Rectifying Unlearning Efficacy and Privacy Evaluation: A New Inference Attack Perspective

Nima Naderloui[1], Shenao Yan[1], Binghui Wang[2], Jie Fu[3], Wendy Hui Wang[3], Weiran Liu[4], Yuan Hong[1]

[1]University of Connecticut, [2]Illinois Institute of Technology, [3]Stevens Institute of Technology, [4]Alibaba Group

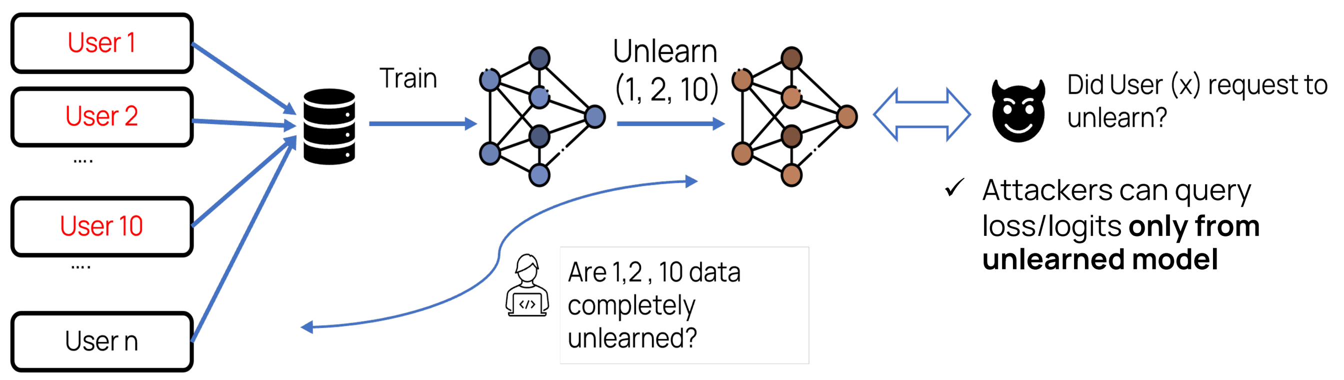
https://github.com/datasec-lab/Ruli

Our paper

Introduction

- ❑ Inexact Unlearning for efficient data removal, privacy protection and safety.
- ❑ Inexact unlearning requires empirical evaluation
- ❑ Unlearning should **protect all samples** and **be close** to **(Retraining)** gold removal standard [1]

Threat model



✓ Efficacy is a MIA test only trusted ends (model providers and users) can perform not real attackers! (MIA resilience ≠unlearning)

What Was Missing

PI. Average-case MIAs Cannot Fully Disclose Unlearning Privacy

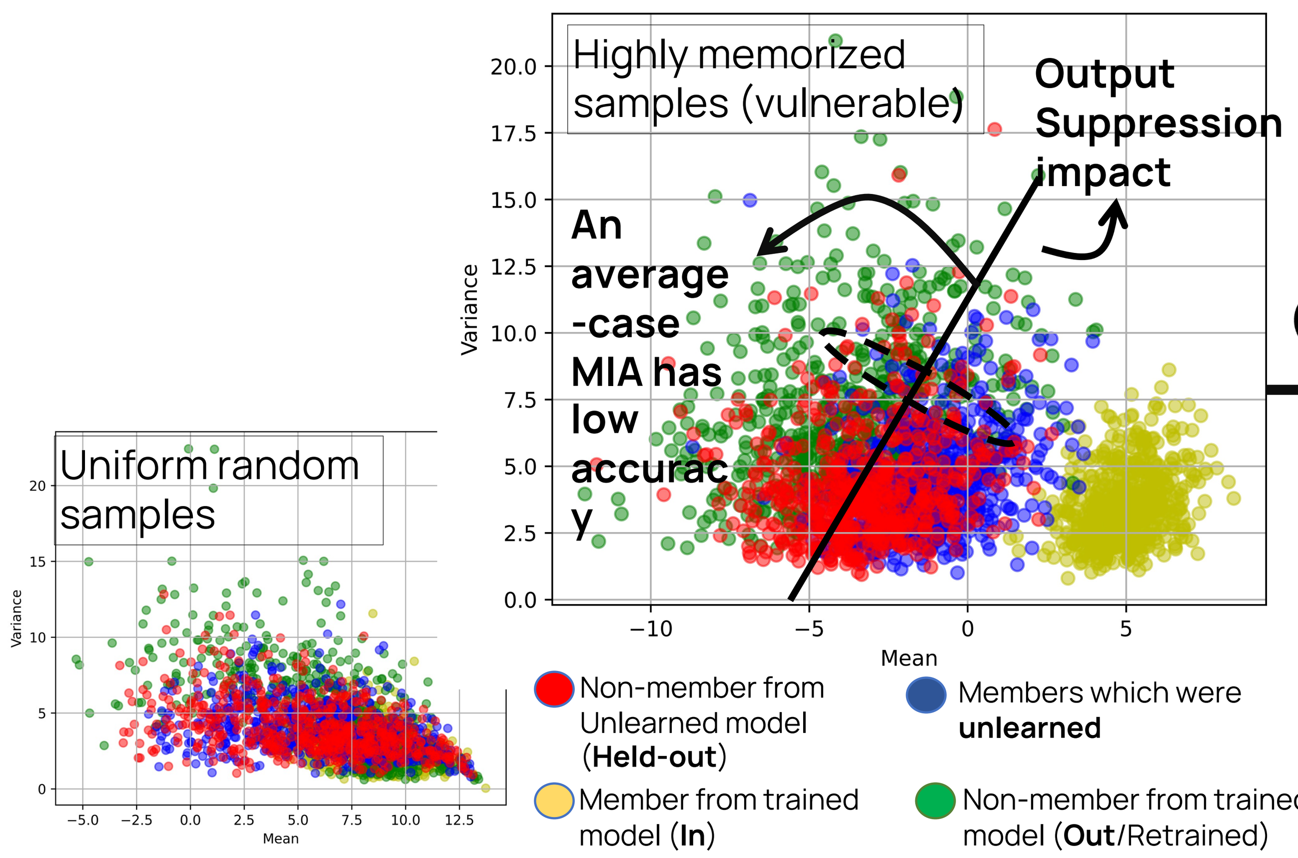
- ❑ An MIA on population would hide the per sample unlearning requirement
- ❑ **We use per-sample MIAs like [1]**

PII. Evaluating Random Samples Underestimates Unlearning Privacy

- ❑ Many samples are well-protected even with no unlearning
- ❑ **We are not interested in well-protected samples**

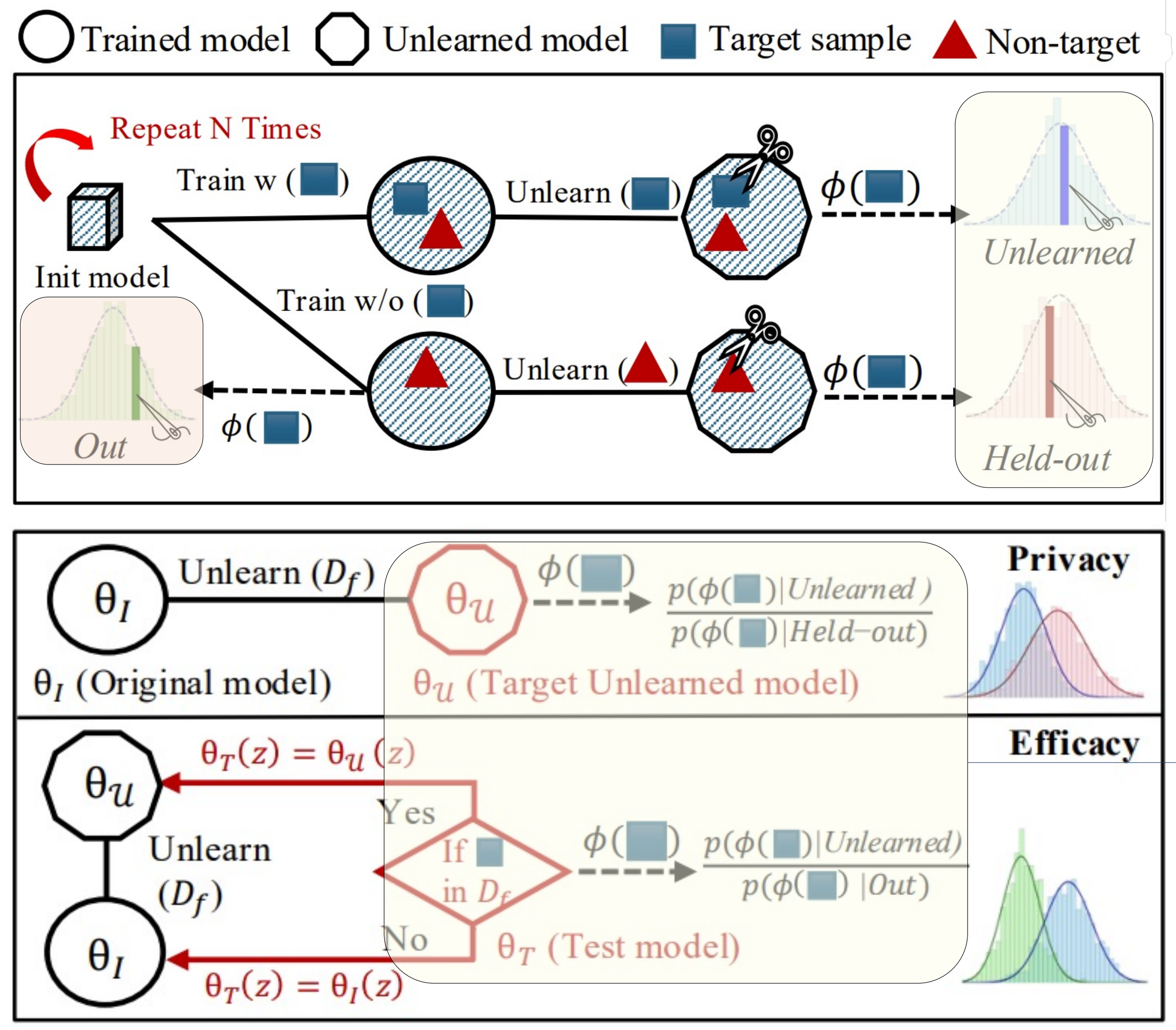
PIII. Incomplete Comparisons with the Retrain Baseline (Efficacy)

- ❑ We need to distinguish whether a sample is *unlearned or Retrained.*
- ❑ Challenge is unlearning suppresses outputs and this is not necessarily unlearning [2] -> MIA resilience ≠unlearning
- ❑ **We need a MIA to calibrate output suppression**



RULI: Workflow and Algorithm

- ✓ A unified per-sample MIA to measure privacy leakage with efficacy with no additional shadow costs
- ✓ With N training and unlearning, we will get N/3 instance per distribution (while keeping unlearning rate low).
- ✓ Valid in Game theoretical backbone



a) Select target data

b) Train shadow models; prepare distributions

c) Query:
Unlearned model -> Privacy
2.Hypothetic Test model -> Efficacy

Game 2: Targeted MIA for unlearning privacy

1. The *challenger* trains a model with $D_{\text{train}} \subseteq \mathcal{D}$ and gets θ_I.
2. The *adversary* chooses a target set D_{target} and sends to *challenger*.
3. The *challenger* unlearns $D_f \cup \{D_{\text{train}} \cap D_{\text{target}}\}$ to get the model $\theta_{\mathcal{U}}$.
4. The *challenger* flips a coin c:
   - If $c =$ head, the challenger chooses a data point z from $D_f \cap D_{\text{target}}$
   - If $c =$ tail, the challenger chooses a data point z from $D_{\text{target}} \setminus D_{\text{train}}$
5. The *challenger* sends the selected data point z to the adversary.
6. Given the unlearned model $\theta_{\mathcal{U}}$, the *adversary* queries z to determine if it is in D_{train} and guess $\hat{c} = \{\text{head, tail}\}$; *adversary* wins if $\hat{c} = c$.

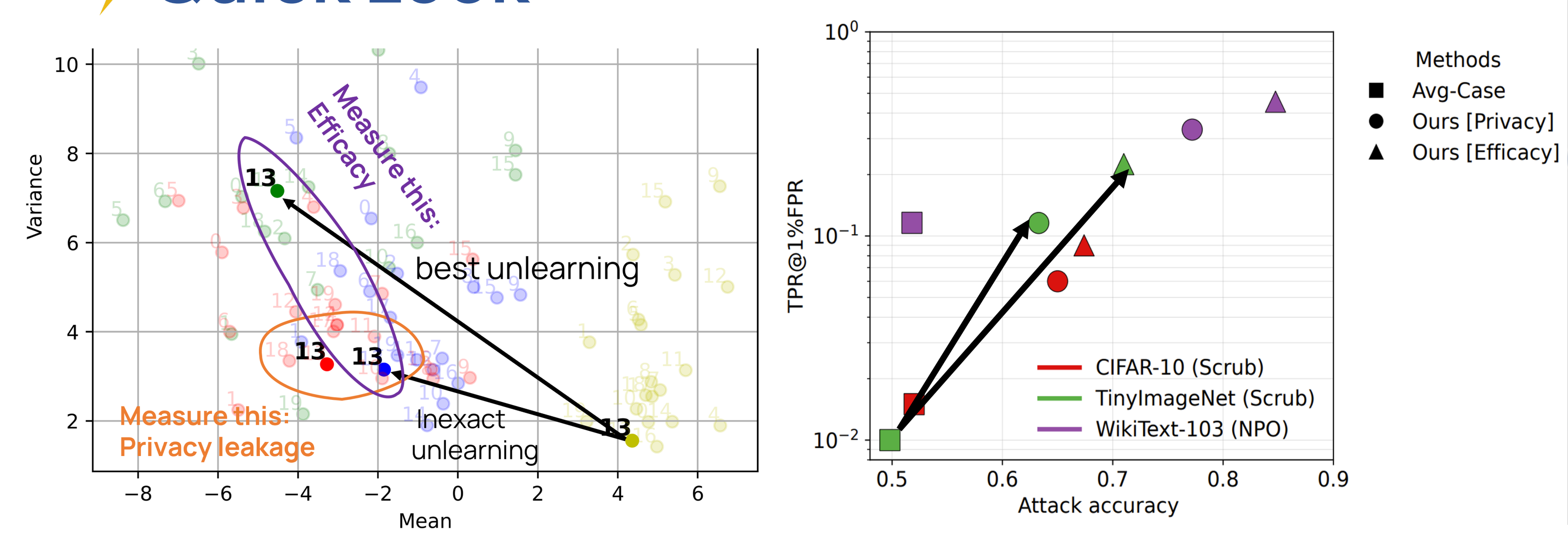
Revisited Game for **PII**

Game 3: MIA for unlearning efficacy

1. The *challenger* trains a model with $D_{\text{train}} \subseteq \mathcal{D}$ and gets θ_I.
2. The *adversary* chooses a target set D_{target} and sends to *challenger*.
3. The challenger unlearns $D_f \cup \{D_{\text{train}} \cap D_{\text{target}}\}$ to get the model $\theta_{\mathcal{U}}$.
4. The challenger flips a coin c:
   - If $c =$ head, the challenger chooses a data point z from $D_f \cap D_{\text{target}}$, and the query result will be given as $f_{\theta_{\mathcal{U}}}(\cdot)$
   - If $c =$ tail, the challenger chooses a data point z from $D_{\text{target}} \setminus D_{\text{train}}$, and the query result will be given as $f_{\theta_I}(\cdot)$
5. The *challenger* sends the selected data point z to the adversary.
6. Given the query from queries z as $f_\theta(\cdot)$, the *adversary* determines if z is in D_f and guess $\hat{c} = \{\text{head, tail}\}$; *adversary* wins if $\hat{c} = c$.

Revisited Game for **PIII**

Quick Look

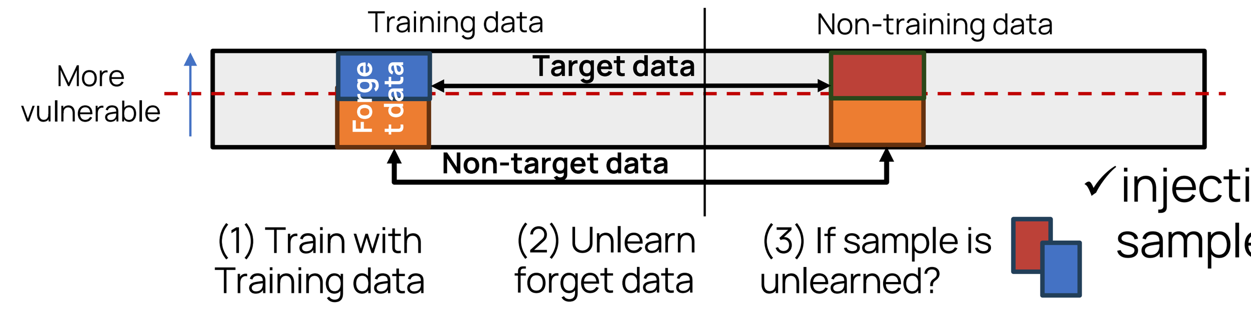


RULI achieves SOTA on privacy leakage and is the first to separate efficacy and privacy for Machine Unlearning

Experiments

Baselines & Settings

- ❑ Targeting SOTA inexact unlearning's
- ❑ Choosing best unlearning parameters for any experiment
- ❑ Targeted population MIA baseline to show impact of **PI**



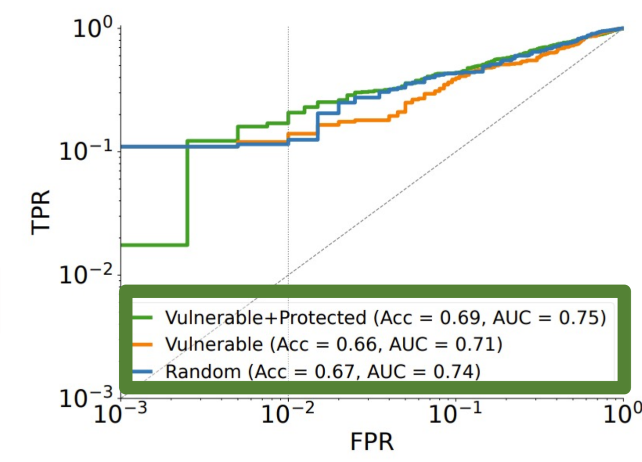
✓injecting vulnerable samples as canaries

Different targets to show **PII:** Uniform random samples (most of existing works), Protected samples, Vulnerable samples only, Vulnerable + protected (Best) , Random from one class [1]

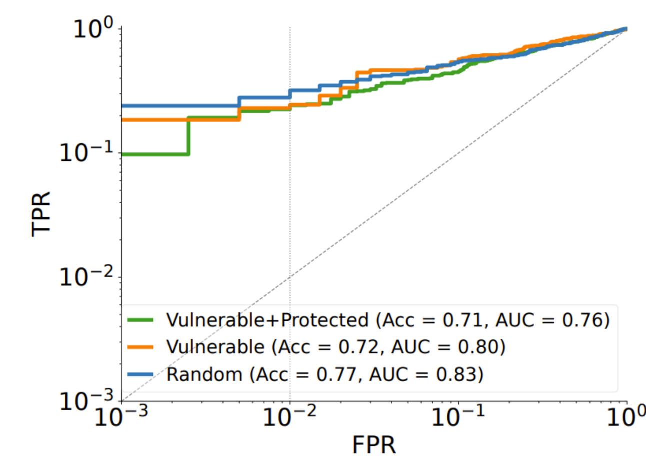
a) Image classification. Unlearn images from trained/finetuned model. CIFAR-10&100/Tiny ImageNet

Target data	Targeted average-case attack (Population attack)				RULI			
	AUC	ACC	TPR@ 1%FPR	TPR@ 5%FPR	AUC	ACC	TPR@ 1% FPR	TPR@ 5%FPR
ℓ_1 Sparse								
Vulnerable only	54.4%	55.1%	2.3%	5.2%	59.6%	56.0%	2.4%	12.4%
Vulnerable as canaries	55.3%	54.7%	0.8%	5.6%	**62.6%**	**57.0%**	**6.3%**	**16.6%**
Random	53.2%	52.8%	0.0%	2.4%	56%	54.4%	0.8%	6.4%
Scrub								
Vulnerable only	52.5%	52.4%	2.0%	5.4%	65.3%	61.5%	11.7%	23.9%
Vulnerable as canaries	56.0%	56.2%	1.0%	6.3%	**69.5%**	**63.6%**	**10.9%**	**27.1%**
Random	49.6%	49.8%	1.0%	2.8%	59.7%	57.0%	6.0%	14.0%

unlearning <1% of data from fine-tuned Swin-small model with Tiny ImageNet (left; privacy leakage, right; efficacy)



(a) ℓ_1 Sparse



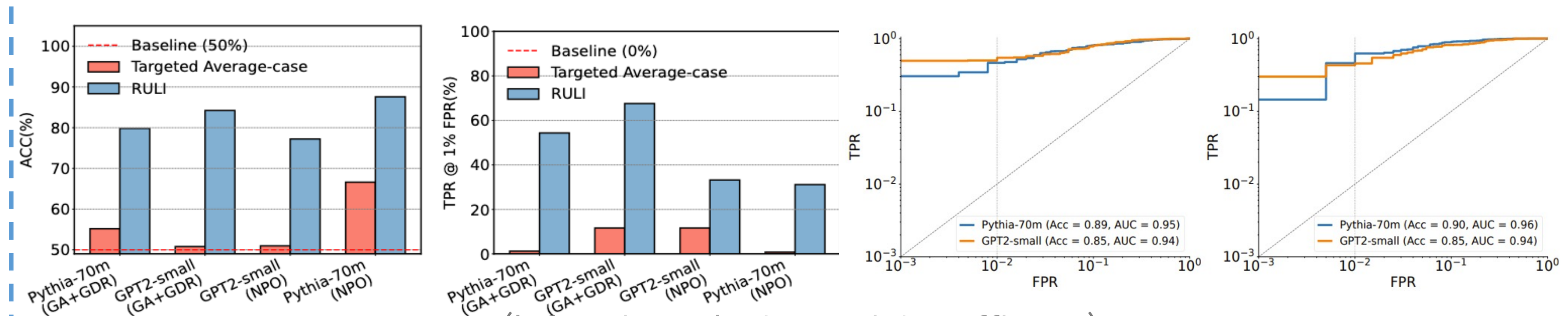
(b) Scrub

- **Per-sample attacks work better (PI)**
- **Our canary injection settings shows higher leakage (PII)**
- **Better MIA resilience does not guarantee Efficacy (PIII)**

•5 % canaries: RULI still finds leaks—TPR@1 % = 8.7 %, Acc = 68.5 % (CIFAR-10).
•*Mitigation: Sequentially unlearn samples with similar memorization.*

b) Language models. Unlearn last 7-gram sequence from WikiText-103

Example: ...The Meridian Historic Districts and Landmarks Commission was created in 1979, and the Meridian Main Street program was founded in 1985.



(left: privacy leakage, right: efficacy)

• Limitation: not feasible to apply RULI to foundation models or model with large knowledge domain

References:
[1] Hayes, Jamie, et al. "Inexact unlearning needs more careful evaluations to avoid a false sense of privacy." In SaTML 2025.
[2] Cooper, A. Feder, et al. "Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice." In GenLaw 2024