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

USENIX Security 2025
Track 3: ML and AI Privacy 2

It's 2025: Has Unlearning Already Won?

- ❑ A large and growing body of work has been introduced for **inexact selective unlearning.**
- ❑ Empirical evaluations indicate the subtle and incremental improvements in recent unlearning works.
- ❑ Did we solve unlearning?! or need to revisit empirical evaluation?!

a) Unlearning request

Image generated by the author (Nima Naderloui) with DALL·E 4o (OpenAI)
Prompt: "Single-panel office cartoon, simple pastel colors, ...<context>... style is similar to classic comics in newspapers. The font is Comic Sans"

It's 2025: Has Unlearning Already Won?

Failure of membership inference attack (MIA) → Better Forgetting [1]

Existing MIAs suggest that unlearning approximates Retraining (Gold standard)

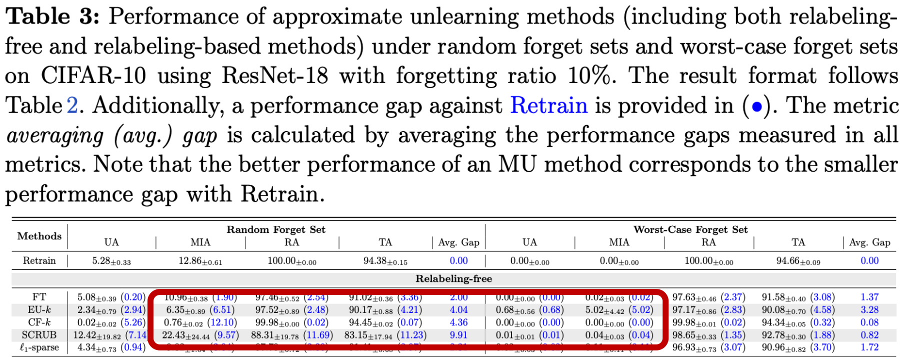


Table 3: Performance of approximate unlearning methods (including both relabeling-free and relabeling-based methods) under random forget sets and worst-case forget sets on CIFAR-10 using ResNet-18 with forgetting ratio 10%. The result format follows Table 2. Additionally, a performance gap against Retrain is provided in (•). The metric *averaging (avg.) gap* is calculated by averaging the performance gaps measured in all metrics. Note that the better performance of an MU method corresponds to the smaller performance gap with Retrain.

Methods	Random Forget Set					Worst-Case Forget Set				
	UA	MIA	RA	TA	Avg. Gap	UA	MIA	RA	TA	Avg. Gap
Retrain	5.28$_{\pm0.33}$	12.86$_{\pm0.61}$	100.00$_{\pm0.00}$	94.38$_{\pm0.15}$	0.00	0.00$_{\pm0.00}$	0.00$_{\pm0.00}$	100.00$_{\pm0.00}$	94.66$_{\pm0.09}$	0.00
Relabeling-free										
FT	5.08$_{\pm0.39}$ (0.20)	10.96$_{\pm0.38}$ (1.90)	97.46$_{\pm0.52}$ (2.54)	91.02$_{\pm0.36}$ (3.36)	2.00	0.00$_{\pm0.00}$ (0.00)	0.02$_{\pm0.03}$ (0.02)	97.63$_{\pm0.46}$ (2.37)	91.58$_{\pm0.40}$ (3.08)	1.37
EU-k	2.34$_{\pm0.79}$ (2.94)	6.35$_{\pm0.89}$ (6.51)	97.52$_{\pm0.89}$ (2.48)	90.17$_{\pm0.88}$ (4.21)	4.04	0.68$_{\pm0.56}$ (0.68)	5.02$_{\pm4.42}$ (5.02)	97.17$_{\pm0.86}$ (2.83)	90.08$_{\pm0.70}$ (4.58)	3.28
CF-k	0.02$_{\pm0.02}$ (5.26)	0.76$_{\pm0.02}$ (12.10)	99.98$_{\pm0.00}$ (0.02)	94.45$_{\pm0.02}$ (0.07)	4.36	0.00$_{\pm0.00}$ (0.00)	0.00$_{\pm0.00}$ (0.00)	99.98$_{\pm0.01}$ (0.02)	94.34$_{\pm0.05}$ (0.32)	0.08
SCRUB	12.42$_{\pm19.82}$ (7.14	22.43$_{\pm24.44}$ (9.57)	88.31$_{\pm19.78}$ (11.69)	83.15$_{\pm17.94}$ (11.23)	9.91	0.01$_{\pm0.01}$ (0.01)	0.04$_{\pm0.03}$ (0.04)	98.65$_{\pm0.33}$ (1.35)	92.78$_{\pm0.30}$ (1.88)	0.82
ℓ_1-sparse	4.34$_{\pm0.73}$ (0.94)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	96.93$_{\pm0.73}$ (3.07)	90.96$_{\pm0.82}$ (3.70)	1.72

One-way MIA acc: low MIA accuracy gap < 3% with "retraining" on top unlearning [2].

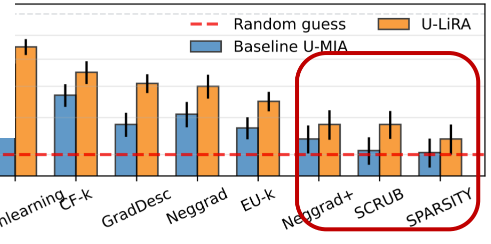
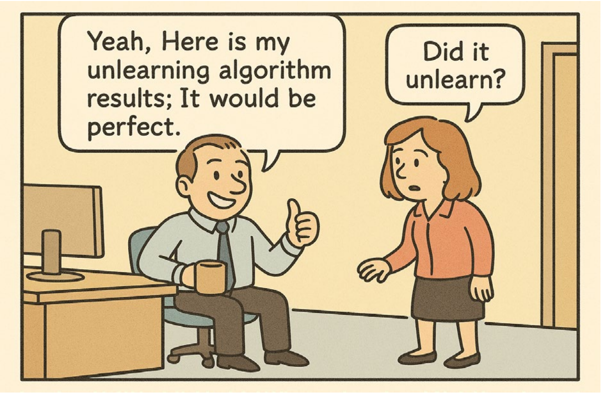


Figure 1 | *Membership inference attack accuracy using a baseline attack and U-LiRA across different unlearning algorithms. Attack and unlearning algorithm descriptions are in Section 4. U-LiRA outperforms the baseline by a large margin across all unlearning algorithms because it creates per-example MIA decision rules.*

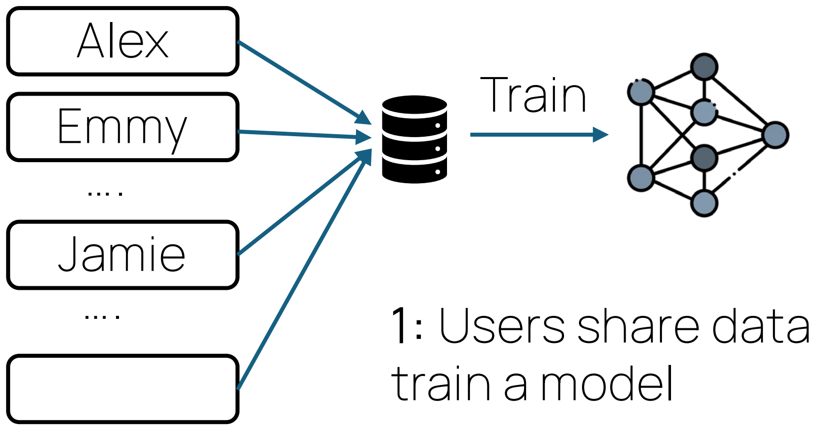
SOTA on privacy leakage: MIA accuracy gap < 10% on top unlearning [3].


Yeah, Here is my unlearning algorithm results; It would be perfect.
Did it unlearn?

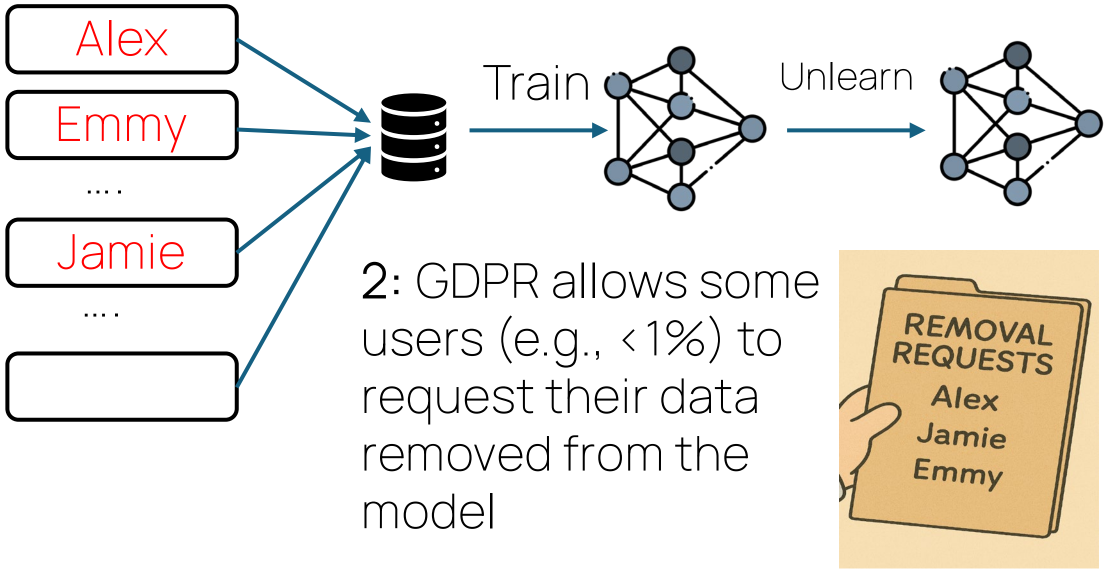
b) Using a fast inexact unlearning

[1] Jagielski, Matthew, et al. "Measuring forgetting of memorized training examples." In ICLR 2023.
[2] Fan, Chongyu, et al. "Challenging forgets: Unveiling the worst-case forget sets in machine unlearning." In ECCV 2024.
[3] Hayes, Jamie, et al. "Inexact unlearning needs more careful evaluations to avoid a false sense of privacy." In SaTML 2025.

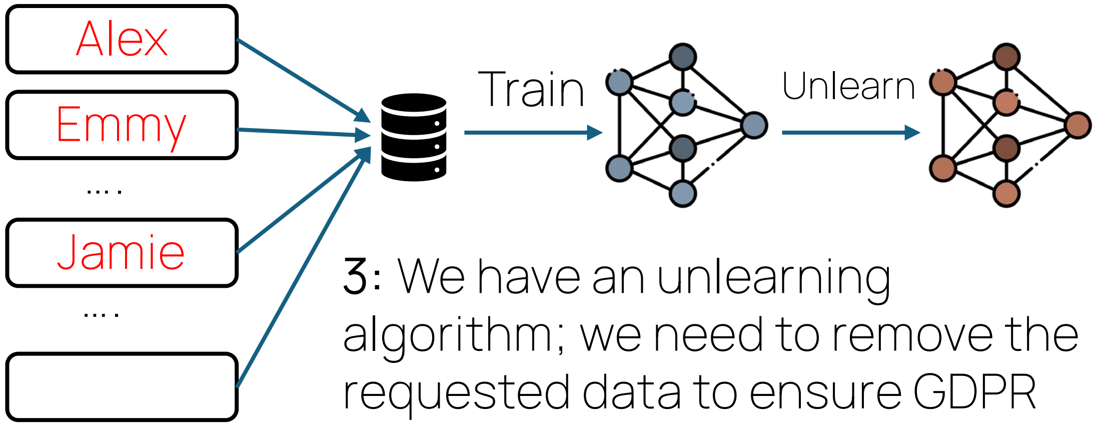
Warm-up: Our Motivation



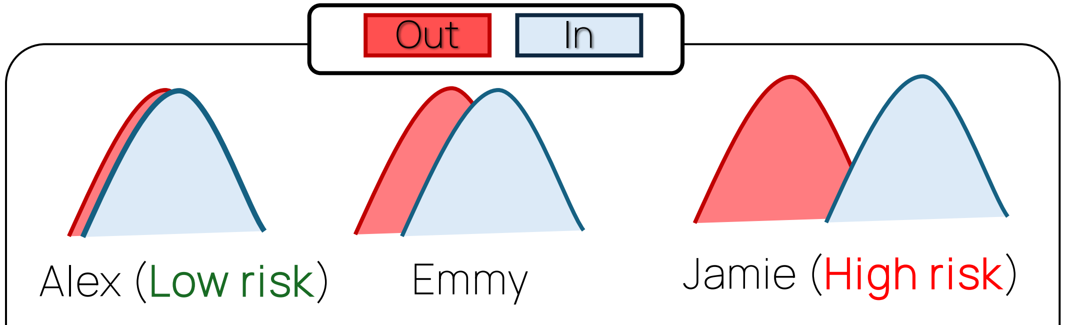
1: Users share data to train a model



2: GDPR allows some users (e.g., <1%) to request their data removed from the model

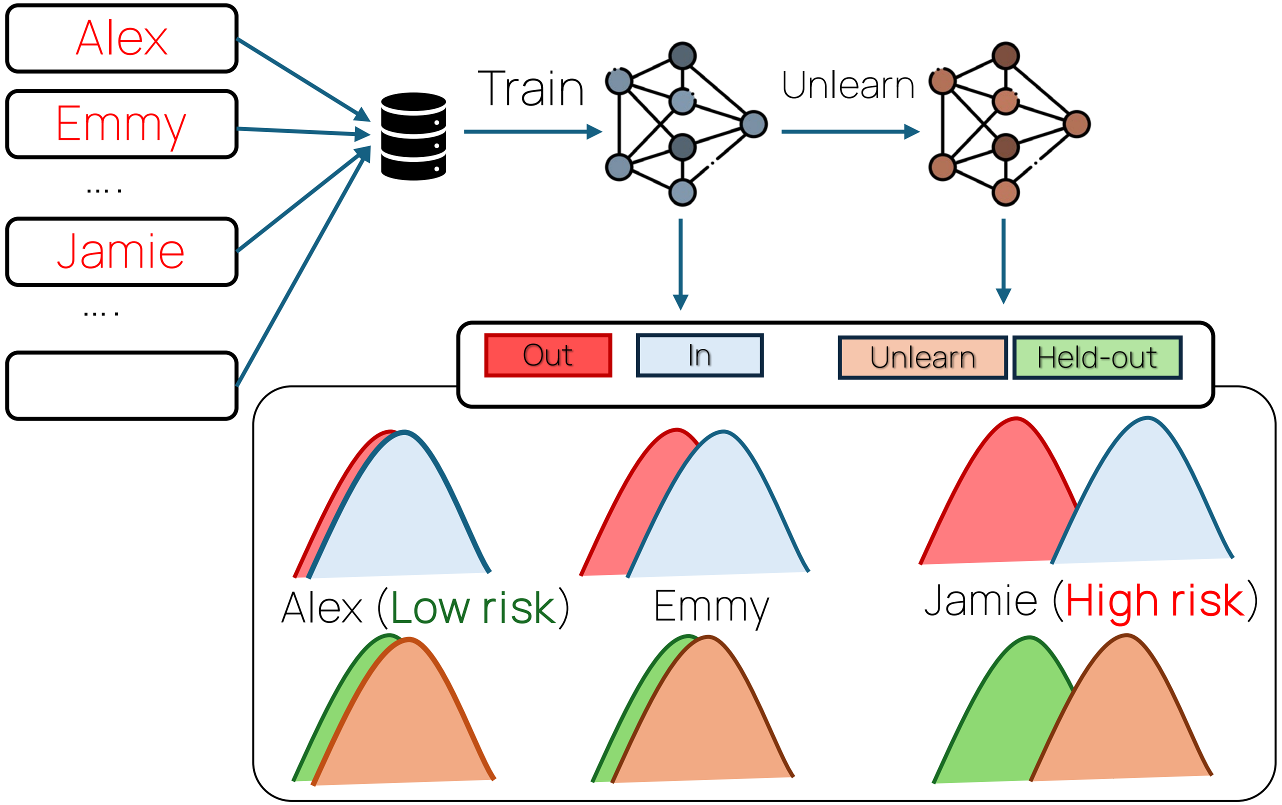


3: We have an unlearning algorithm; we need to remove the requested data to ensure GDPR compliance



Recalling: Memorization is not uniform! Some samples are more vulnerable than others.

Threat Model and Definitions



Threat Model: *adversary only has access to the final unlearned model*

In: distribution of **trained models** where a sample is *member*

Out: distribution of **trained models** where sample is *non-member*

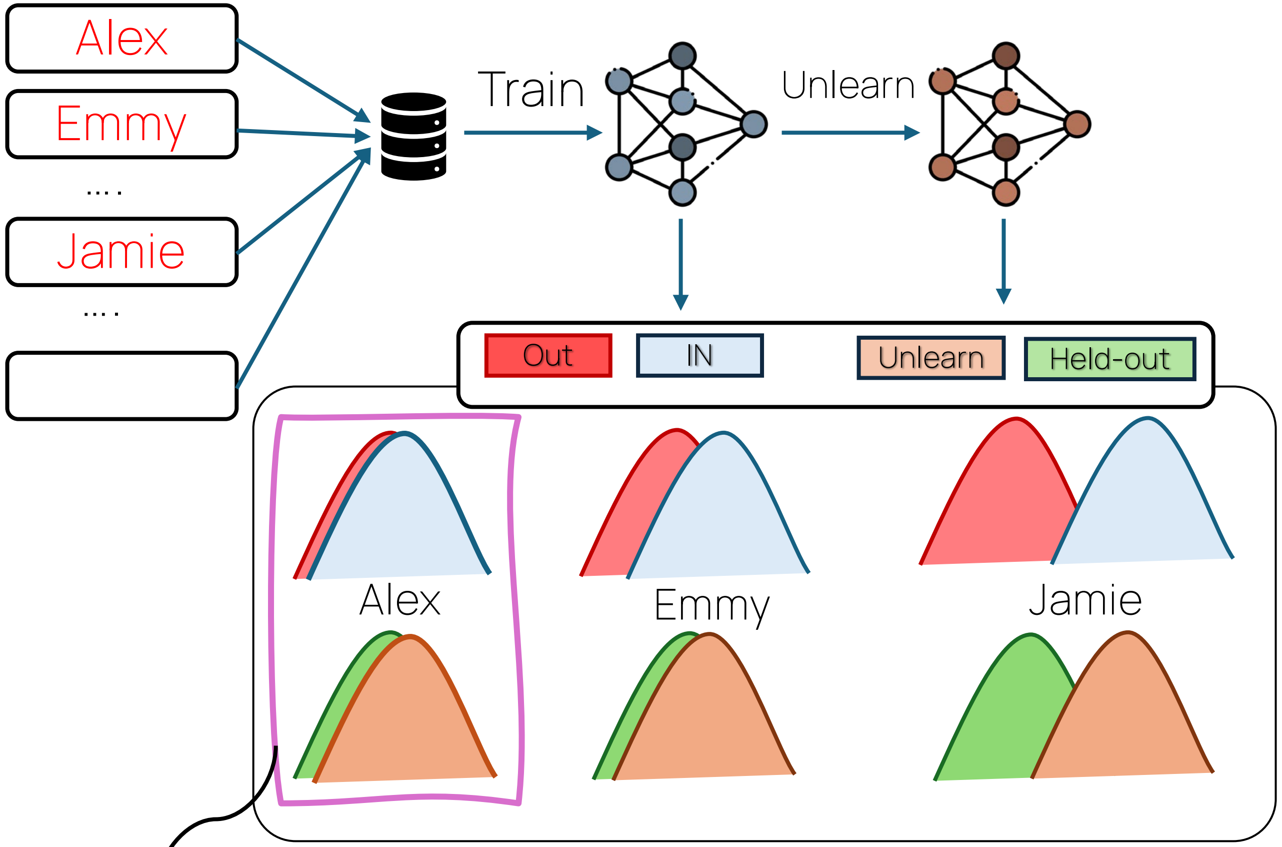
Unlearn: distribution of **unlearned models** where a sample is *unlearned*

Held-out: distribution of **unlearned models** where sample is *non-member*

If *Unlearn* ≈ *Held-out*, privacy is preserved.
"Privacy"

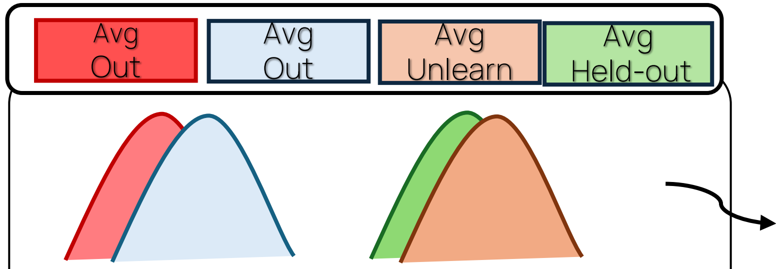
If Unlearn ≈ *Out*, unlearning is effective.
"Efficacy" (Indistinguishability to Retraining)

What is missing today



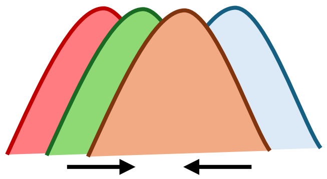
3. Many samples are like this; well- protected already.

"Let's do not evaluate them"



Average-case MIAs (or model accuracy) underestimate per-sample's unlearning requirements.

1. "Better to be per-sample like [3]"



Unlearning suppresses model output! (makes MIA-resilience, but not a removal guarantee)

2. MIA resilience differs from unlearning guarantee! Need to find a way to measure efficacy

Efficacy: "MIA to identify if any sample is unlearned or retrained"

[3] Hayes, Jamie, et al. "Inexact unlearning needs more careful evaluations to avoid a false sense of privacy." In SaTML 2025.

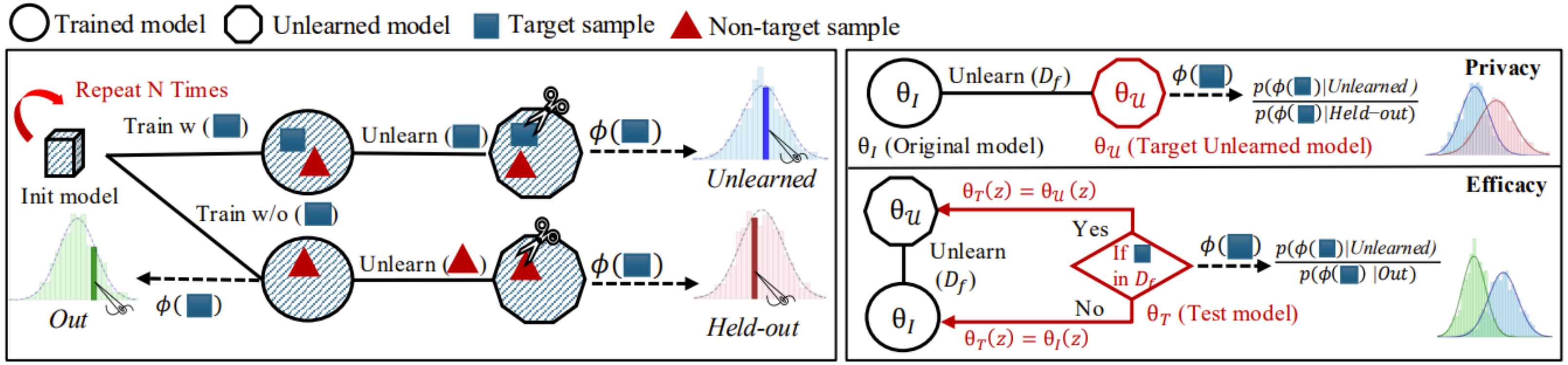
Rectified Unlearning Evaluation Framework via Likelihood Inference (RULI)

1. We introduced an algorithm to train shadow models; got all distributions required per-sample

We optimized our algorithm's parallelization to reduce shadow-model costs.

2. We introduced a hypothetic *Test model* to measure Efficacy;

This calibrates output suppression impact.



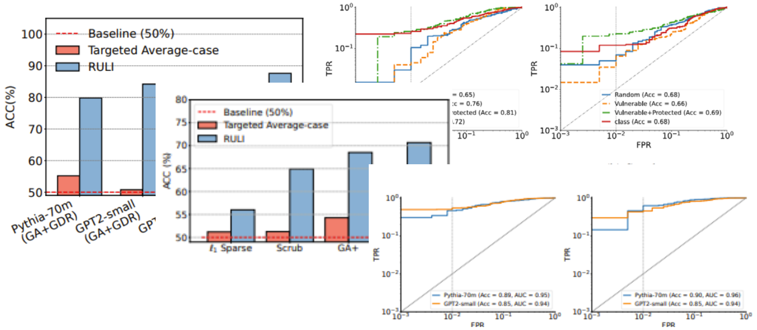
(a) Shadow model training and unlearning

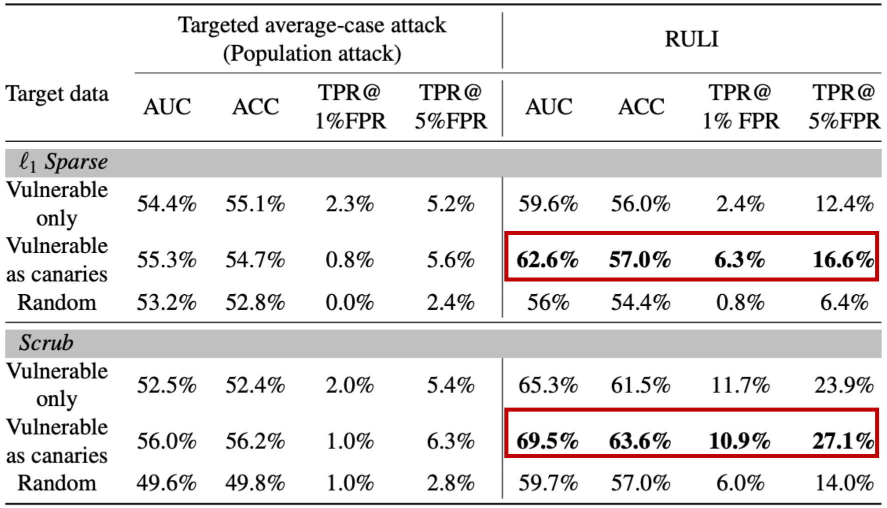
(b) Privacy and efficacy measurements

3. We targeted vulnerable sample and inject them as **canaries** to challenge/evaluate unlearning.

Our Results

- ❑ We assume we can always find the **best unlearning parameters** per unlearning request.
- ❑ Canary injection usually **leaks** more than purely unlearning vulnerable samples!
- ❑ We also tried similar experiments on CIFAR-10, CIFAR-100, and 7-gram unlearning from WikiText-103.





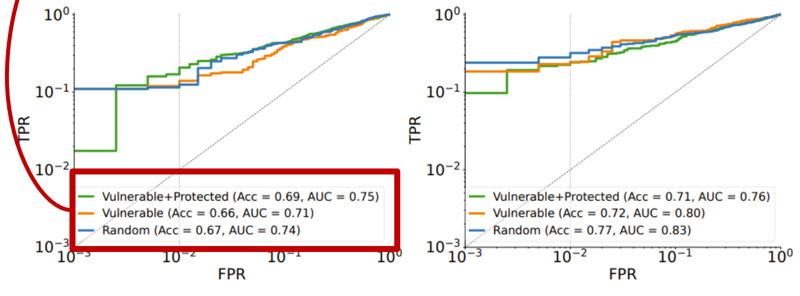
Target data	Targeted average-case attack (Population attack)				RULI			
	AUC	ACC	TPR@ 1%FPR	TPR@ 5%FPR	AUC	ACC	TPR@ 1% FPR	TPR@ 5%FPR
ℓ_1 Sparse								
Vulnerable only	54.4%	55.1%	2.3%	5.2%	59.6%	56.0%	2.4%	12.4%
Vulnerable as canaries	55.3%	54.7%	0.8%	5.6%	**62.6%**	**57.0%**	**6.3%**	**16.6%**
Random	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%

~12.6% higher MIA success 6.3x higher privacy risk than retraining

~19.5% higher MIA success; 10.9x privacy risk than retraining

Tiny ImageNet unlearning; swin-small model; unlearning <1% of the data. 500 samples: 250 Out and 250 Unlearned

Up to 69% MIA success distinguishing unlearned vs retrained



(a) ℓ_1 Sparse (b) Scrub

This is one example; further results are in the paper.

Last words …

Thanks for your attention!

More details about our design and validations?

Let's discuss this more in the following poster session

Or contact us via email: nima.naderloui@uconn.edu

Code available on: https://github.com/datasec-lab/Ruli