Rectifying Unlearning Efficacy and Privacy Evaluation: A New Inference Prospective

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?

Every model has a "fast" unlearning fix.

A large and growing body of works have been introduced for inexact selective unlearning and, improvements are incremental

Empirical evaluations indicate that unlearning is approaching seamless “perfection”.

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?

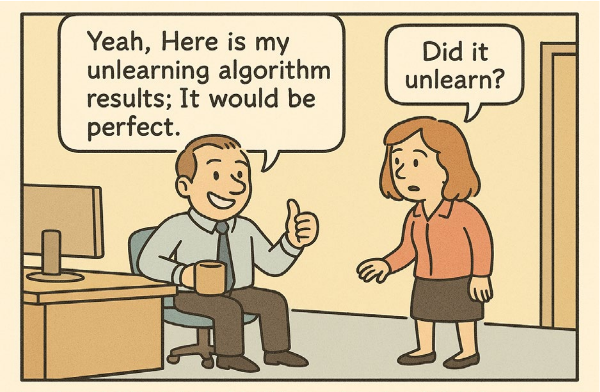
Failure of membership inference attack (MIA) → Better Forgetting [1]

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 UA	Random Forget Set MIA	Random Forget Set RA	Random Forget Set TA	Random Forget Set Avg. Gap	Worst-Case Forget Set UA	Worst-Case Forget Set MIA	Worst-Case Forget Set RA	Worst-Case Forget Set TA	Worst-Case Forget Set Avg. Gap
Retrain	$5.28_{\pm 0.33}$	$12.86_{\pm 0.61}$	$100.00_{\pm 0.00}$	$94.38_{\pm 0.15}$	0.00	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$94.66_{\pm 0.09}$	0.00
Relabeling-free										
FT	$5.08_{\pm 0.39}$ (0.20)	$10.96_{\pm 0.38}$ (1.90)	$97.46_{\pm 0.52}$ (2.54)	$91.02_{\pm 0.36}$ (3.36)	2.00	$0.00_{\pm 0.00}$ (0.00)	$0.02_{\pm 0.03}$ (0.02)	$97.63_{\pm 0.46}$ (2.37)	$91.58_{\pm 0.40}$ (3.08)	1.37
EU-k	$2.34_{\pm 0.79}$ (2.94)	$6.35_{\pm 0.89}$ (6.51)	$97.52_{\pm 0.89}$ (2.48)	$90.17_{\pm 0.88}$ (4.21)	4.04	$0.68_{\pm 0.56}$ (0.68)	$5.02_{\pm 4.42}$ (5.02)	$97.17_{\pm 0.86}$ (2.83)	$90.08_{\pm 0.70}$ (4.58)	3.28
CF-k	$0.02_{\pm 0.02}$ (5.26)	$0.76_{\pm 0.02}$ (12.10)	$99.98_{\pm 0.00}$ (0.02)	$94.45_{\pm 0.02}$ (0.07)	4.36	$0.00_{\pm 0.00}$ (0.00)	$0.00_{\pm 0.00}$ (0.00)	$99.98_{\pm 0.01}$ (0.02)	$94.34_{\pm 0.05}$ (0.32)	0.08
SCRUB	$12.42_{\pm 19.82}$ (7.14)	$22.43_{\pm 24.44}$ (9.57)	$88.31_{\pm 19.78}$ (11.69)	$83.15_{\pm 17.94}$ (11.23)	9.91	$0.01_{\pm 0.01}$ (0.01)	$0.04_{\pm 0.03}$ (0.04)	$98.65_{\pm 0.33}$ (1.35)	$92.78_{\pm 0.30}$ (1.88)	0.82
ℓ_1-sparse	$4.34_{\pm 0.73}$ (0.94)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	$96.93_{\pm 0.73}$ (3.07)	$90.96_{\pm 0.82}$ (3.70)	1.72

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



b) Using a fast inexact unlearning

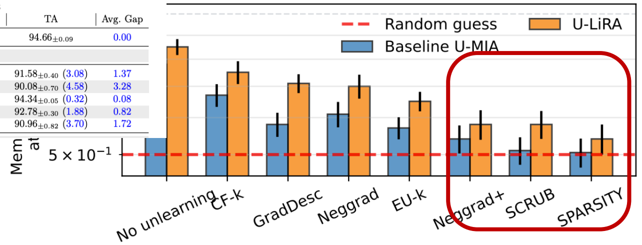
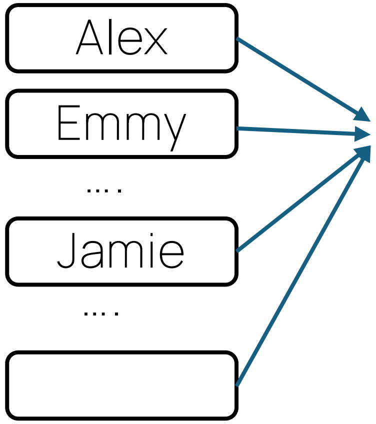


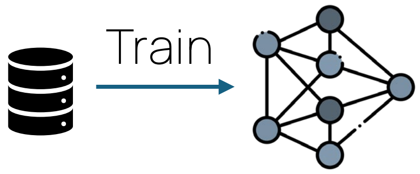
Figure 1 | *Membership inference attack accuracy using a baseline attack and U-LiRA across different unlearning algorithms. Attack and unlearning algorithm descriptions are in Section 4. U-LiRA outperforms the baseline by a large margin across all unlearning algorithms because it creates per-example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

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

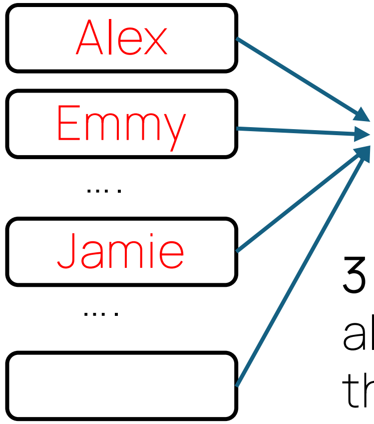
[1] Jagielski, Matthew, et al. "Measuring forgetting of memorized training examples." In ICLR 2023.
[2] Fan, Chongyu, et al. "Challenging forgets: Unveiling the worst-case forget sets in machine unlearning." In ECCV 2024.
[3] Hayes, Jamie, et al. "Inexact unlearning needs more careful evaluations to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

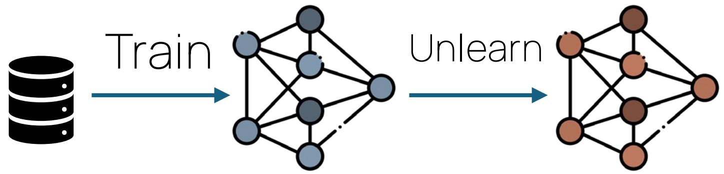
Warmup up: Our motivation



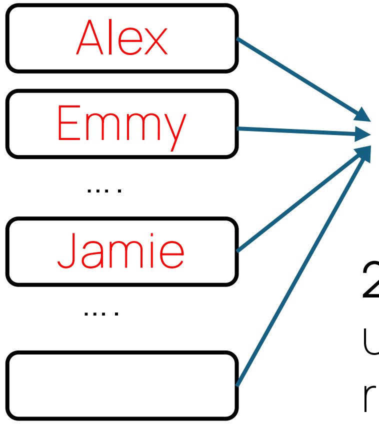


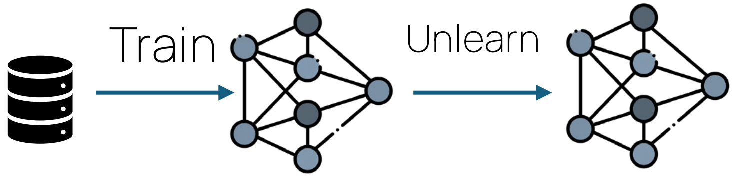
1: Users share data to train a model



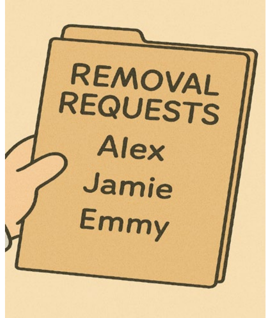


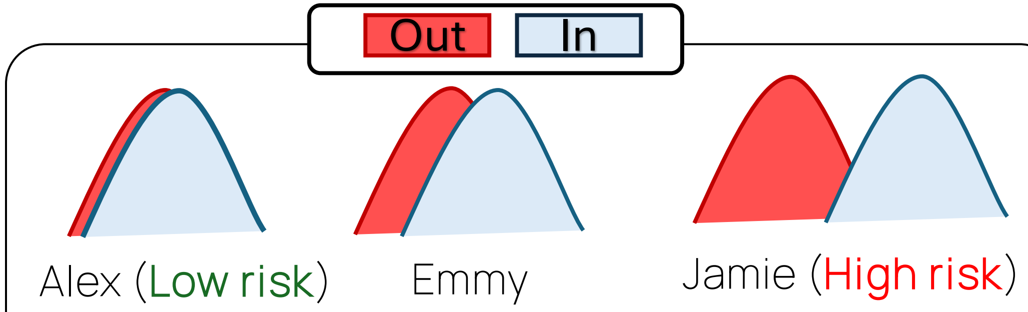
3: We have an unlearning algorithm; we need to now remove the requested data to ensure GDPR compliance





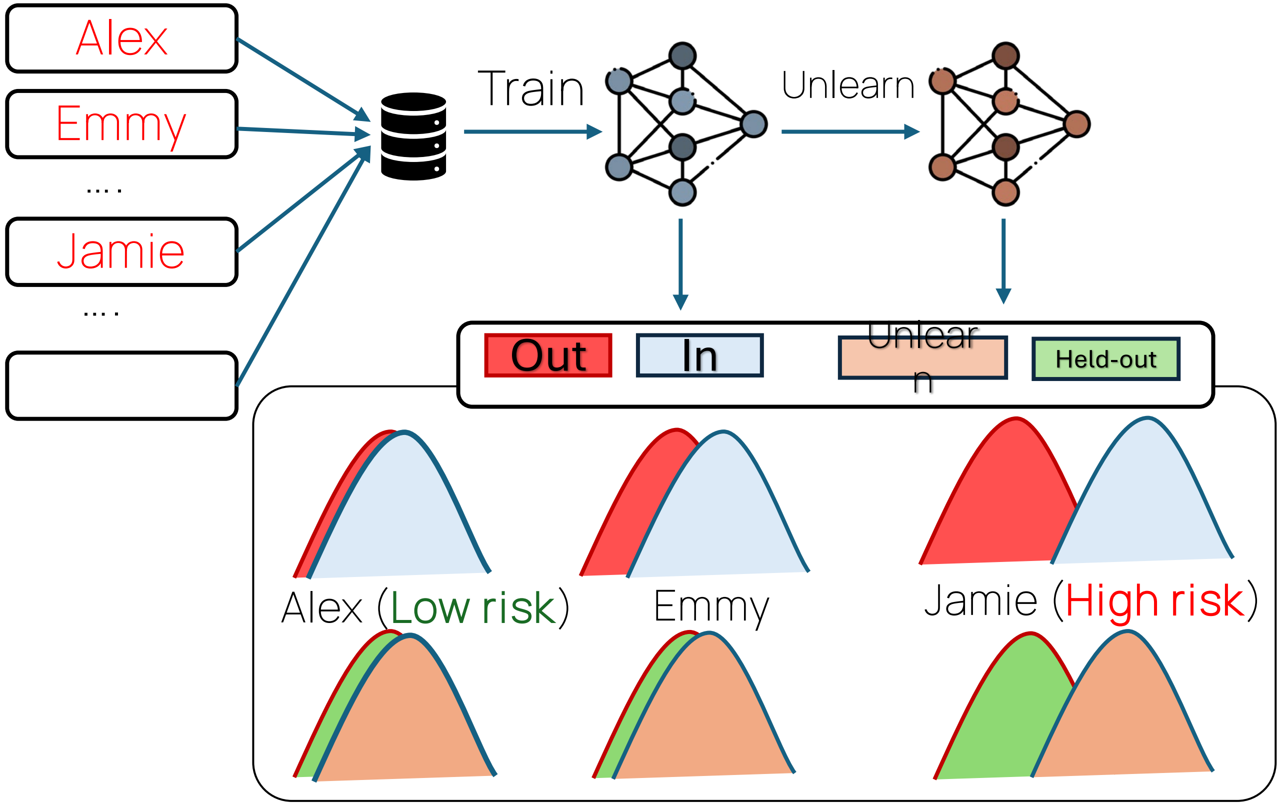
2: GDPR allows some users (e.g., <1%) to request their data removed from the model





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

What is missing today: Our motivation



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

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

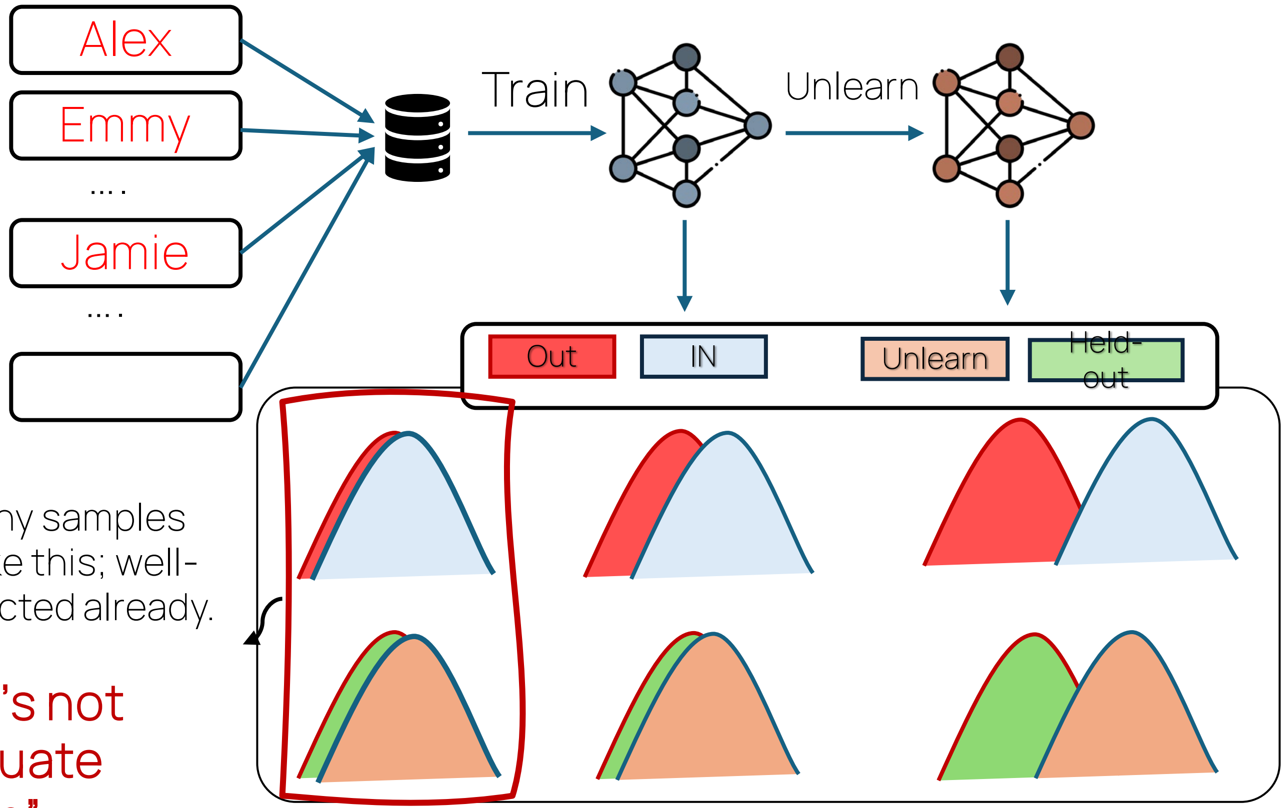
In: distribution of **trained models** where a sample is *member*

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*

What is missing today: Our motivation

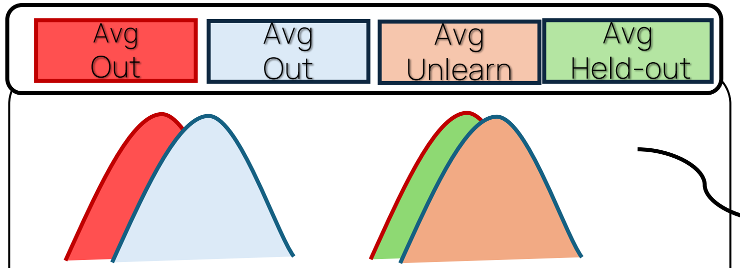


3. Many samples are like this; well-protected already.

"Let's not evaluate them"

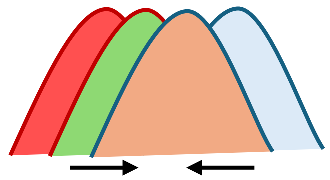
If *Unlearn* ~= *Held-out* privacy is protected (Privacy leakage)

If Unlearn ~= *Out* unlearning is effective (Efficacy) (Indistinguishability to Retraining)



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

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



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

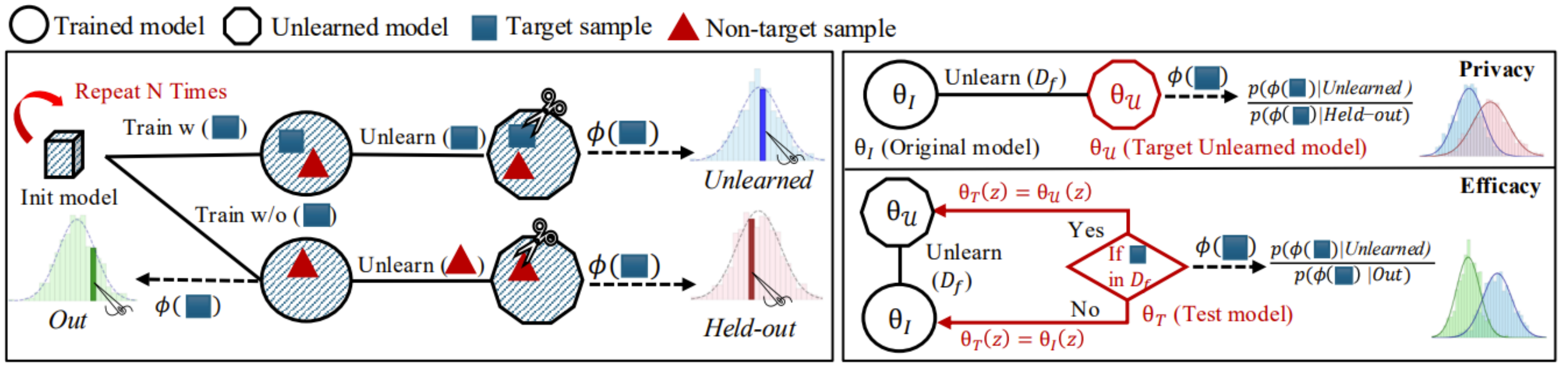
"MIA to detect if sample is unlearned or **retrained (Efficacy)**"

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

Our framework: RULI

1. We introduce an algorithm to train shadow models; get all distributions required per-sample
We tried optimizing parallelization our algorithm to minimize the shadow costs!

2. We introduce 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 target vulnerable samples; went further and inject them as **canaries** to challenge unlearning.

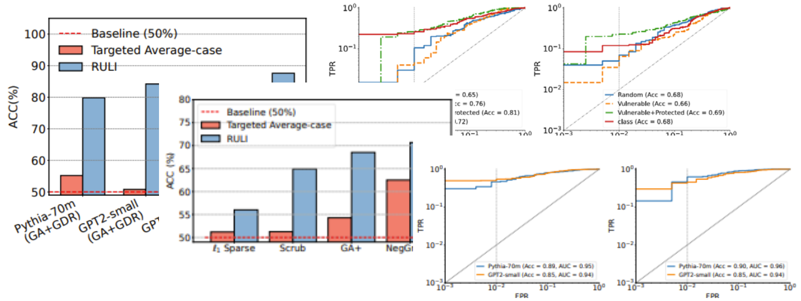
Our Results

- We performed our attack on best-inexact unlearning baselines.

- We assume we can always find **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. For generalizability, unlearning random 7-gram from GPT-2; similar trends exists!



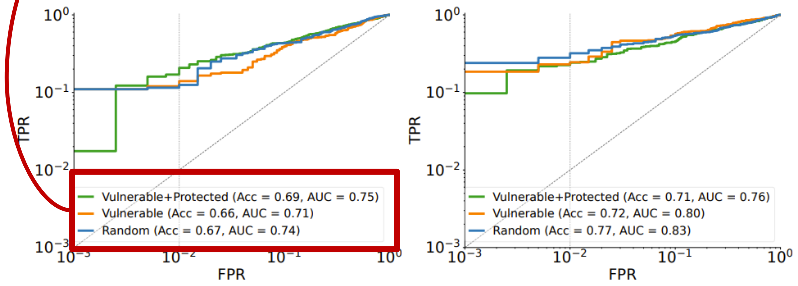
Target data	Targeted average-case attack (Population attack)				RULI			
	AUC	ACC	TPR@ 1%FPR	TPR@ 5%FPR	AUC	ACC	TPR@ 1% FPR	TPR@ 5%FPR
ℓ_1 Sparse								
Vulnerable only	54.4%	55.1%	2.3%	5.2%	59.6%	56.0%	2.4%	12.4%
Vulnerable as canaries	55.3%	54.7%	0.8%	5.6%	**62.6%**	**57.0%**	**6.3%**	**16.6%**
Random	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
x6.3 higher privacy risk than retraining

~19.5% higher MIA success;
x10.9 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

An example of our results; Please find more results available in our paper!

Last words …

- ❑ Our research *does not refute or critique* idea of inexact unlearning or existing evaluations. However, we claim:
- ❑ We need to *rectify & improve* our evaluation toward **stronger evaluations** to understand limitations of unlearning for privacy.

Thanks for your attention!

But we did not tell everything?!
Interested in more details about our design and experiments?
RULI is not perfect! Wondering about **RULI's limitations** and edge cases?

Let's discuss more in the following poster session

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