#NDSSSymposium2025

Provably Unlearnable Data Examples

Derui Wang[1,3], Minhui Xue[1,3], Bo Li[2], Seyit Camtepe[1,3], and Liming Zhu[1]

1.CSIRO's Data61, Australia

2. University of Chicago, USA

3. Cyber Security Cooperative Research Centre, Australia

Outline

- Background and Motivation
  - Learnability of public data poses risks.
  - There is NO rigid guarantee for data learnability control.
- Proposed Theory: Certified Learnability of Data
  - The first certification framework towards the effectiveness and robustness of unlearnable examples.
  - Main theorem, algorithms, and properties.
- Experiments
  - Certified learnability towards certifiable pirate models.
  - Protection against pirate models beyond the certifiable ones.
- Applications and Things to be Improved

Learnability of Public Data Poses Risks

2017 IEEE Symposium on Security and Privacy

Membership Inference Attacks Against Machine Learning Models

Reza Shokri
Cornell Tech
shokri@cornell.edu

Marco Stronati*
INRIA
marco@stronati.org

Congzheng Song
Cornell
cs2296@cornell.edu

Vitaly Shmatikov
Cornell Tech
shmat@cs.cornell.edu

Published as a conference paper at ICLR 2017

DELVING INTO TRANSFERABLE ADVERSARIAL EXAMPLES AND BLACK-BOX ATTACKS

Yanpei Liu*, Xinyun Chen*
Shanghai Jiao Tong University

Chang Liu, Dawn Song
University of the California, Berkeley

Pirate ML models are trained on public data to act as adversarial domain experts to expose proprietary/sensitive knowledge.

Data Exploitation will Become Easier

Fine-Tuning Llama 3 and Using It Locally: A Step-by-Step Guide

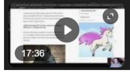
We'll fine-tune Llama 3 on a dataset of patient-doctor conversations, creating a model tailored for medical dialogue. After merging, converting, and quantizing the model, it will be ready for private local use via the Jan application.

May 30, 2024 · 19 min read

Fine-Tuning DeepSeek-R1 on Consumer Hardware: A Step-by-Step Guide 🤖 ✨ 🔥

Mistral-7B Fine-Tuning: A Step-by-Step Guide

Fine-tune Stable Diffusion with LoRA for as low as $1



Fine-tuning large models doesn't have to be complicated and expensive. In this tutorial, I provide a step-by-step demonstration of the ...

YouTube · Julien Simon · 9 Oct 2023

Pirate ML models are trained on public data to act as adversarial domain experts to expose proprietary/sensitive knowledge. Open LLMs/VLMs/LVMs and PEFT may make it worse.

Data Exploitation will Become Easier

Fine-Tuning Llama 3 and Using It Locally: A Step-by-Step Guide

We'll fine-tune Llama 3 on a dataset of patient-doctor conversations, creating a model tailored for medical dialogue. After merging, converting, and quantizing the model, it will be ready for private local use via the Jan application.

May 30, 2024 · 19 min read

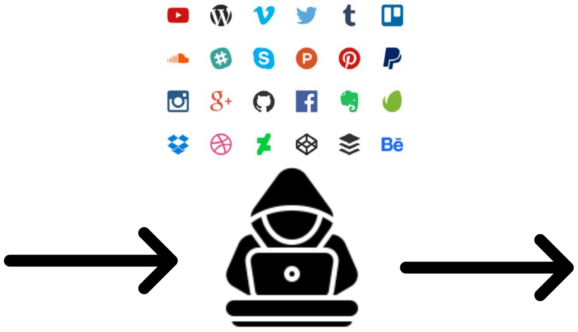
Fine-Tuning DeepSeek-R1 on Consumer Hardware: A Step-by-Step Guide 🤖 ✨ 🔥

Mistral-7B Fine-Tuning: A Step-by-Step Guide

Fine-tune Stable Diffusion with LoRA for as low as $1

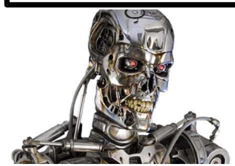
Fine-tuning large models doesn't have to be complicated and expensive. In this tutorial, I provide a step-by-step demonstration of the ...

17:36 YouTube · Julien Simon · 9 Oct 2023



?

Less ~~data~~ is More ~~harm~~!



Pirate ML models are trained on public data to act as adversarial domain experts to expose proprietary/sensitive knowledge. Open LLMs/VLMs/LVMs and PEFT may make it worse.

Data Exploitation will Become Easier

Fine-Tuning Llama 3 and Using It Locally: A Step-by-Step Guide

We'll fine-tune Llama 3 on a dataset of patient-doctor conversations, creating a model tailored for medical dialogue. After merging, converting, and quantizing the model, it will be ready for private local use via the Jan application.

May 30, 2024 · 19 min read

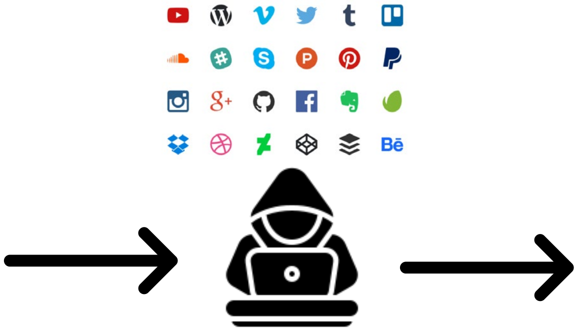
Fine-Tuning DeepSeek-R1 on Consumer Hardware: A Step-by-Step Guide 🤖 ✨ 🔥

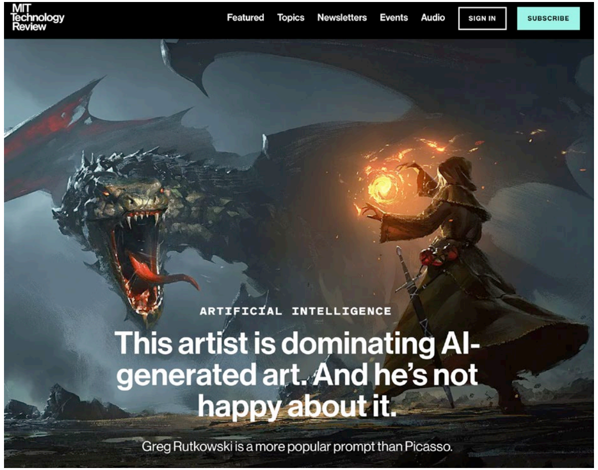
Mistral-7B Fine-Tuning: A Step-by-Step Guide

Fine-tune Stable Diffusion with LoRA for as low as $1

Fine-tuning large models doesn't have to be complicated and expensive. In this tutorial, I provide a step-by-step demonstration of the ...

YouTube · Julien Simon · 9 Oct 2023





6,651 reads

Creating a Domain Expert LLM: A Guide to Fine-Tuning

Pirate ML models are trained on public data to act as adversarial domain experts to expose proprietary/sensitive knowledge. Open LLMs/VLMs/LVMs and PEFT may make it worse.

Legislation from Various Nations Says No, But……

General Data Protection Regulation

GDPR

EU Artificial Intelligence Act

samtec®

California Privacy Notice

Administration

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

Voluntary AI Safety Standard

Guiding safe and responsible use of artificial intelligence in Australia

Date published: 5 September 2024

人工智能安全治理框架

AI Safety Governance Framework

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC
2024年9月

There is NO strict guarantee on the learnability of the data.

Data Learnability Control via Unlearnable Examples*

+ Data points are perturbed to unlearnable examples (UEs) before publication.

+ ML models trained on UEs perform poorly on unperturbed data (*i.e.*, low utility).

* This line of work is also referred to as "Perturbative Availability Poison" or "Shortcut Learning".

Data Learnability Control via Unlearnable Examples*

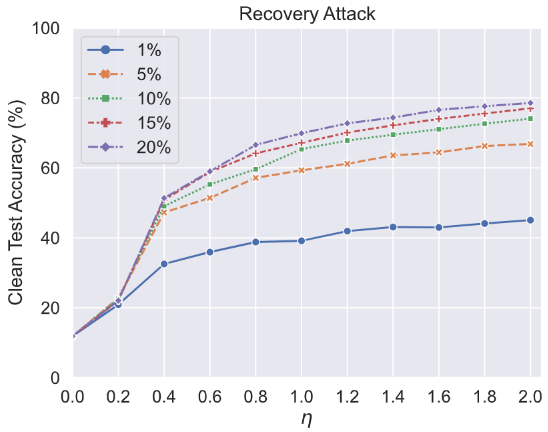
+ Data points are perturbed to unlearnable examples (UEs) before publication.

+ ML models trained on UEs perform poorly on unperturbed data (*i.e.*, low utility).

- Generalization issues caused by diverse pirate models and training strategies:
  No rigid guarantee on the maximally attainable learning results for adversaries.

- Evaluating UEs is challenging due to training stochasticity:
  Testing accuracy alone is insufficient!

- A new threat: Recovery Attack.



An attacker can slightly perturb the weights of a pirate model trained on UEs using projected SGD and 5%~10% of clean data to restore its utility.

* This line of work is also referred to as "Perturbative Availability Poison" or "Shortcut Learning".

Data Learnability Control via Unlearnable Examples*

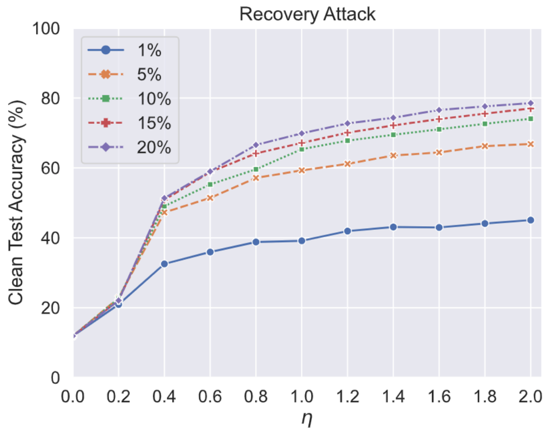
+ Data points are perturbed to unlearnable examples (UEs) before publication.

+ ML models trained on UEs perform poorly on unperturbed data (*i.e.*, low utility).

- Generalization issues caused by diverse pirate models and training strategies:
  No rigid guarantee on the maximally attainable learning results for adversaries.

- Evaluating UEs is challenging due to training stochasticity:
  Testing accuracy alone is insufficient!

- A new threat: Recovery Attack.



An attacker can slightly perturb the weights of a pirate model trained on UEs using projected SGD and 5%~10% of clean data to restore its utility.

* This line of work is also referred to as "Perturbative Availability Poison" or "Shortcut Learning".

Unlearnable examples are vulnerable.

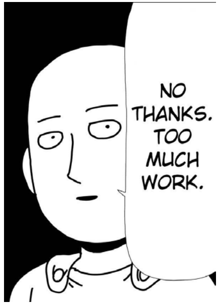
Certified Data Learnability Control is Needed

Given a set of UEs, can we establish an upper bound on the utility of pirate models trained on them?

Certified Data Learnability Control is Needed

Given a set of UEs, can we establish an upper bound on the utility of pirate models trained on them?

Training with different models and learning algorithms millions of times to estimate it is not a wise move!



~~Influence Function, Shapley Value, ……~~
No existing tools are available.

Certified Data Learnability Control is Needed

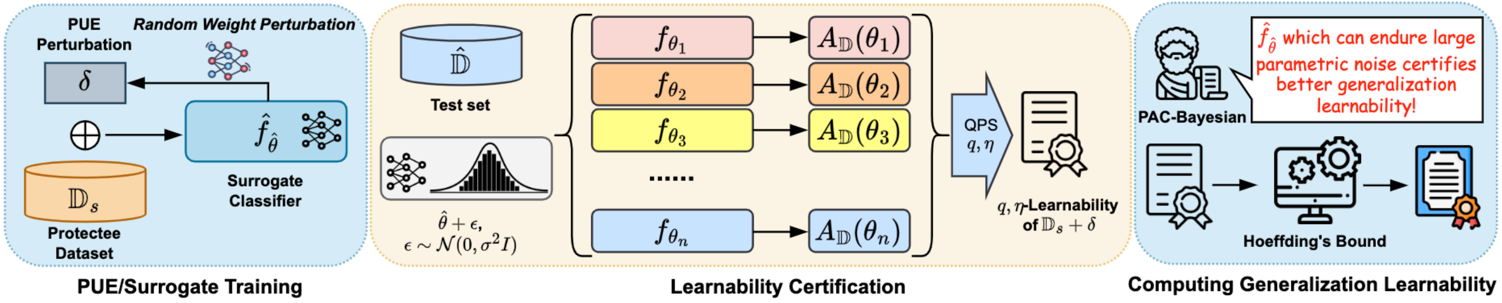
Given a set of UEs, can we establish an upper bound on the utility of pirate models trained on them?

Training with different models and learning algorithms millions of times to estimate it is not a wise move!

It is possible to derive such an upper bound guaranteed with a high probability, for pirate classifiers under some constraints.

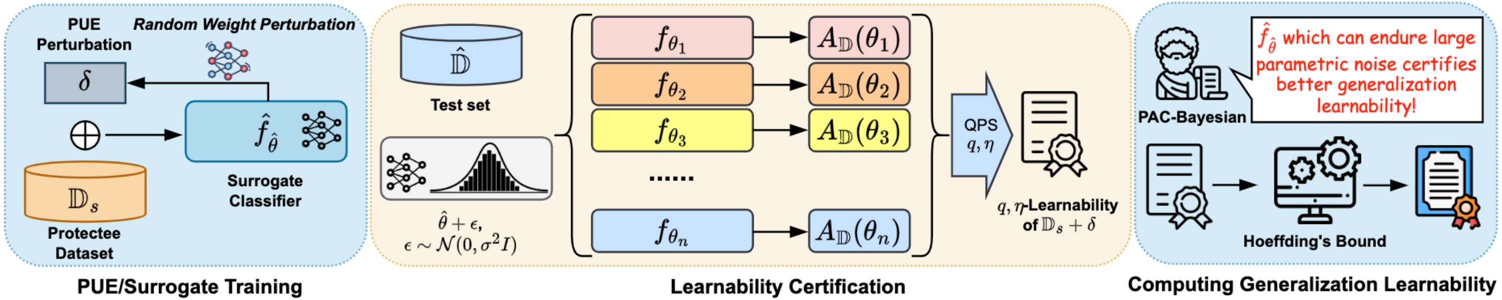
Certified Data Learnability

- Framework:



Certified Data Learnability

- Framework:



Definition 3 ((q,η)-*Learnability*). *Suppose a learning function Γ selects $\hat{\theta}$ based on an unlearnable dataset $\mathbb{D}_s \oplus \delta$. The certified (q,η)-Learnability of $\mathbb{D}_s \oplus \delta$ is*

$$l_{(q,\eta)}(\hat{\Theta};\mathbb{D}_s \oplus \delta) = \inf \{t \mid \Pr_{\epsilon}[A_{\mathbb{D}}(\hat{\theta}+\epsilon) \leq t] \geq \overline{q}\}, \quad (10)$$

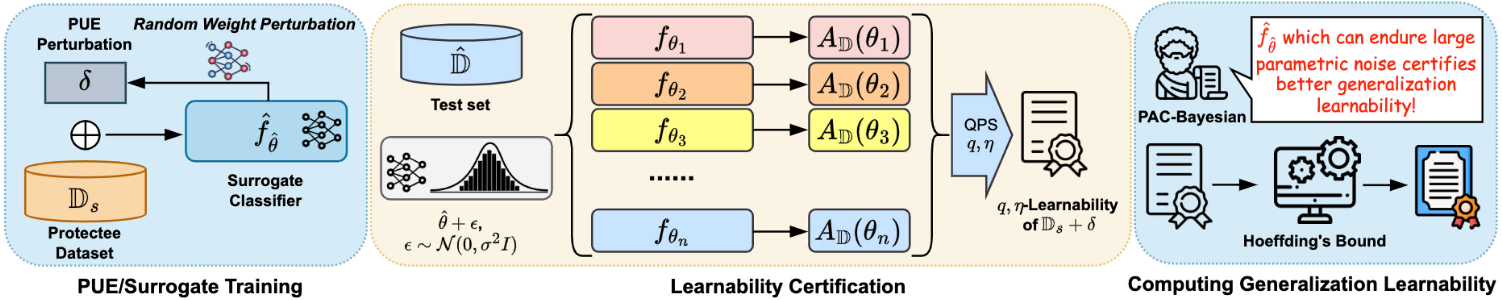
Guarantee

where $\overline{q} = \Phi(\Phi^{-1}(q) + \frac{\eta}{\sigma})$. For any θ^* drawn from the certified parameter set $\hat{\Theta} := \{\theta \mid \theta \sim \mathcal{N}(\hat{\theta}+\upsilon, \sigma^2 I),\ \|\upsilon\| \leq \eta\}$, there is $A_{\mathbb{D}}(\theta^*) \leq l_{(q,\eta)}(\hat{\Theta};\mathbb{D}_s \oplus \delta)$ with probability no less than q.

Constraint

Certified Data Learnability

- Framework:



Definition 3 ((q,η)-*Learnability*). *Suppose a learning function* Γ *selects* $\hat{\theta}$ *based on an unlearnable dataset* $\mathbb{D}_s \oplus \delta$. *The certified* (q,η)-*Learnability of* $\mathbb{D}_s \oplus \delta$ *is*

$$l_{(q,\eta)}(\hat{\Theta};\mathbb{D}_s \oplus \delta) = \inf \{t | \Pr_{\epsilon}[A_{\mathbb{D}}(\hat{\theta}+\epsilon) \le t] \ge \overline{q}\}, \quad (10)$$

Guarantee

where $\overline{q} = \Phi(\Phi^{-1}(q) + \frac{\eta}{\sigma})$. *For any* θ^* *drawn from the certified parameter set* $\hat{\Theta} := \{\theta \mid \theta \sim \mathcal{N}(\hat{\theta}+\upsilon, \sigma^2 I),\ \|\upsilon\| \le \eta\}$, *there is* $A_{\mathbb{D}}(\theta^*) \le l_{(q,\eta)}(\hat{\Theta};\mathbb{D}_s \oplus \delta)$ *with probability no less than* q.

Constraint

This is the first certification framework towards the effectiveness and robustness of UEs.

How to Certify It

- Main theorem:

The QPS function computes the model utility at a given quantile (q) among all attainable model utilities.

Definition 2 (*Quantile Parametric Smoothing function*)**.** *Given a dataset $\mathbb{D}$ from the space $\mathcal{X} \times \mathcal{Y}$, an $\mathbb{D}$-parameterized function $A_{\mathbb{D}} : \Theta \to [0,1]$ with an input $\theta \in \Theta$, and a parametric smoothing noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ under a standard deviation of σ, a Quantile Parametric Smoothing function $h_q(\theta)$ is defined as:*

$$h_q(\theta) = \inf \{t \mid \Pr_{\epsilon}[A_{\mathbb{D}}(\theta + \epsilon) \leq t] \geq q\}, \tag{7}$$

where $q \in [0,1]$ is a probability.

How to Certify It

- Main theorem:

The QPS function computes the model utility at a given quantile (q) among all attainable model utilities.

Definition 2 (*Quantile Parametric Smoothing function*). *Given a dataset $\mathbb{D}$ from the space $\mathcal{X} \times \mathcal{Y}$, an $\mathbb{D}$-parameterized function $A_{\mathbb{D}} : \Theta \to [0,1]$ with an input $\theta \in \Theta$, and a parametric smoothing noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ under a standard deviation of σ, a Quantile Parametric Smoothing function $h_q(\theta)$ is defined as:*

$$h_q(\theta) = \inf \{t \mid \Pr_{\epsilon}[A_{\mathbb{D}}(\theta + \epsilon) \leq t] \geq q\}, \tag{7}$$

where $q \in [0,1]$ is a probability.

The perturbation bound of QPS certifies the learnability within a certified parameter set that scales with η.

Theorem 1 (*Perturbation bound on QPS*). *Let $\Gamma : \mathcal{X} \times \mathcal{Y} \to \hat{\theta} \in \Theta$ be a learning function selecting $\hat{\theta}$ from the parameter space Θ based on a dataset defined in $\mathcal{X} \times \mathcal{Y}$. Given an target dataset $\mathbb{D}$ and a quantile smoothed function $h_q(\hat{\theta})$ centered at a Gaussian $\mathcal{N}(\hat{\theta}, \sigma^2 I)$, then there exists an upper bound for $h_q(\hat{\theta} + \upsilon)$. Specifically,*

$$h_q(\hat{\theta} + \upsilon) \leq \inf \{t \mid \Pr_{\epsilon}[A_{\mathbb{D}}(\hat{\theta} + \epsilon) \leq t] \geq \overline{q}\}, \ \forall \ \|\upsilon\| \leq \eta, \tag{9}$$

where $\overline{q} := \Phi(\Phi^{-1}(q) + \frac{\eta}{\sigma})$. $\Phi(\cdot)$ is the standard Gaussian CDF and $\Phi^{-1}(\cdot)$ is the inverse of the CDF. $\|\upsilon\|$ is the ℓ_2 norm of the parameter shift υ from $\hat{\theta}$.

How to Certify It (cont'd)

- Certification algorithms:

Algorithm 1: Quantile Upper Bound

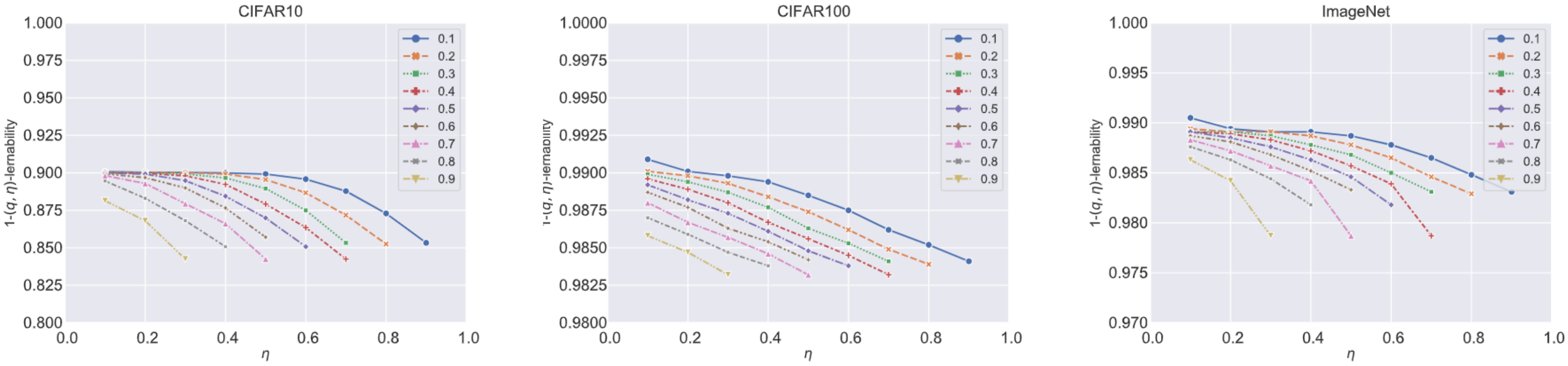
func QUPPERBOUND
Input: noise draws n, α, σ, η, quantile q.
Output: Index of the value in the q-th quantile
$\overline{q} \leftarrow \Phi(\Phi^{-1}(q) + \frac{\eta}{\sigma})$
$\underline{k},\ \overline{k} \leftarrow \lceil n * \overline{q} \rceil,\ n$
$k^* \leftarrow 0$
for $k \in \{\underline{k}, \underline{k}+1, ..., \overline{k}\}$ **do**
  if BINOMIAL$(n, k, \overline{q}) > 1 - \alpha$ **then**
    $k^* \leftarrow k$
  else
    Continue
if $k^* \neq 0$ **then**
  Output: k^*
else
  ABSTAIN
func BINOMIAL
Input: Sampling number n, k, $\overline{q}$.
CONF $\leftarrow \sum_{i=1}^{k} \binom{n}{k} (\overline{q})^i (1-\overline{q})^{n-i}$
Output: CONF

Algorithm 2: (q, η)-Learnability Certification

Input: Accuracy function $A_{\mathbb{D}}$, surrogate weights $\hat{\theta}$, test set $\mathbb{D}$, n, σ, q, η.
Output: (q, η)-Learnability
Initialize a
for $i \in 1, ..., n$ **do**
  $\theta \leftarrow \hat{\theta} + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2 I)$
  Evaluate $A_{\mathbb{D}}(\theta)$ on $\mathbb{D}$
  Append $A_{\mathbb{D}}(\theta)$ to a
$a \leftarrow Sort(a)$
$k \leftarrow$ QUPPERBOUND$(n, \alpha, \sigma, \eta, q)$
$t \leftarrow a_k$
Output: t

Check our paper for more details.

Properties of the Certification



Larger values of q and η help certify higher learnability.

#NDSSSymposium2025

Certification Results

- We propose Random Weight Perturbation (RWP) for certification surrogate training:

TABLE I: Certified (q, η)-Learnability under Different Training Methods (%, $\sigma = 0.25$)

Data	Method	$\eta \times 100$								
		0.1	0.5	1.0	5.0	10.0	15.0	20.0	25.0	30.0
CIFAR10	PUE-B	**10.62**	**10.67**	**10.71**	**11.07**	**11.86**	**12.50**	**13.20**	**14.67**	**15.75**
	EMN	5.69	5.70	5.74	5.91	6.24	6.43	6.96	8.61	10.27
CIFAR100	PUE-B	**1.32**	**1.32**	**1.33**	**1.37**	**1.41**	**1.47**	**1.53**	**1.59**	**1.68**
	EMN	0.43	0.43	0.44	0.47	0.52	0.59	0.70	0.77	0.89
ImageNet	PUE-B	**1.46**	**1.46**	**1.48**	**1.54**	**1.63**	**1.79**	**1.85**	**1.97**	**2.13**
	EMN	1.34	1.34	1.37	1.41	1.45	1.49	1.54	1.58	1.67

TABLE II: Certified (q, η)-Learnability under Different Training Methods (%, $\sigma = 0.8$)

Data	Method	η									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CIFAR10	PUE-B	**10.68**	**10.86**	**11.18**	**11.37**	**11.52**	**12.00**	**12.71**	**12.87**	**13.72**	**15.17**
	EMN	5.86	6.11	6.28	6.52	7.00	7.27	7.36	7.87	8.38	9.02
CIFAR100	PUE-B	**1.13**	**1.15**	**1.16**	**1.20**	**1.23**	**1.27**	**1.31**	**1.36**	**1.51**	**1.56**
	EMN	0.47	0.48	0.51	0.55	0.59	0.63	0.66	0.67	0.69	0.70
ImageNet	PUE-B	**1.19**	**1.24**	**1.26**	**1.28**	**1.32**	**1.37**	**1.45**	**1.50**	**1.61**	**1.80**
	EMN	1.12	1.14	1.16	1.20	1.25	1.29	1.40	1.40	1.48	1.48

RWP surrogates (*PUE-B*) produce higher certified (q, η)-Learnability on the same set of UEs

RWP creates better certification surrogates.

Certification Results

- Provably Unlearnable Examples (PUEs) are generated based on the online surrogate:

TABLE III: Certified (q, η)-Learnability under Different PAP Noises (%, $\sigma = 0.25$, online)

Data	Method	$\eta \times 100$								
		0.1	0.5	1.0	5.0	10.0	15.0	20.0	25.0	30.0
CIFAR10	PUE-10	**10.24**	**10.29**	**10.31**	**10.69**	**11.14**	**11.82**	**12.29**	**13.12**	**13.59**
	PUE-1	10.86	10.97	11.04	11.62	12.12	12.64	13.35	14.22	14.66
	PUE-B	10.62	10.67	10.71	11.07	11.86	12.50	13.20	14.67	15.75
CIFAR100	PUE-10	**1.29**	**1.29**	**1.31**	**1.35**	**1.41**	**1.43**	**1.46**	**1.51**	**1.65**
	PUE-1	1.35	1.36	1.36	1.42	1.48	1.53	1.57	1.69	1.87
	PUE-B	1.32	1.32	1.33	1.37	**1.41**	1.47	1.53	1.59	1.68
ImageNet	PUE-10	**1.45**	**1.45**	**1.45**	**1.50**	**1.61**	**1.68**	**1.76**	**1.84**	**1.95**
	PUE-1	1.58	1.58	1.58	1.67	1.73	1.81	1.95	2.00	2.19
	PUE-B	1.46	1.46	1.48	1.54	1.63	1.79	1.85	1.97	2.13

TABLE IV: Certified (q, η)-Learnability under Different PAP Noises (%, $\sigma = 0.8$, online)

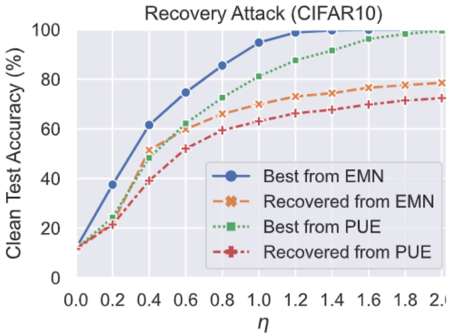
Data	Method	η									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CIFAR10	PUE-10	**10.10**	**10.24**	**10.42**	**10.80**	**11.10**	**11.57**	**11.89**	**12.45**	**12.81**	13.72
	PUE-1	10.60	10.73	11.07	11.21	11.38	11.67	12.04	12.71	13.00	**13.59**
	PUE-B	10.68	10.86	11.18	11.37	11.52	12.00	12.71	12.87	13.72	15.17
CIFAR100	PUE-10	**1.11**	**1.14**	**1.16**	**1.20**	**1.23**	**1.25**	**1.27**	**1.35**	1.48	**1.52**
	PUE-1	1.13	1.15	1.18	1.21	1.24	1.27	1.31	**1.35**	**1.44**	1.60
	PUE-B	1.13	1.15	**1.16**	**1.20**	1.23	1.27	1.31	1.36	1.51	1.56
ImageNet	PUE-10	**1.17**	**1.19**	**1.24**	**1.26**	**1.28**	**1.32**	**1.37**	**1.45**	**1.54**	**1.61**
	PUE-1	1.22	1.24	1.30	1.35	1.37	1.41	1.45	1.52	**1.54**	1.68
	PUE-B	1.19	1.24	1.26	1.28	1.32	1.37	1.45	1.50	1.61	1.80

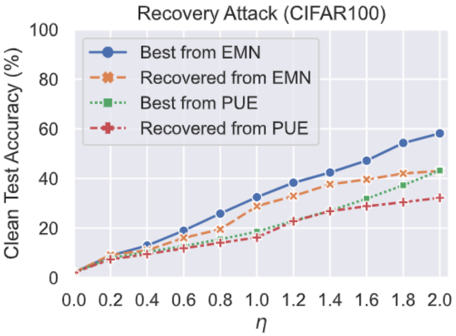
Under the same training method, PUEs (*PUE-10*) have lower certified (q, η)-Learnability compared to the baseline (*PUE-B*)

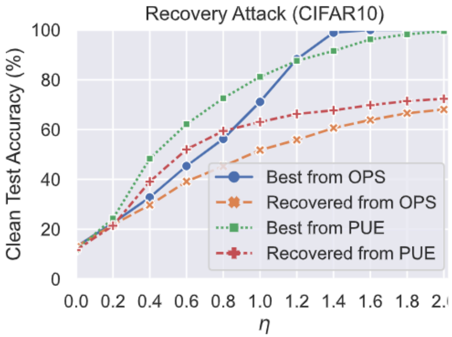
PUEs lead to more robust protection against certifiable adversaries.

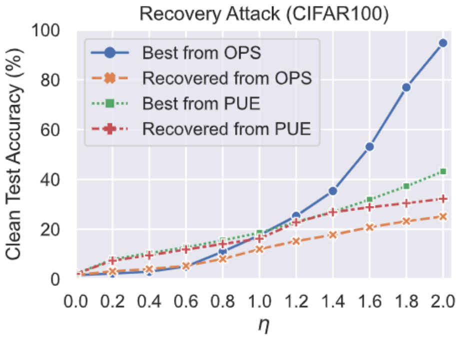
Protection Against Pirate Classifiers Beyond the Certified Parameter Set

- Hardness results of recovery attacks:









PUEs are also more robust against general adversaries.

Applying Our Work and Directions for Improvements

- Applications:
  - ✓ Serve as an **evaluation metric** for UEs.
  - ✓ Provide a **data availability guarantee** before publishing the data.
  - ✓ Generate **robust PUEs**.

Applying Our Work and Directions for Improvements

- Applications:
  - ✓ Serve as an **evaluation metric** for UEs.
  - ✓ Provide a **data availability guarantee** before publishing the data.
  - ✓ Generate **robust PUEs**.

- Directions for Improvements:
  - ❑ **Coverage**: Expand the certified parameter set.
  - ❑ **Tightness**: Certify higher learnability.
  - ❑ **Efficiency**: Accelerate PUE generation and surrogate training.

Applying Our Work and Directions for Improvements

- Applications:
  - ✓ Serve as an **evaluation metric** for UEs.
  - ✓ Provide a **data availability guarantee** before publishing the data.
  - ✓ Generate **robust PUEs**.

- Directions for Improvements:
  - ❑ **Coverage**: Expand the certified parameter set.
  - ❑ **Tightness**: Certify higher learnability.
  - ❑ **Efficiency**: Accelerate PUE generation and surrogate training.

Thank you for your attention!
Questions?

Derui (Derek) Wang
Research Scientist | **Data61, CSIRO** Australia's National Science Agency
derek.wang@data61.csiro.au