# Low-Cost High-Power Membership Inference Attacks

Sajjad Zarifzadeh, Philippe Liu, **Reza Shokri**

Data Privacy and Trustworthy Machine Learning Lab
National University of Singapore

## Data Privacy in Machine Learning

Models should not **leak** training data

↓

Allow inferring what could not otherwise
be learned about a data record when
it is excluded from the training set

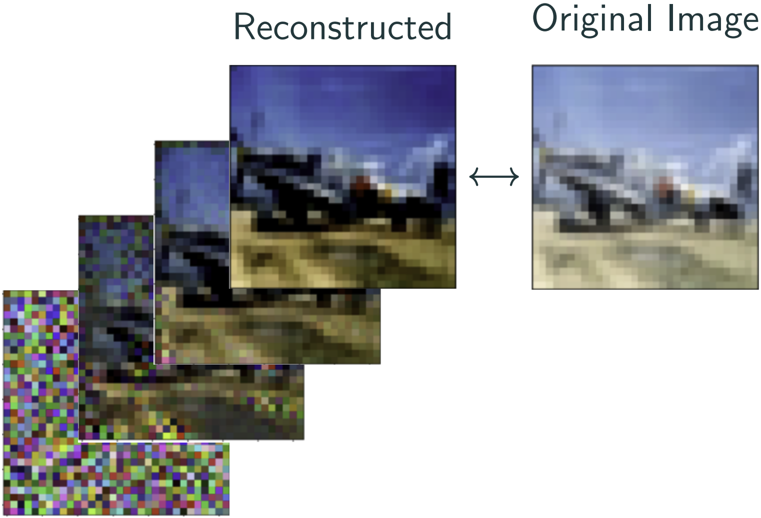
# Data Privacy in Machine Learning

Models should not **leak** training data

↓

Allow inferring what could not otherwise
be learned about a data record when
it is excluded from the training set

## Leakage![1]




[1] [Ye, Borovykh, Hayou, and Shokri] Leave-one-out Distinguishability in Machine Learning, ICLR 2024

## Measuring Information Leakage: Membership Inference Game

Sample data $x_0, x_1, x_2, \cdots, x_n \sim \pi$

Sample secret bit $\boxed{b \sim \{0, 1\}}$

Train a model $\boxed{x_b \quad x_2 \cdots x_n} \xrightarrow{\textit{training}} \theta$

## Measuring Information Leakage: Membership Inference Game

Sample data $\color{red}{x_0}, x_1, x_2, \cdots, x_n \sim \pi$

Sample secret bit $\boxed{b \sim \{0, 1\}}$

Train a model $\boxed{x_b \; x_2 \cdots x_n} \xrightarrow{\textit{training}} \theta$

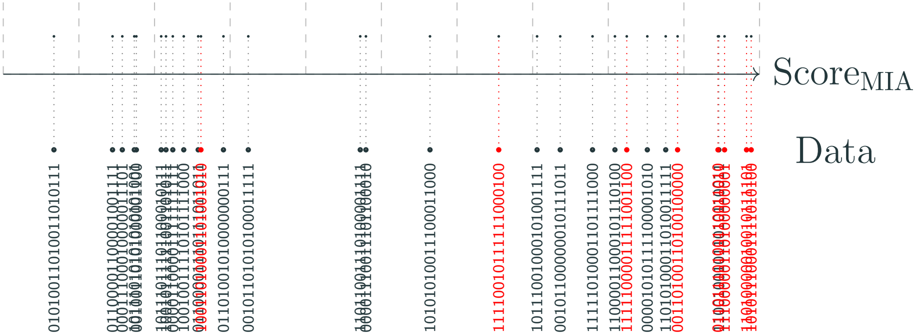
- Send $\theta$ and $x_0$ to adversary.
- Adversary **wins** if it correctly infers membership of $x_0$.
- Adversary's success is due to model's leakage.

## Membership Inference Attack (MIA)[2]

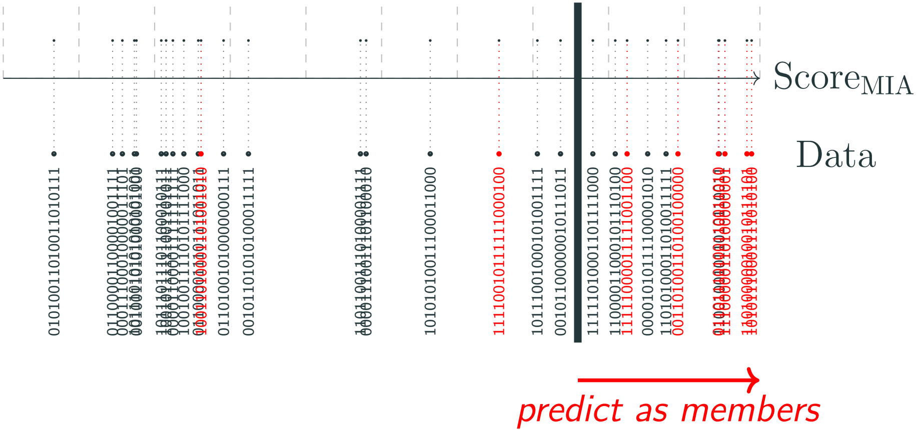
Given a model $\theta$ and a data point $x$, **infer if $x$ was part of the training set of $\theta$.**

[2][Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, IEEE S&P 2017

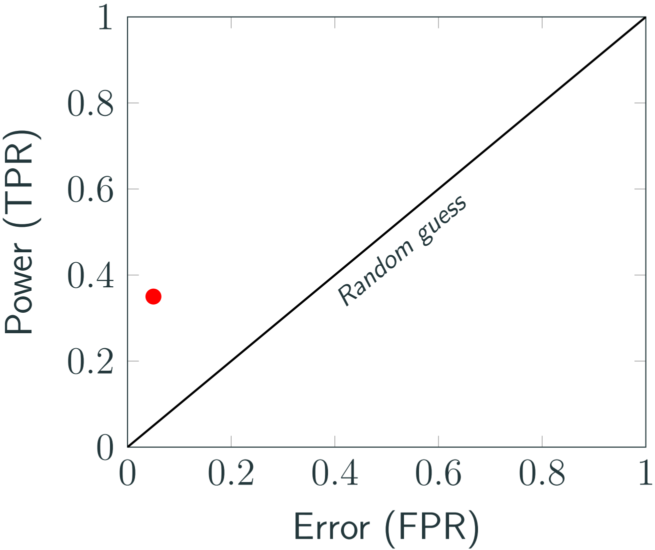
# How MIA helps partition the data universe



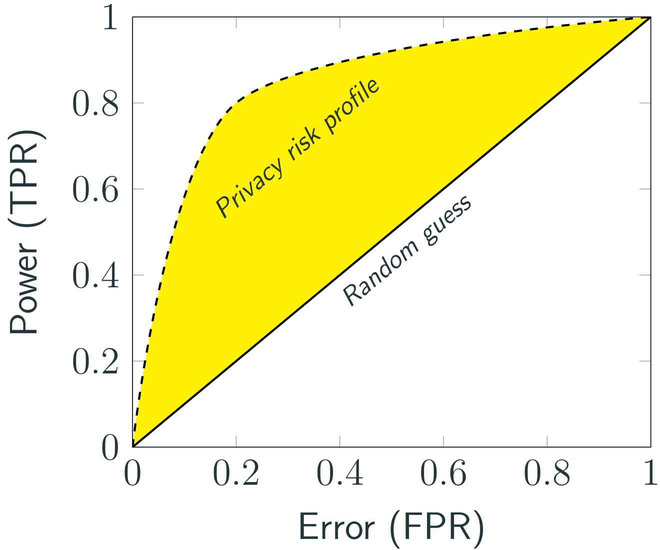
# How MIA helps partition the data universe



## TPR-FPR Tradeoff Curve (corresponding to a MIA game)



# TPR-FPR Tradeoff Curve (corresponding to a MIA game)



## Applications of MIA

- Privacy **auditing** tools (e.g., privacy-meter.com)
- Methods for quantitative analysis of **memorization**
- Oracles in **reconstruction** attacks

## Prior Work

- Over 8000 papers since [Homer et al., 2008]
- No single prior attack outperforms all others in every scenario
- Attacks outperform each other in different parts of the TPR-FPR tradeoff curve
- Some methods fail against well-generalized models
- Some methods fail against large models
- Many methods fail at detecting both in-distribution members and out-of-distribution non-members
- Many attacks are computationally very costly (as they require training so many reference models)

## Prior Work

- Over 8000 papers since [Homer et al., 2008]
- No single prior attack outperforms all others in every scenario
- Attacks outperform each other in different parts of the TPR-FPR tradeoff curve
- Some methods fail against well-generalized models
- Some methods fail against large models
- Many methods fail at detecting both in-distribution members and out-of-distribution non-members
- Many attacks are computationally very costly (as they require training so many reference models)

## Prior Work

- Over 8000 papers since [Homer et al., 2008]
- No single prior attack outperforms all others in every scenario
- Attacks outperform each other in different parts of the TPR-FPR tradeoff curve
- Some methods fail against well-generalized models
- Some methods fail against large models
- Many methods fail at detecting both in-distribution members and out-of-distribution non-members
- Many attacks are computationally very costly (as they require training so many reference models)

## Prior Work

- Over 8000 papers since [Homer et al., 2008]
- No single prior attack outperforms all others in every scenario
- Attacks outperform each other in different parts of the TPR-FPR tradeoff curve
- Some methods fail against well-generalized models
- Some methods fail against large models
- Many methods fail at detecting both in-distribution members and out-of-distribution non-members
- Many attacks are computationally very costly (as they require training so many reference models)

## Prior Work

- Over 8000 papers since [Homer et al., 2008]
- No single prior attack outperforms all others in every scenario
- Attacks outperform each other in different parts of the TPR-FPR tradeoff curve
- Some methods fail against well-generalized models
- Some methods fail against large models
- Many methods fail at detecting both in-distribution members and out-of-distribution non-members
- Many attacks are computationally very costly (as they require training so many reference models)

# Prior Work

- Over 8000 papers since [Homer et al., 2008]
- No single prior attack outperforms all others in every scenario
- Attacks outperform each other in different parts of the TPR-FPR tradeoff curve
- Some methods fail against well-generalized models
- Some methods fail against large models
- Many methods fail at detecting both in-distribution members and out-of-distribution non-members
- Many attacks are computationally very costly (as they require training so many reference models)

## Prior Work

- Over 8000 papers since [Homer et al., 2008]
- No single prior attack outperforms all others in every scenario
- Attacks outperform each other in different parts of the TPR-FPR tradeoff curve
- Some methods fail against well-generalized models
- Some methods fail against large models
- Many methods fail at detecting both in-distribution members and out-of-distribution non-members
- Many attacks are computationally very costly (as they require training so many reference models)

## Expectations from a MIA method

MIA must be **efficient** (to make the privacy auditing practical), **precise** (to accurately reflect the risk), and **robust** (to be a reliable auditing method under various settings).

We design a Robust Membership Inference Attack (RMIA) with these objectives

## Expectations from a MIA method

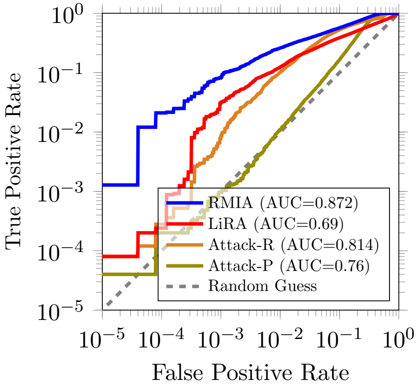
MIA must be **efficient** (to make the privacy auditing practical), **precise** (to accurately reflect the risk), and **robust** (to be a reliable auditing method under various settings).

We design a Robust Membership Inference Attack (RMIA) with these objectives

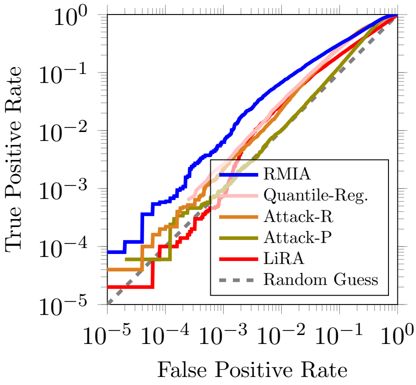
| # Ref Models | Attack | CIFAR-10 | | |
|---|---|---|---|---|
| | | AUC | TPR @ FPR | |
| | | | 0.01 % | 0.0 % |
| 0 | Attack-P [Ye et al., 2022, Yeom et al., 2018] | $58.19 \pm 0.33$ | $0.01 \pm 0.01$ | $0.00 \pm 0.01$ |
| 1* | Quantile-Reg. [Bertran et al., 2023] | $61.45 \pm 0.29$ | $0.08 \pm 0.05$ | $0.03 \pm 0.03$ |
| 1 | Attack-R [Ye et al., 2022] | $63.65 \pm 0.27$ | $0.07 \pm 0.04$ | $0.02 \pm 0.02$ |
| | LiRA [Carlini et al., 2022] | $53.20 \pm 0.23$ | $0.48 \pm 0.10$ | $0.25 \pm 0.11$ |
| | **RMIA [Zarifzadeh et al., 2024]** | $\mathbf{68.64 \pm 0.43}$ | $\mathbf{1.19 \pm 0.27}$ | $\mathbf{0.51 \pm 0.32}$ |
| 2 | Attack-R [Ye et al., 2022] | $63.35 \pm 0.30$ | $0.32 \pm 0.15$ | $0.08 \pm 0.06$ |
| | LiRA [Carlini et al., 2022] | $54.42 \pm 0.34$ | $0.67 \pm 0.24$ | $0.27 \pm 0.12$ |
| | LiRA [Carlini et al., 2022] (Online) | $63.97 \pm 0.35$ | $0.76 \pm 0.24$ | $0.43 \pm 0.21$ |
| | **RMIA [Zarifzadeh et al., 2024]** | $\mathbf{70.13 \pm 0.37}$ | $\mathbf{1.71 \pm 0.23}$ | $\mathbf{0.91 \pm 0.30}$ |
| 4 | Attack-R [Ye et al., 2022] | $63.52 \pm 0.29$ | $0.65 \pm 0.21$ | $0.21 \pm 0.20$ |
| | LiRA [Carlini et al., 2022] | $54.60 \pm 0.25$ | $0.97 \pm 0.44$ | $0.57 \pm 0.40$ |
| | LiRA [Carlini et al., 2022] (Online) | $67.00 \pm 0.33$ | $1.38 \pm 0.37$ | $0.51 \pm 0.35$ |
| | **RMIA [Zarifzadeh et al., 2024]** | $\mathbf{71.02 \pm 0.37}$ | $\mathbf{2.91 \pm 0.64}$ | $\mathbf{2.13 \pm 0.47}$ |
| 127 | Attack-R [Ye et al., 2022] | $64.41 \pm 0.41$ | $1.52 \pm 0.33$ | $0.80 \pm 0.43$ |
| | LiRA [Carlini et al., 2022] | $55.18 \pm 0.37$ | $1.37 \pm 0.32$ | $0.72 \pm 0.31$ |
| | **RMIA [Zarifzadeh et al., 2024]** | $\mathbf{71.71 \pm 0.43}$ | $\mathbf{4.18 \pm 0.61}$ | $\mathbf{3.14 \pm 0.87}$ |
| 254 | LiRA [Carlini et al., 2022] (Online) | $72.04 \pm 0.47$ | $3.39 \pm 0.86$ | $2.01 \pm 0.78$ |
| | **RMIA [Zarifzadeh et al., 2024] (Online)** | $\mathbf{72.25 \pm 0.46}$ | $\mathbf{4.31 \pm 0.47}$ | $\mathbf{3.15 \pm 0.61}$ |

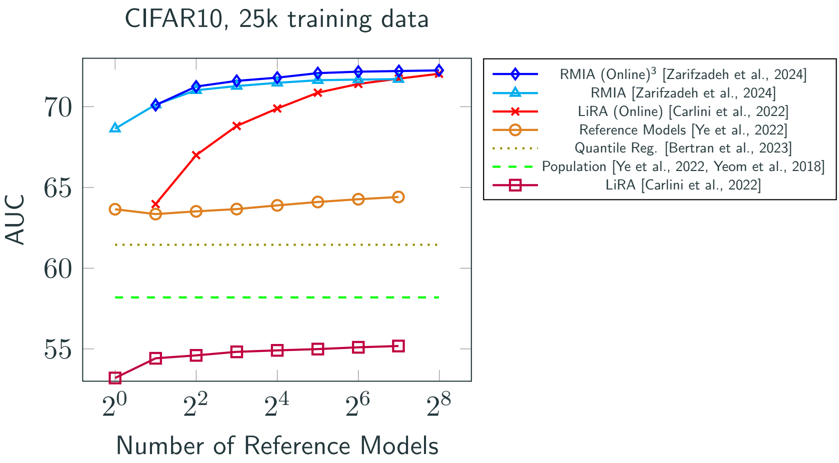
# Attacking larger models with 1 reference/attack model

CIFAR100




ImageNet

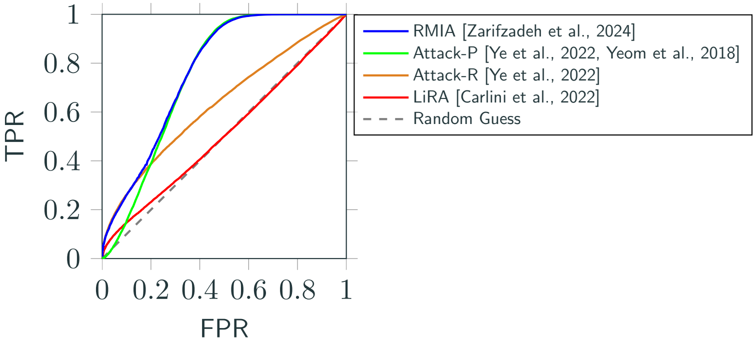





[3] In the online setting, for every membership inference $\mathrm{MIA}(x;\theta)$, the adversary trains half of his reference models on datasets that contain $x$. We consider these impractical yet powerful methods as *proof of concept* attacks.

# In-distribution members and out-of-distribution non-members

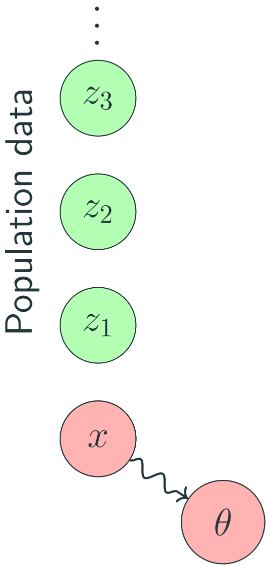
In a reconstruction attack, an adversary can use MIA as an oracle on extremely large number of samples which are not necessarily generated from the same distribution as the training data. MIA should filter out the OOD non-members while detecting in distribution members.



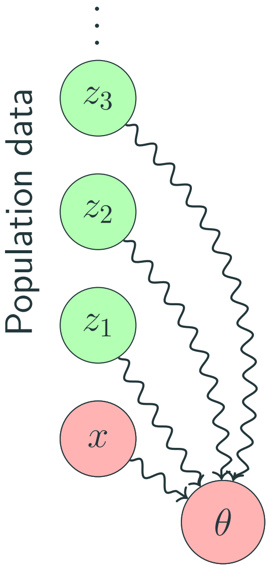

Results are for CIFAR-10 models and non-members from CINIC-10.

# How does RMIA work?

# One Hypothesis: $x$ was in the Training Set that Resulted in $\theta$

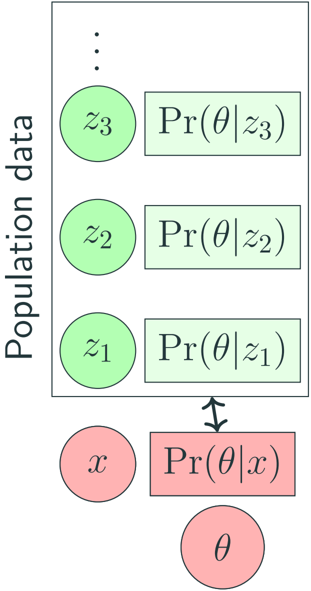


# A Fine-Grained Model of the Null Hypothesis



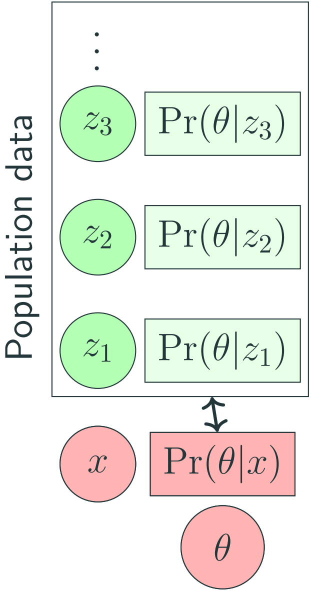

Null hypothesis: composition of worlds where a random population data point $z$ (and not $x$) was in the training set that resulted in $\theta$.

## A Fine-Grained Model of the Null Hypothesis




Null hypothesis: composition of worlds where a random population data point $z$ (and not $x$) was in the training set that resulted in $\theta$.

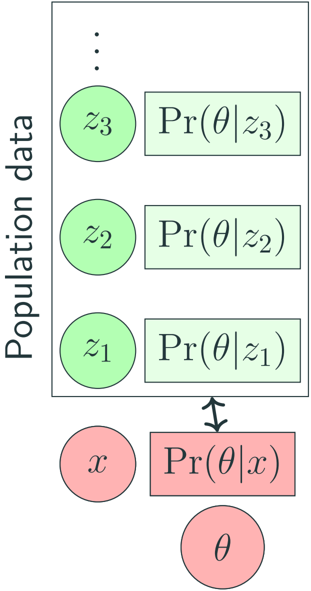
## A Fine-Grained Model of the Null Hypothesis




Null hypothesis: composition of worlds where a random population data point $z$ (and not $x$) was in the training set that resulted in $\theta$.

Design **pairwise likelihood ratio tests** to check the membership of a data point $x$ relative to $z$.

## A Fine-Grained Model of the Null Hypothesis




Null hypothesis: composition of worlds where a random population data point $z$ (and not $x$) was in the training set that resulted in $\theta$.

Design **pairwise likelihood ratio tests** to check the membership of a data point $x$ relative to $z$.

$$\mathrm{LR}_\theta(x, z) = \frac{\Pr(\theta|x)}{\Pr(\theta|z)} > 1 \ ?$$

# Composing the Pairwise LR Tests

We **compose** the pairwise tests:

$$\boxed{\mathrm{Score}_{\mathrm{MIA}}(x;\theta) = \Pr_{z\sim\pi}\Big(\mathrm{LR}_\theta(x,z) \geq 1\Big)}$$

MIA corresponding to a given FPR returns "member" if:

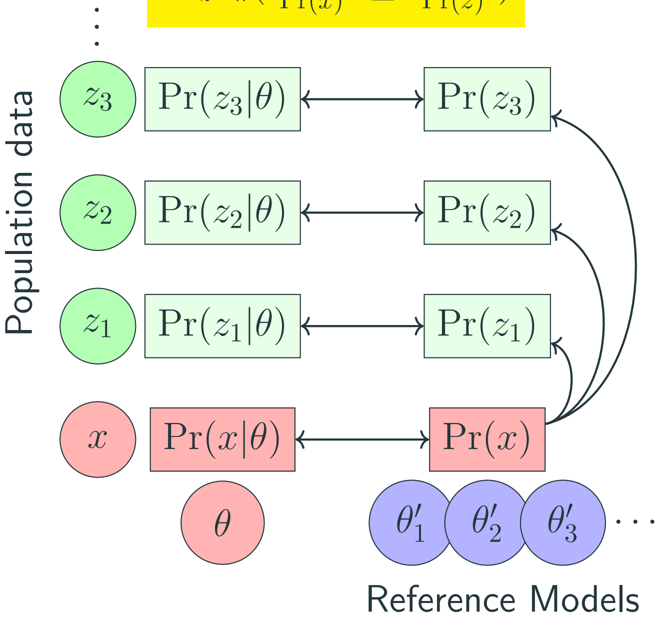
$$\mathrm{Score}_{\mathrm{MIA}}(x;\theta) \geq (1 - \mathrm{FPR})$$

## Computing the Pairwise Likelihood Ratios

$$\mathrm{LR}_\theta(x, z) = \frac{\Pr(\theta|x)}{\Pr(\theta|z)} = \left(\frac{\Pr(x|\theta)}{\Pr(x)}\right) \cdot \left(\frac{\Pr(z|\theta)}{\Pr(z)}\right)^{-1}$$

$\Pr(x)$ is the mean of $\Pr(x|\theta')$ over **reference models** $\theta'$.

$$\Pr_{z\sim\pi}\left(\frac{\Pr(x|\theta)}{\Pr(x)} \geq \frac{\Pr(z|\theta)}{\Pr(z)}\right)$$



## Summary of Results

- RMIA outperforms all prior attacks in every configuration, for every benchmark dataset and models used in MIA literature.
- TPR-FPR curves obtained for RMIA dominate the curves obtained from other methods for all FPR
- RMIA is low-cost, and can achieve close to its maximum power while using only a few reference models
- Why? Other methods appear to be uncalibrated and average versions of RMIA.

| **Method** | RMIA | LiRA | Attack-R | Attack-P |
|---|---|---|---|---|
| **MIA Score** | $\Pr_z\left(\frac{\Pr(\theta\|x)}{\Pr(\theta\|z)} \geq 1\right)$ | $\frac{\Pr(\theta\|x)}{\Pr(\theta\|\bar{x})}$ | $\Pr_{\theta'}\left(\frac{\Pr(x\|\theta)}{\Pr(x\|\theta')} \geq 1\right)$ | $\Pr_z\left(\frac{\Pr(x\|\theta)}{\Pr(z\|\theta)} \geq 1\right)$ |

## Summary of Results

- RMIA outperforms all prior attacks in every configuration, for every benchmark dataset and models used in MIA literature.
- TPR-FPR curves obtained for RMIA dominate the curves obtained from other methods for all FPR
- RMIA is low-cost, and can achieve close to its maximum power while using only a few reference models
- Why? Other methods appear to be uncalibrated and average versions of RMIA.

| **Method** | RMIA | LiRA | Attack-R | Attack-P |
|---|---|---|---|---|
| **MIA Score** | $\Pr_z\left(\frac{\Pr(\theta\|x)}{\Pr(\theta\|z)} \geq 1\right)$ | $\frac{\Pr(\theta\|x)}{\Pr(\theta\|\bar{x})}$ | $\Pr_{\theta'}\left(\frac{\Pr(x\|\theta)}{\Pr(x\|\theta')} \geq 1\right)$ | $\Pr_z\left(\frac{\Pr(x\|\theta)}{\Pr(z\|\theta)} \geq 1\right)$ |

# References i


Bertran, M., Tang, S., Kearns, M., Morgenstern, J., Roth, A., and Wu, Z. S. (2023).
**Scalable membership inference attacks via quantile regression.**
In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS'23).

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. (2022).
**Membership inference attacks from first principles.**
In IEEE Symposium on Security and Privacy (S&P'22), page 1897–1914.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. (2008).
**Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays.**
PLoS Genetics, 4(8).

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., and Shokri, R. (2022).
**Enhanced membership inference attacks against machine learning models.**

In Proceedings of the 29th ACM SIGSAC Conference on Computer and Communications Security (CCS'22), page 3093–3106.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. (2018).
**Privacy risk in machine learning: Analyzing the connection to overfitting.**
In 2018 IEEE 31st Computer Security Foundations Symposium (CSF'18), pages 268–282. IEEE.

Zarifzadeh, S., Liu, P., and Shokri, R. (2024).
**Low-cost high-power membership inference attacks.**
In International conference on machine learning (ICML'24).