \geqslant 1-\alpha \right\}\,,$$

where $\hat{\pi} : \mathcal{X} \to [0,1]$ is an estimator of the propensity score function π, learned on independent data. The estimator of the causal AUC is still the area under the estimated ROC curve. Next proposition then generalizes the identification result established in Proposition 6.2 to the other causal evaluation metrics.

Proposition D.2. *Under Assumption 6.1, if $(X_i, A_i, Y_i)_{i\in[n]}$ are obtained according to Algorithm 3, the causal evaluation metrics are identifiable. Let the propensity score π write as:*

$$\forall x \in \mathcal{X}\,, \quad \pi(x) \overset{def}{=} \mathbb{P}\left(A_i = 1 | X_i = x\right)\,.$$

We have:

$$\begin{aligned}
\tau_{\mathrm{ATE}} &= \mathbb{E}\left[A_i Y_i\right]\mathbb{P}\left(A_i=1\right)^{-1}\\
&\quad - \mathbb{E}\left[\frac{\pi(X_i)(1-A_i)}{1-\pi(X_i)}Y_i\right]\mathbb{P}\left(A_i=1\right)^{-1}\,,\\
\tau_{\mathrm{TPR}}(t) &= \mathbb{E}\left[A_i \mathbb{1}_{\{Y_i \geqslant t\}}\right]\mathbb{P}\left(A_i=1\right)^{-1}\\
\tau_{\mathrm{FPR}}(t) &= \mathbb{E}\left[\frac{\pi(X_i)(1-A_i)}{1-\pi(X_i)}\mathbb{1}_{\{Y_i \geqslant t\}}\right]\mathbb{P}\left(A_i=1\right)^{-1}\,.
\end{aligned}$$

E G-Formula and AIPW beyond the ATE

Letting $\hat{\mu}_0$ be a regression model that estimates conditional MIA response $\mu_0 : x \in \mathcal{X} \mapsto \mathbb{E}\left[Y_i(0) | X_i = x\right]$, the G-Formula estimator of the ATE (which is in fact an ATT) is defined in Equation (4). Outcome regression model $\hat{\mu}_0$ is typically learned on an independent dataset or using crossfitting, by fitting a model $Y_i \sim X_i$ on non-member datapoints (satisfying $A_i = 0$). We now generalize this approach to the causal TPR and FPR, in order to obtain a causal ROC curve and a causal AUC, without having to resort to propensity scores.

TPR. For the causal TPR, the G-Formula estimator $\hat{\tau}^{(\mathrm{G})}_{\mathrm{TPR}}(t)$ is unchanged (same as $\hat{\tau}^{(\mathrm{IPW})}_{\mathrm{TPR}}(t)$ in Equation (7)). The same for the AIPW estimator.

FPR. For the causal TPR, the G-Formula estimator $\hat{\tau}^{(\mathrm{G})}_{\mathrm{FPR}}(t)$ requires to learn a family of regressors $\{\hat{\mu}_{0,t}, t \in \mathbb{R}\}$ where $\hat{\mu}_{0,t}$ is an estimator of:

$$\mu_{0,t} : x \in \mathcal{X} \mapsto \mathbb{P}\left(Y_i(0) \geqslant t | X_i = x\right) .$$

Since $\mu_{0,t}(x) = \mathbb{P}\left(Y_i \geqslant t | X_i = x, A_i = 0\right)$, $\hat{\mu}_{0,t}$ can be learned by using a classifier and outputting conditional probabilities, that we fit on the non-members $\left\{(X_i, \mathbb{1}_{\{Y_i \geqslant t\}}) | A_i = 0\right\}$. This needs to be done for all $t \in \mathbb{R}$ or for a discretized set. The causal FPR at threshold t can the be estimated using:

$$\hat{\tau}^{(\mathrm{G})}_{\mathrm{FPR}}(t) = \frac{1}{n_1} \sum_{A_i = 1} \hat{\mu}_{0,t}(X_i) .$$

Another approach would be to use *distributional regression*, which is typical of comparison-based estimands in causal inference [20]. An estimator $\hat{\mathbb{P}}_0(x, \cdot)$ of the probability distribution $\mathbb{P}_{Y|X_i=x, A_i=0}$ could be learn, where $\hat{\mathbb{P}}(x, t)$ is an estimator of $\mathbb{P}\left(Y_i \geqslant t | X_i = x\right)$.

Finally, the AIPW estimator of the causal FPR writes as:

$$\hat{\tau}^{(\mathrm{AIPW})}_{\mathrm{FPR}}(t) = \frac{1}{n_1} \sum_{i=1}^{n} \left[A_i \hat{\mu}_{0,t}(X_i) + \frac{(1 - A_i)\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \left(\mathbb{1}_{\{Y_i \geqslant t\}} - \hat{\mu}_{0,t}(X_i) \right) \right]$$

F Proofs in the 1-Run Randomized setting

In all the proofs, everything is done conditionally on the initial data $\mathcal{D}$, and we leave the conditioning with respect to $\mathcal{D}$ implicit. For a vector $\bar{a} \in \{0,1\}^n$, we let $n_1(\bar{a}) = \#\{i \,:\, a_i = 1\}$ and similarly for $n_0(\bar{a})$. In particular $n_1(\bar{A}) = n_1$. We'll use the fact that, on an event of probability at least $1 - 4e^{-t}$, it holds

$$\left| n_a - \frac{n}{2} \right| \leqslant \sqrt{\frac{tn}{2}} \quad \text{for} \quad a \in \{0,1\}.$$

This is a straightforward application of Hoeffding's inequality. In particular, for $t \leqslant n/8$, it holds on the same event $\frac{n}{4} \leqslant n_a \leqslant \frac{3n}{4}$ for both $a \in \{0,1\}$.

F.1 Proof of Theorem 5.4

F.1.1 Proof with perfect interpolation

Proof. First, note that thanks to the interpolation assumption,

$$\frac{1}{n_1} \sum_{i : A_i = 1} Y_i = 0 = \mathbb{E}\left[\mathbb{E}\left[Y_i | A_i = 1, X_i\right]\right] .$$

We thus need to work only with $\frac{1}{n_0} \sum_{i:A_i=0} Y_i$. We let $a \in \{0,1\}^n$ with $n_0(\bar{a}) \neq 0$, and define, letting $\overline{X}_{\bar{a}} := \{X_j\}_{\{j:\, a_j = 1\}}$,

$$\mu^{(0)}_{\bar{a}}(\overline{X}_{\bar{a}}) := \mathbb{E}\left[Y_i | \bar{A} = \bar{a}, \overline{X}_{\bar{a}}\right],$$

for any i such that $a_i = 0$ (we also integrate wrt the randomness of $\mathcal{A}$ in the expectation, after conditioning on $\mathcal{A}$ to apply Hoeffding). Thanks to Hoeffding's inequality, it holds

$$\mathbb{P}\left(\left| \frac{1}{n_0(\bar{a})} \sum_{i : a_i = 0} Y_i - \mu^{(0)}_{\bar{a}}(\overline{X}_{\bar{a}}) \right| > \sqrt{\frac{t}{2n_0(\bar{a})}} \,\middle|\, \overline{A} = \bar{a}, \overline{X}_{\bar{a}} \right) \leqslant 2e^{-t} .$$

Now for two adjacent datasets $\overline{X}_{\bar{a}}$ and $\overline{X}'_{\bar{a}}$, it holds:

$$|\mu^{(0)}_{\bar{a}}(\overline{X}_{\bar{a}}) - \mu^{(0)}_{\bar{a}}(\overline{X}'_{\bar{a}})| = |\mathbb{E}_{\mathcal{A}}\mathbb{E}_{\mathcal{P}_T}[\mathcal{L}(\mathcal{A}(\mathcal{D} \cup \overline{X}_{\bar{a}}), X)] - \mathbb{E}_{\mathcal{A}}\mathbb{E}_{\mathcal{P}_T}[\mathcal{L}(\mathcal{A}(\mathcal{D} \cup \overline{X}'_{\bar{a}}), X)]| \leqslant \alpha,$$

thanks to the average stability assumption (which implicitly encompasses adjacency for *Replace* and *Add/Remove*). Thus, using McDiarmid's inequality, it holds

$$\mathbb{P}\left(|\mu_{\bar{a}}^{(0)}(\overline{X}_{\bar{a}}) - \mathbb{E}[Y_i|\overline{A} = \bar{a}]| > \sqrt{\frac{n_1(\bar{a})t\alpha^2}{2}} \,\middle|\, \overline{A} = \bar{a}\right) \leqslant 2e^{-t}.$$

Thus

$$\mathbb{P}\left(|\hat{\tau}_{\mathrm{ATE}} - \mathbb{E}[Y_i|\overline{A} = \bar{a}]| \leqslant \sqrt{\frac{2t}{n_0(\bar{a})}} + \sqrt{\frac{n_1(\bar{a})t\alpha^2}{2}} \,\middle|\, \overline{A} = \bar{a}\right) \geqslant 1 - 4e^{-t}.$$

We conclude the proof by introducing

$$g(\bar{a}_{-i}) := \mathbb{E}[Y_i|A_i = 0, \overline{A}_{-i} = a_{-i}].$$

The previous bounds allow for a control of $|\hat{\tau}_{\mathrm{ATE}} - g(\bar{A}_{-i})|$. Using again the α-stability, one find that for two adjacent vectors $\bar{A}_{-i}, \bar{A}'_{-i}$, it holds $|g(\bar{A}_{-i}) - g(\bar{A}'_{-i})| \leqslant \alpha$. Using McDiarmid's inequality again, we conclude

$$\mathbb{P}\left(|g(\overline{A}_{-i}) - \mathbb{E}[Y_i|A_i = 0]| > \sqrt{\frac{(n-1)t\alpha^2}{2}}\right) \leqslant 2e^{-t}.$$

Piecing all the inequalities together yields the result.

□

F.1.2 Proof with β-uniform training stability

Proof. The only difference lies in the way we handle the term

$$\frac{1}{n_1}\sum_{i:A_i=1} Y_i = \frac{1}{n_1}\sum_{i:A_i=1} \mathcal{L}(\mathcal{A}(\mathcal{D}_{\mathrm{train}}), X_i).$$

Like before, we let $\bar{a} \in \{0,1\}^n$ with $n_1(\bar{a}), n_0(\bar{a}) \neq 0$. We define

$$\mu_{\bar{a}}^{(1)}(\overline{X}_{\bar{a}}) := \frac{1}{n_1(\bar{a})}\sum_{i:a_i=1} \mathcal{L}(\mathcal{A}(\mathcal{D} \cup \overline{X}_{\bar{a}}), X_i)$$

Thanks to β-uniform training stability, it holds that for any adjacent $\overline{X}'_{\bar{a}}$:

$$\begin{aligned}|\mu_{\bar{a}}^{(1)}(\overline{X}_{\bar{a}}) - \mu_{\bar{a}}^{(1)}(\overline{X}'_{\bar{a}})| &\leqslant \frac{1}{n_1(\bar{a})}\sum_{i:a_i=1} |\mathcal{L}(\mathcal{A}_\theta(\mathcal{D} \cup \overline{X}_{\bar{a}}), X_i) - \mathcal{L}(\mathcal{A}_\theta(\mathcal{D} \cup \overline{X}'_{\bar{a}}), X'_i)| \\ &\leqslant \frac{1}{n_1(\bar{a})} + \beta.\end{aligned}$$

Using again McDiarmid's inequality, we find that

$$\mathbb{P}\left(\left|\mu_{\bar{a}}^{(1)}(\overline{X}_{\bar{a}}) - \mathbb{E}[Y_i|\overline{A} = \bar{a}]\right| > \left(\frac{1}{n_1(\bar{a})} + \beta\right)\sqrt{\frac{n_1(\bar{a})t}{2}} \,\middle|\, \overline{A} = \bar{a}\right) \leqslant 2e^{-t},$$

where i is such that $a_i = 1$. Like in the previous case, we show that, using this time β-uniform training stability

$$\mathbb{P}\left(|\mathbb{E}[Y_i|A_i = 1, \overline{A}_{-i}] - \mathbb{E}[Y_i|A_i = 1]| > \sqrt{\frac{(n-1)t\beta^2}{2}}\right) \leqslant 2e^{-t}.$$

which allows to conclude. □

F.2 Proof of Theorem C.1

Proof. Using the same reasoning as in the proof with 0-uniform training stability, we can handle the term with $A_i = 0$. For $\bar{a} \in \{0,1\}^n$, we let $\theta \in \Theta$ be a generic model parameter, and $\hat{\theta}_{\bar{a}}$ be the one obtained after training on $\mathcal{D} \cup \overline{X}_{\bar{a}}$. We define

$$f_\theta(X) := \mathcal{L}(\theta, X),$$

so that in particular $Y_i = f_{\hat{\theta}_{\bar{a}}}(X_i)$ on the event $\overline{A} = \bar{a}$. Note that $\hat{\theta}_{\bar{a}}$ is $\overline{X}_{\bar{a}}$-measurable. Because $\theta \mapsto f_\theta(x)$ is L-Lipschitz for all $x \in \mathcal{X}$, the empirical process

$$\theta \mapsto \xi_{\bar{a}}(\theta) := \frac{1}{n_1(\bar{a})} \sum_{a_i=1} f_\theta(X_i) - \mathbb{E}_{\mathcal{P}_T}[f_\theta(X)]$$

is $2L$-Lipschitz and centered. Letting $\mathcal{N}_\varepsilon$ be a minimal ε-net over $\mathrm{B}_{\mathbb{R}^D}(0, B)$, if holds that $|\mathcal{N}_\varepsilon| \leqslant (2B/\varepsilon)^D$, hence

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |\xi_{\bar{a}}(\theta)| > s \;\middle|\; \overline{A} = \bar{a}\right) \leqslant 2\left(\frac{2B}{\varepsilon}\right)^D \exp\left(-\frac{n_1(\bar{a})(s - 2L\varepsilon)^2}{2}\right).$$

Letting

$$s = \sqrt{\frac{2\lambda_t}{n_1(\bar{a})}} + 2L\varepsilon \quad \text{and} \quad \varepsilon = \frac{1}{2L}\sqrt{\frac{2\lambda_t}{n_1(\bar{a})}} \quad \text{with} \quad \lambda_t := t + \frac{1}{2}D\log n_1(\bar{a}) + D\log(4BL),$$

one finds

$$\mathbb{P}\left(\left|\frac{1}{n_1(\bar{a})}\sum_{a_i=1} Y_i - \mathbb{E}_{\mathcal{P}_T}[f_{\hat{\theta}_{\bar{a}}}(X)|\overline{X}_{\bar{a}}, \overline{A} = \bar{a}]\right| > 2\sqrt{\frac{2\lambda_t}{n_1(\bar{a})}} \;\middle|\; \overline{X}_{\bar{a}}, \overline{A} = \bar{a}\right) \leqslant 2e^{-t}.$$

Now notice that $\mathbb{E}_{\mathcal{P}_T}[f_{\hat{\theta}_{\bar{a}}}(X)|\overline{X}_{\bar{a}}, \overline{A} = \bar{a}] = \mu_{\bar{a}}^{(0)}(\overline{X}_{\bar{a}})$. Using a previous bound, we show that

$$\mathbb{P}\left(\left|\frac{1}{n_1(\bar{a})}\sum_{a_i=1} Y_i - \frac{1}{n_0(\bar{a})}\sum_{a_i=0} Y_i\right| > 2\sqrt{\frac{2\lambda_t}{n_1(\bar{a})}} + \sqrt{\frac{t}{2n_0(\bar{a})}} \;\middle|\; \overline{X}_{\bar{a}}, \overline{A} = \bar{a}\right) \leqslant 4e^{-t}.$$

For this we conclude easily, since $|\hat{\tau}_{\mathrm{ATE}}|$ is upper-bounded by 1:

$$|\tau_{\mathrm{ATE}}| = |\mathbb{E}[\hat{\tau}_{\mathrm{ATE}}]| \leqslant \mathbb{E}[|\hat{\tau}_{\mathrm{ATE}}|] \leqslant \varepsilon + \mathbb{P}(|\hat{\tau}_{\mathrm{ATE}}| > \varepsilon), \quad \text{for all } \varepsilon > 0.$$

□

G Proofs in the observational setting

Lemma G.1. *Let*

$$\bar{\tau}_{\mathrm{ATE}} := \frac{1}{n_1}\sum_{A_i=1} Y_i - \frac{1}{n_0}\sum_{A_i=0} \frac{\pi(X_i)}{1-\pi(X_i)} Y_i\,.$$

It holds, with probability $1 - 2e^{-t}$*:*

$$|\hat{\tau}_{\mathrm{ATE}} - \bar{\tau}_{\mathrm{ATE}}| \leq \Delta_{\hat{\pi}} + \frac{1}{\eta}\sqrt{\frac{t}{n}}\,.$$

Proof. We have:

$$\begin{aligned}
|\hat{\tau}_{\mathrm{ATE}} - \bar{\tau}_{\mathrm{ATE}}| &= \left|\frac{1}{n_0}\sum_{A_i=0}\left\{\frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)} - \frac{\pi(X_i)}{1-\pi(X_i)}\right\} Y_i\right| \\
&\leqslant \Delta_{\hat{\pi}} + \underbrace{\frac{1}{n_0}\sum_{A_i=0}\left|\frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)} - \frac{\pi(X_i)}{1-\pi(X_i)}\right| - \Delta_{\hat{\pi}}}_{:=\xi}\,.
\end{aligned}$$

Now like in the previous section, we control ξ by conditioning on $\bar{A} = \bar{a}$, applying Höeffding's inequality for rvs in $[0, C/\eta]$, and finally use the fact that $n_0 \geqslant n/4$ w.h.p. □

All the proofs in the observational setting mimic the structures of those of the previous section. Thanks to Lemma G.1, one only needs to bound $|\bar{\tau}_{\mathrm{ATE}} - \tau_{\mathrm{ATE}}|$. To ease the notation, we introduce $\zeta(X_i) := \frac{\pi(X_i)}{1-\pi(X_i)}$. As the term involving the Y_i's for $A_i = 1$ is the same as in the previous section, we only focus on bounding

$$\frac{1}{n_0} \sum_{A_i=0} \zeta(X_i) Y_i - \mathbb{E}[\zeta(X_i) Y_i | A_i = 0].$$

The bound is the same for each settings and the proof goes as follows. We let again $a \in \{0,1\}^n$ with $n_0(\bar{a}) \neq 0$, and define, letting $\overline{X}_{\bar{a}} := \{X_j\}_{\{j:\, a_j=1\}}$,

$$\mu^{(0)}_{\bar{a}}(\overline{X}_{\bar{a}}) := \mathbb{E}\left[\zeta(X_i) Y_i | \bar{A} = \bar{a}, \overline{X}_{\bar{a}}\right],$$

for any i such that $a_i = 0$. Thanks to Hoeffding's inequality, it holds

$$\mathbb{P}\left(\left| \frac{1}{n_0(\bar{a})} \sum_{i:a_i=0} \zeta(X_i) Y_i - \mu^{(0)}_{\bar{a}}(\overline{X}_{\bar{a}}) \right| > \sqrt{\frac{t}{2 n_0(\bar{a}) \eta^2}} \,\middle|\, \overline{A} = \bar{a}, \overline{X}_{\bar{a}} \right) \leqslant 2e^{-t}.$$

Now for two adjacent datasets $\overline{X}_{\bar{a}}$ and $\overline{X}'_{\bar{a}}$, it holds:

$$\begin{aligned}
|\mu^{(0)}_{\bar{a}}(\overline{X}_{\bar{a}}) - \mu^{(0)}_{\bar{a}}(\overline{X}'_{\bar{a}})| &= |\mathbb{E}_{\mathcal{A}} \mathbb{E}_{\mathcal{P}_0}[\zeta(X) \mathcal{L}(\mathcal{A}(\mathcal{D} \cup \overline{X}_{\bar{a}}), X)] - \mathbb{E}_{\mathcal{A}} \mathbb{E}_{\mathcal{P}_0}[\zeta(X) \mathcal{L}(\mathcal{A}(\mathcal{D} \cup \overline{X}'_{\bar{a}}), X)]| \\
&= |\mathbb{E}_{\mathcal{A}} \mathbb{E}_{\mathcal{P}_T}[\mathcal{L}(\mathcal{A}(\mathcal{D} \cup \overline{X}_{\bar{a}}), X)] - \mathbb{E}_{\mathcal{A}} \mathbb{E}_{\mathcal{P}_T}[\mathcal{L}(\mathcal{A}(\mathcal{D} \cup \overline{X}'_{\bar{a}}), X)]| \\
&\leqslant \alpha,
\end{aligned}$$

thanks to the average stability assumption. Thus, using McDiarmid's inequality, it holds

$$\mathbb{P}\left(|\mu^{(0)}_{\bar{a}}(\overline{X}_{\bar{a}}) - \mathbb{E}[\zeta(X_i) Y_i | \overline{A} = \bar{a}]| > \sqrt{\frac{n_1(\bar{a}) t \alpha^2}{2}} \,\middle|\, \overline{A} = \bar{a} \right) \leqslant 2e^{-t}.$$

Thus

$$\mathbb{P}\left(|\bar{\tau}_{\mathrm{ATE}} - \mathbb{E}[\zeta(X_i) Y_i | \overline{A} = \bar{a}]| \leqslant \sqrt{\frac{2t}{n_0(\bar{a}) \eta^2}} + \sqrt{\frac{n_1(\bar{a}) t \alpha^2}{2}} \,\middle|\, \overline{A} = \bar{a} \right) \geqslant 1 - 4e^{-t}.$$

We conclude the proof by introducing

$$g(\bar{a}_{-i}) := \mathbb{E}[\zeta(X_i) Y_i | A_i = 0, \overline{A}_{-i} = a_{-i}].$$

The previous bounds allow for a control of $|\bar{\tau}_{\mathrm{ATE}} - g(\bar{A}_{-i})|$. Using again the α-stability, one find that for two adjacent vectors $\bar{A}_{-i}, \bar{A}'_{-i}$, it holds $|g(\bar{A}_{-i}) - g(\bar{A}'_{-i})| \leqslant \alpha$. Using McDiarmid's inequality again, we conclude

$$\mathbb{P}\left(|g(\overline{A}_{-i}) - \mathbb{E}[\zeta(X_i) Y_i | A_i = 0]| > \sqrt{\frac{(n-1) t \alpha^2}{2}} \right) \leqslant 2e^{-t}.$$

Piecing all the inequalities together yields the result.

H Additional Experimental Analysis

H.1 Additional details on synthetic data experiment

Data generation process. Training, member and non-member data is generated as follows. Training datapoints $X_i = (a_i, b_i)$ are generated with $a_i \sim \mathcal{N}(0, I_d)$ and $b_i \sim \mathcal{N}(a_i^\top w_\star, 1) | a_i$, for some random fixed unitary vector $w_\star \in \mathbb{R}^d$. This distribution corresponds to $\mathcal{P}_T$. Non-member datapoints taken from the same distribution (no distribution shift), or taken with $a_i \sim \mathcal{N}(\mu, I_d)$ for some $\mu \in \mathbb{R}^d$, depending on the setting (inducing distribution shift between members and non-members). This shifted distribution corresponds to $\mathcal{P}_0$. We consider linear regression with the square loss on such data, and consider the loss-based MIA (that thresholds loss values).

Simulation parameters. Data simulation parameters are as follows:

1. Dimension d, which is not the same in the two training algorithms considered next;
2. Training dataset of size 2000, made of samples from $\mathcal{P}_T$;
3. μ is a random unitary vector;
4. $w_\star$ is a random standard Gaussian Vector conditioned on having $w_\star^\top \mu / \|\mu\| = 0.90$ (correlation between $w_\star$ and μ).

Training algorithms $\mathcal{A}$. We consider the two different algorithms to attack for the MIA:

1. We first consider a setting where our assumptions hold and for which MIAs should have good-to-moderate performances: overparametrized ridge regression, with $\lambda = 10^3$ for regularization parameter, with an ambiant dimension $d = 2500$. Definitions 5.2 and 5.3 will hold in this setting due to uniform stability.
2. We then consider a setting for which MIAs should not be able to detect any signal and for which our assumptions hold. We thus consider a gradient-based optimizer that adds Gaussian noise at each step and clips gradient of norm larger than 1, therefore limiting sensitivity of each stochastic gradient batch. We add ℓ^2 regularization with $\lambda = 10$, and $d = 400$ in this setting. The optimizer is DP-SGD, with Gaussian noise $\mathcal{N}(0, 3I_d)$, learning rate of 0.01, 75 epochs and batchsize of 128. This corresponds to DP parameters $(\varepsilon, \delta) = (0.602, 0.01)$. Stability will hold in this setting, while interpolation will not (due to DP-SGD noise). β-uniform training stability will hold, as a consequence of uniform stability.

MIA evaluation settings. We consider the three different MIA evaluation settings, corresponding to Algorithms 1 to 3.

1. *Multi Run setting*: 400 training of each algorithm to attack. For each run, the training dataset is made of 2000 samples from $\mathcal{P}_T$ (that stay fixed accross all runs), together with another independent sample from $\mathcal{P}_T$ for the first 200 runs (members) and none for the last 200 runs. 200 non-members datapoints are then sampled from $\mathcal{P}_T$.
2. $1-$*Run setting*, with only one training run. 2000 samples are drawn from $\mathcal{P}_T$, forming the training set on which the training algorithm is runned. These 2000 training points are the member dataset. 2000 non-members are then independently sampled from the same distribution $\mathcal{P}_T$;
3. $0-$*Run setting*: training set and members are the same as in the $1-$Run setting. The 2000 non-members are here sampled from thethe shifted distribution $\mathcal{P}_0$.

In the multi-run setting there are 200 members and 200 non-members, while in the one-run and zero-run settings there are 2000 of each.

Results. Results for all metrics are reported in Table 2, ROC curves in Figure 3 (A-B). *Raw* metrics mean that they were not adjusted by propensity scores while *Corrected* means that they were (if not specified, correction with true propensity scores, *Learned*: correction with propensity scores estimated via logistic regression). Raw metrics in the 1-Run and 0-Run settings are thus the causal metrics, as well as the corrected ones in the zero-run setting. The raw metrics in the zero-run setting are *not* causal metrics, and our experiments highlight the bias induced by not taking into account distribution shifts. Confidence intervals are made over 100 runs of independent samples from the same data-generating process.

Interpretation. In both settings (DP-SGD, Ridge), Figure 3 (A-B) shows the ROC curves of the Multi-Run and One-Run settings, that don't have any distribution shift (resp. blue and orange). Then, the ROC curves of the Zero-run setting are shown without correction (green) and with propensity score correction (red). In all these plots, the non-drifted multi-run and one-run ROC curves are essentially the same: interferences are negligible in this case. Due to distribution shift, the zero-run curve shows easier attack surfaces. This is actually a false interpretation since after correction, the corrected zero-run curve matches that of the non-drifted curves. As expected, MIAs are more performant in the Ridge setting than the DP-SGD setting.

In Figure 3 (B), we plotted the $y = \min(1, e^{\varepsilon}x + \delta, 1 - e^{-\varepsilon}(1 - \delta - x))$ curve. Under $(\varepsilon, \delta)-$DP, a MIA cannot have a ROC curve that crosses that black line in the multi-run setting [16]. Here, we observe that not adjusting for distribution shift breaks this and yields a ROC curve above this line.

The metrics in Table 2 show the same results through four different causal metrics introduced in Section 3.2. Correction to target the right causal metrics make the zero-run performances match the non-drifted settings. For all metrics and in all settings, the IPW correction makes the causal metrics all align together.

Model	Evaluation	AUC	$\sup_t \tau_{\mathrm{ATE}}(t)$	ATE	TPR0.2
Ridge	Multi-run (Raw)	0.561 ± 0.016	0.102 ± 0.023	-287.694 ± 248.990	0.241 ± 0.027
Ridge	One-run (Raw)	0.561 ± 0.019	0.103 ± 0.027	-293.668 ± 189.226	0.243 ± 0.031
Ridge	Zero-run (Raw)	0.669 ± 0.020	0.269 ± 0.031	-1049.710 ± 138.808	0.350 ± 0.041
Ridge	Zero-run (Corrected, Oracle)	0.562 ± 0.026	0.108 ± 0.036	-297.327 ± 151.472	0.201 ± 0.068
Ridge	Zero-run (Corrected, Learned)	0.595 ± 0.025	0.159 ± 0.034	-496.127 ± 106.100	0.228 ± 0.083
DP-SGD	Multi-run (Raw)	0.504 ± 0.018	0.024 ± 0.024	0.633 ± 50.951	0.207 ± 0.024
DP-SGD	One-run (Raw)	0.503 ± 0.017	0.023 ± 0.020	-0.949 ± 46.320	0.205 ± 0.022
DP-SGD	Zero-run (Raw)	0.616 ± 0.016	0.183 ± 0.026	-322.243 ± 41.150	0.299 ± 0.032
DP-SGD	Zero-run (Corrected, Oracle)	0.503 ± 0.024	0.029 ± 0.026	-1.985 ± 68.971	0.165 ± 0.083
DP-SGD	Zero-run (Corrected, Learned)	0.512 ± 0.059	0.063 ± 0.084	-23.022 ± 158.351	0.174 ± 0.111

Table 2: Causal metrics in the three settings (with oracle or learned propensity scores in the zero-run setting), and non-causal metrics in the zero-run setting to highlight the distribution-shift induced bias. $\tau_{\mathrm{ATE}}(t)$ is defined as $\tau_{\mathrm{TPR}}(t) - \tau_{\mathrm{FPR}}(t)$ (Youden's J metric at a given threshold), and $\sup_t \tau_{\mathrm{ATE}}(t)$ is thus the largest gap between (causal) TPR and FPR. TPR0.2 is the TPR at FPR $\alpha = 0.2$.

H.2 Additional details on CIFAR10

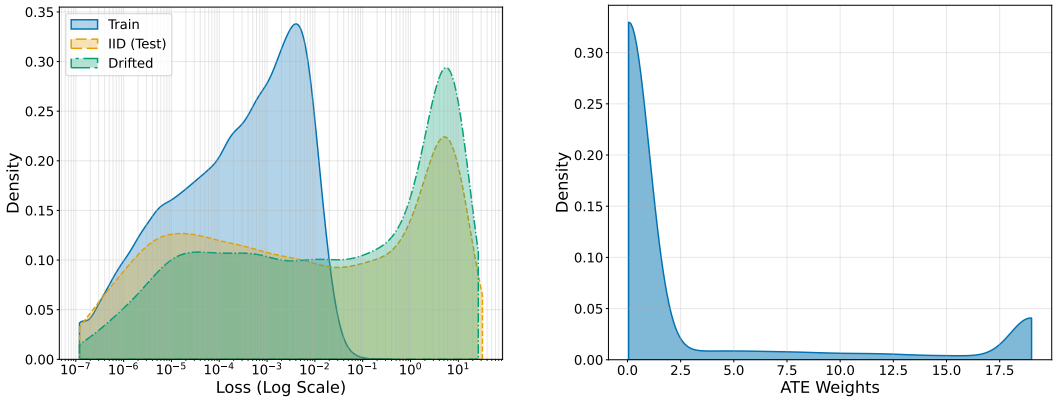
We now present in more details our CIFAR10 experimental setup.

Dataset Construction. We define two base domains: the original clean CIFAR-10 data ($\mathcal{D}_{\text{clean}}$) and a corrupted domain ($\mathcal{D}_{\text{noise}}$) generated by adding Gaussian noise with variance $\sigma = 0.2$ to the images. To ensure valid overlap, we construct the datasets as follows: i) The Members ($\mathcal{D}_T$) training set is constructed such that 90% of samples are drawn from $\mathcal{D}_{\text{clean}}$ and 10% from $\mathcal{D}_{\text{noise}}$ and ii) The non-members ($\mathcal{D}_0$) are drawn from a distribution of 10% samples from $\mathcal{D}_{\text{clean}}$ and 90% from $\mathcal{D}_{\text{noise}}$. This design creates a covariate shift: members are predominantly clean, while non-members are predominantly noisy, yet it guarantees that the density ratio remains bounded, since every sample has a non-zero probability of appearing in both sets.

Target Model. On this dataset, we train a ResNet-16 [25] architecture trained on. The model is optimized using Stochastic Gradient Descent (SGD) with a batch size of 128 and a fixed learning rate of 0.01. Training is performed for 10000 optimization steps. To isolate the effects of distribution shift and memorization without additional confounders, we train the model without any data augmentation.

Propensity Score Estimation. To correct for this distribution shift, we estimate the propensity score $\pi(x) = \mathbb{P}(A = 1|X = x)$. We treat this as a binary classification problem, distinguishing between the member distribution $\mathcal{D}_T$ and the non-member distribution $\mathcal{D}_0$. We fine-tune a ResNet model [25] pre-trained on ImageNet [15] to minimize the binary cross-entropy loss. We perform this estimation using cross-validation: the propensity model is fitted on a held-out portion of the data mixtures, ensuring that the estimated weights $\hat{\pi}(X_i)$ are independent of the MIA scores Y_i used for the final evaluation.

Results. We compare the performance of Membership Inference Attacks across the three evaluation regimes. Figure 3 (C) displays the ROC curves for the loss-based attack. The *IID (Test)* curve (blue) represents the idealized setting where $\mathcal{D}_0$ follows the same distribution as the training set ($\mathcal{D}_T$). The *Drifted* curve (orange) shows the naive zero-run evaluation where non-members are drawn from the shifted (noisy) distribution. This induces an inflation in attack performance (AUC ≈ 0.75), misleadingly suggesting high memorization. However, this advantage is spurious: it stems from the model's high loss on noisy non-members rather than true privacy leakage. By applying our propensity score correction (green curve), we recover the true causal memorization signal (AUC ≈ 0.66), which aligns closely with the valid IID baseline (AUC ≈ 0.65). This confirms that our method successfully removes the bias.



(a) Loss Distributions (b) Propensity Weights

Figure 5: **Additional metrics. (a)** The loss distribution for drifted non-members is shifted, making them more distinguishable from members without true memorization. **(b)** The distribution of estimated ATE weights shows how our method corrects this by upweighting the minority of non-members that are distributionally similar to the training set.

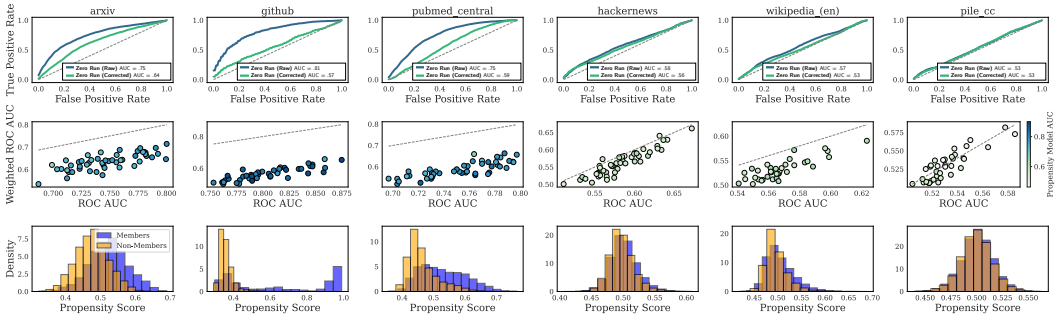


Figure 6: Results of the MIA experiment on the `Pythia-12b` model. The **first row** displays the ROC curve for 6 different datasets corresponding to the **columns**. The **second row** compares the unweighted (x-axis) to the weighted ROC AUC (y-axis). Every dot represents a different test-train split of the MIA data, the color indicating the AUC of the propensity score model. Lastly, the **third row** shows the distribution of the propensity score, which is a proxy for distribution shift between members and non-members.

To better understand the source of bias in the zero-run setting, we analyze the distributions of losses and the estimated propensity weights for the CIFAR10 case. Figure 5a visualizes the density of loss values for training members (Train), IID test non-members, and Drifted non-members. The *Drifted* distribution (orange) is shifted significantly to the right (higher loss) compared to the *IID* distribution (blue). Because standard MIAs predict membership based on low loss, the separation between the low-loss members and high-loss drifted non-members is artificially large, leading to the inflated AUC observed in the main text. The overlap between the Train and IID Test distributions is bigger, reflecting the attack's true performance.

Propensity Weights. Figure 5b shows the distribution of the inverse probability weights (IPW) applied to the non-member samples during our correction. The distribution is heavy-tailed, assigning high weights to the rare non-members that resemble the training distribution (low noise) and low weights to the abundant noisy samples.

H.3 Additional Details on the Pythia LLM Experiment

Datasets and Model. Building on Meeus et al. [50], we consider the Pythia suite of models, specifically the `12b` parameters model. This model been trained on subsets of the MIMIR dataset

Dataset	Members	Non-members
arxiv	1000	500
github	1000	268
hackernews	1000	646
pubmed_central	1000	491
pile_cc	1000	1000
wikipedia_(en)	1000	1000

Table 3: Number of members and non-members for every corpus of MIMIR.

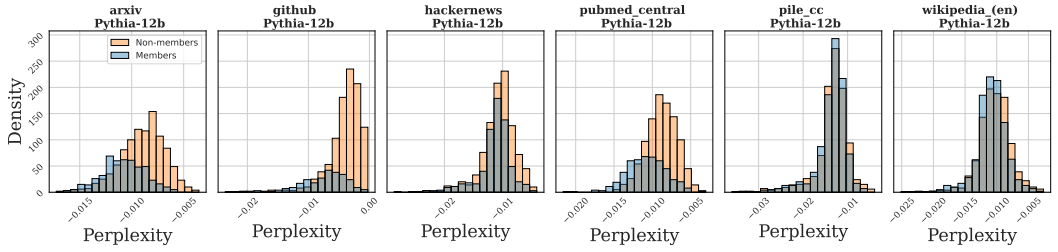


Figure 7: Perplexity for members and non-members for `Pythia-12b`.

with information about members and non-members being openly available. We consider specifically 6 subsets for which membership information is available (see Table 3 for details on these datasets).

Inference of the Propensity Score. To estimate the propensity score $\pi(\cdot)$ for every text, we follow Meeus et al. [50] and use a bag-of-words tokenizer in combination with a random forest classifier. The classifier is trained on 80% of the data, and we use the remaining 20% for evaluation. Note that we balance all datasets prior to training to have the same number of members and non-members. Propensity scores are clipped to the range $[0.01, 0.99]$ to avoid numerical errors. Exact details of the tokenizer and the propensity score model are given in Table 4.

Membership attacks. On the testing split, we query the perplexity of the model p_i for every datapoint X_i. We then compute the ROC curve for both the standard and reweighted perplexity.

H.4 Additional Details on White-House Press Release LLM Experiment

Data scrapping. We scrap the titles of all press statements made by the Biden and Trump II administration. For the Biden administration, we scrap all titles starting on the the 15th of January 2021 from the website `https://bidenwhitehouse.archives.gov`. For the Trump II administration, we scrap all titles starting from the 20th of January 2025 from `https://www.whitehouse.gov/news/`. We will make this data publicaly available. A 2d-PCA of the data is displayed in Figure 9 and shows a clear distribution shift between the two corpuses. We additionaly display a Umap in Figure 10 along with the most frequent words in every cluster.

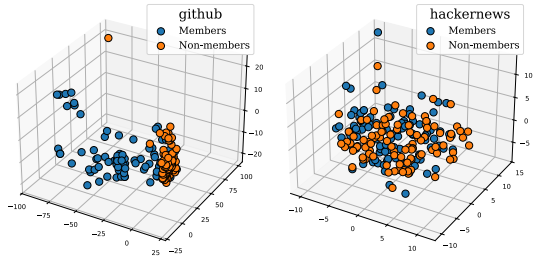


Figure 8: Mean-pooled embbedings produced by `Pythia-70m` of elements of `github` v.s. `hackernews`.

Parameter	MIMIR	Trump-Biden
`Tokenizer`	`CountVectorizer`	—
`min_df`	0.01	0.13
`max_df`	1	—
`max_features`	1000	—
`Propensity Score Model`	`RandomForest`	—
`n_estimators`	100	—
`max_depth`	5	—
`min_samples_leaf`	5	—

Table 4: Parameters for propensity score inference on the MIMIR and Trump-Biden datasets. We report two values for parameters only if they differ.

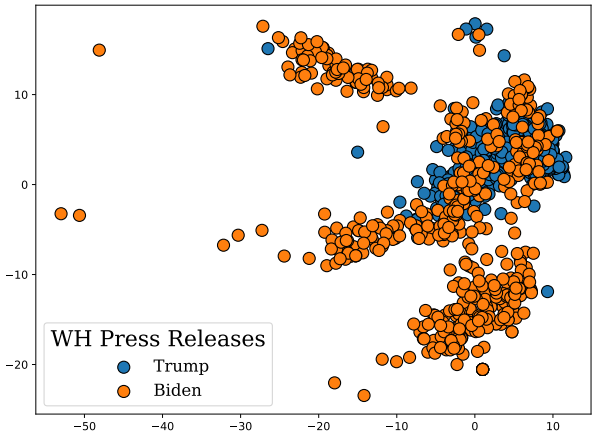


Figure 9: PCA of the mean-pooled internal representation of the non-fine-tuned `Pythia-70m` model of the two corpuses.

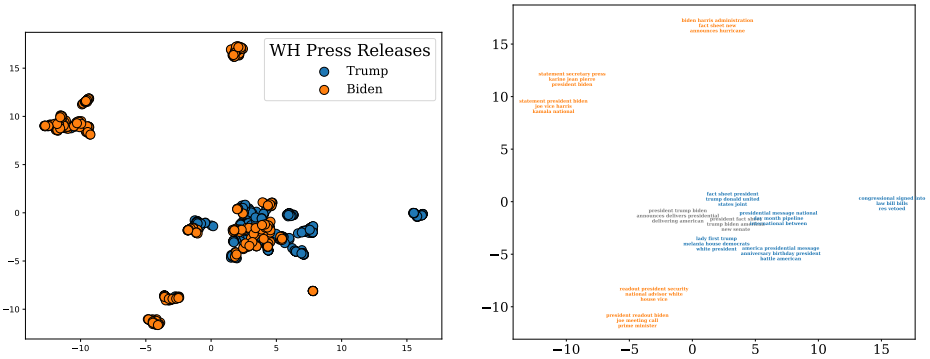


Figure 10: Umap of the titles of press releases from the Biden administration (in orange) and the Trump administration (in blue) on the **left**. The **right** plot shows the most frequent words for every cluster in the original umap. We color the words in grey if the percentage of titles in this cluster to belong to the Trump administration is between 30% and 70%. If it is bellow 30%, the words are colored in orange, and above 70% they are colored in blue.

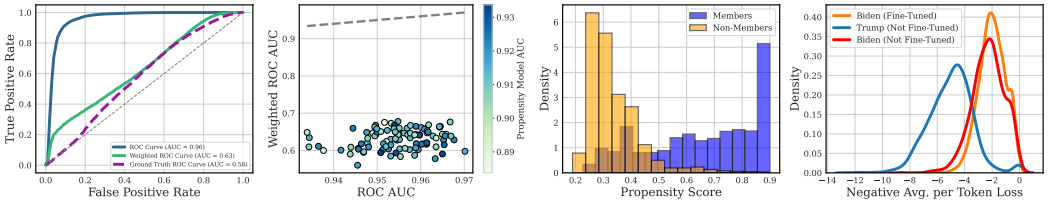


Figure 11: Results of the MIA experiment on the fine-tuned `Pythia-70m` model. The **first plot** shows the ROC curves of the vanilla, debiased and pseudo-ground-truth MIAs. The **second plot** compares the ROC AUC for the vanilla (x-axis) v.s. the debiased MIA (y-axis). Each dot represents one random split of the data into a training and a test set, the color representing the AUC of the propensity score classifier (who classifies based on the text). The **third plot** shows the propensity scores as infered by the propensity score model. Finally, the **last plot** shows a KDE plot of the negative per token mean loss of the fine-tuned model for all three datasets.

Parameter	Value
Number of epochs	2
Batch size	16
Weight decay	0
Optimizer	AdamW
Intial Learning Rate	5×10^{-4}
Learning Rate Schedule	Linear
`LoRA` r	16
`LoRA` α	32
`LoRA` dropout	0.05
Targeted modules	`q_proj, v_proj, query_key_value, k_proj`

Table 5: Parameters for the fine-tuning experiment. Unreported parameters are defaulted to the parameters of `transformers.TrainingArguments` and `peft.LoRAConfig`.

Fine-Tuning. We randomly split the titles from the Biden administration into two halfes of equal size. We then fine-tune the `Pythia-70m` model for 2 epoches. Parameters for the fine-tuning are given in Table 5.

Inference of the Propensity Score. We use the same propensity score model than for the first LLM experiment. Parameters are reported in Table 4.

Membership attacks. We conduct the MIAs in the same way than in the previous experiment.

Results. Complete results can be found in Figure 11.

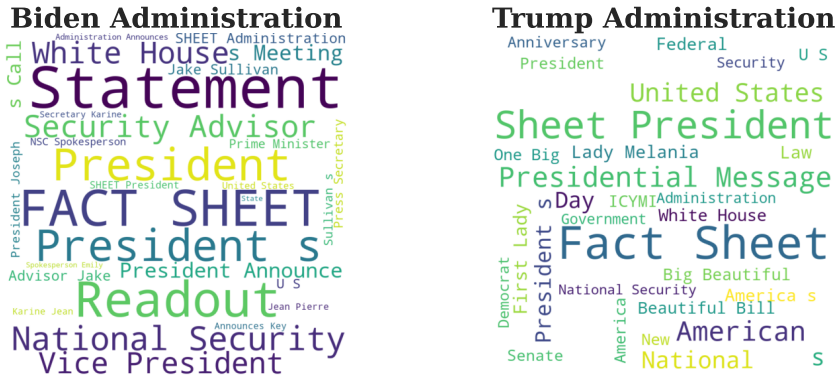


Figure 12: Word clouds for the titles of the press releases of the Biden administration (**left**) and the Trump I administration (**right**). To generate these pointclouds, the names of the Presidents and their Vice-Presidents have been masked by [NAME] and [FIRSTNAME] placeholders, which we have redacted from these word clouds for clarity.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main contributions are described in Sections 3.1 and 4 to 6 and backed up by the experiments in Section 7.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We acknowledged the main limitation of our work, which is its dependency on well-learned propensity scores.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Definitions 5.2 and 5.3 are standard or close to standard learning theory definitions/assumptions, Assumption 2.1 are standard causal inference assumptions.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Synthetic data generation process and parameters are discussed and experiments on real data is described too, in Section 7.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do not provide open access to code, but experiments are well detailed and easy to reproduce.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: As this is central to our method with cross-fitting, we describe this part well in Section 7.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Confidence intervals for performance metrics are reported when possible.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We did not record the total execution time. Although some experiments are more time consuming than others, compute resources are not excessive. For instance, fine-tuning `Pythia-70m` on the Trump-Biden data required less than 5 minutes of compute time on one single GPU.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: None of the ethics concerns apply.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We introduce a framework to causally interpret MIA performances, which we don't believe can have societal negative impacts, since the goal is to safe guard against misinterpretations.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: N/A

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See references in Section 7.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: Our new assets include scrapped press releases from the White House under two different administrations. These assets are not copyrighted, nor do they require consent or anonymization.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: N/A

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: N/A

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs only used for standard writing/editing tasks and code debugging.