CSCI 699: Privacy Preserving Machine Learning - Week 5

Privacy Auditing and Membership Inference

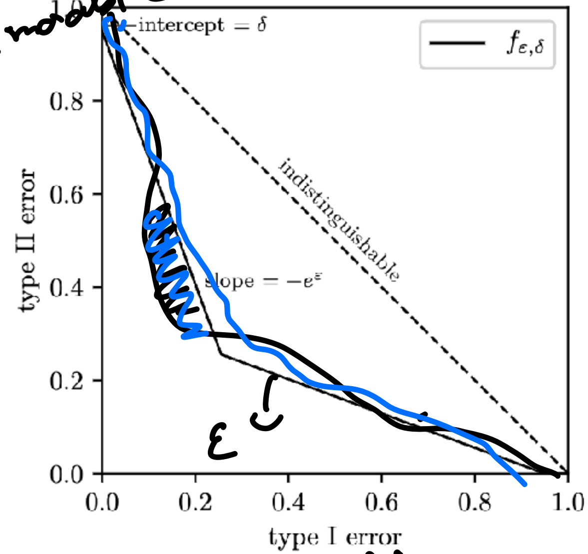
Sai Praneeth Karimireddy, Sep 29 2025

Recap: Privacy Auditing

(x', y') — added mislabelled

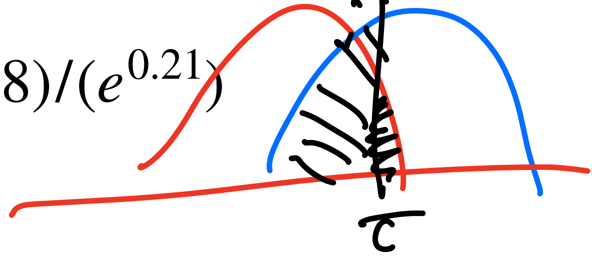
$D \xrightarrow{+} D'$

10K 10K



- We have claimed β = 0.00174 and
  α = 1- 4.922/100 = 0.95078.
- We have claimed privacy of (0.21, 10–5)-DP. $\ell(x', y')$
- $\beta \geq \max(1 - 10^{-5} - e^{0.21} 0.95078\ ,\ (1 - 10^{-5} - 0.95078)/(e^{0.21}))$
  = 0.03988885074

Exact method

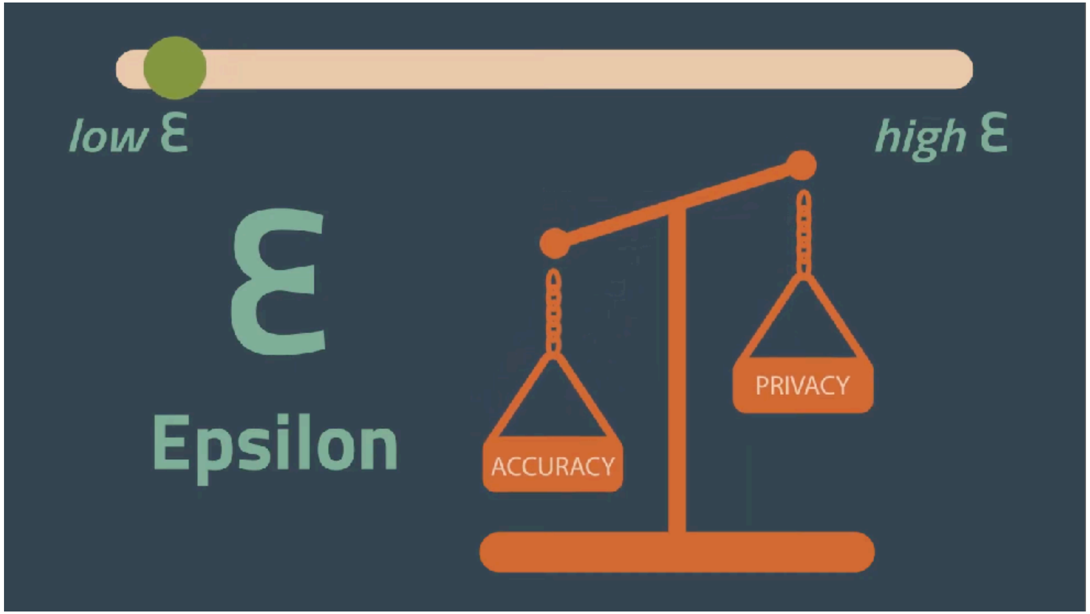


- Can be due to sampling?
- By Clopper-Pearson, $\alpha^{+} \leq$ 0.95509, $\beta^{-} \geq$ 0.00274 with $p = 10^{-10}$
- Later, they found a bug and retracted the paper. Very common in DP!!

Overview

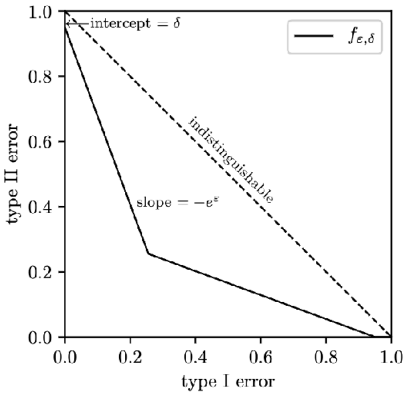
- Efficient Privacy Auditing
- Relaxations of Differential Privacy
  - memorization
- Unlearning
- Local DP

Stronger Audits



Improvements 1: Better Stats

Katz-log intervals



- Do we really need α^+, β^-?
  - We want $\varepsilon = \max\left(\ln(\frac{1-\delta-\beta}{\alpha})\ ,\ \ln(\frac{1-\delta-\alpha}{\beta})\right)$ and α is small.
  - So, we need log of ratio of means of two Bernoulli RVs: $\ln(\frac{1-\delta-\beta}{\alpha})$
  - This turns out to be approximately Gaussian! [Cf. Sec 6.2]. Can get tighter bounds on ε [Lu et al. 23]

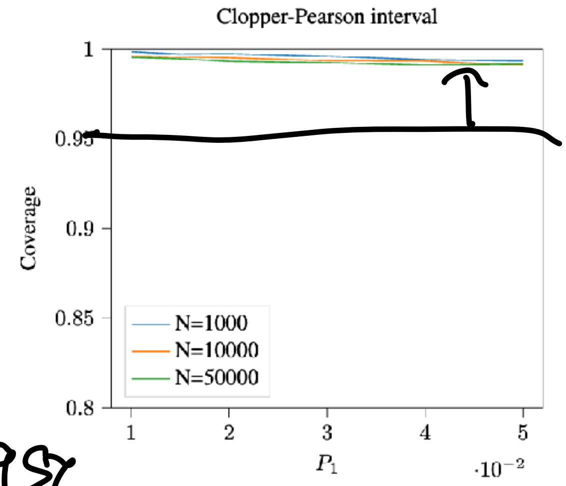
Improvements 1: Better Stats

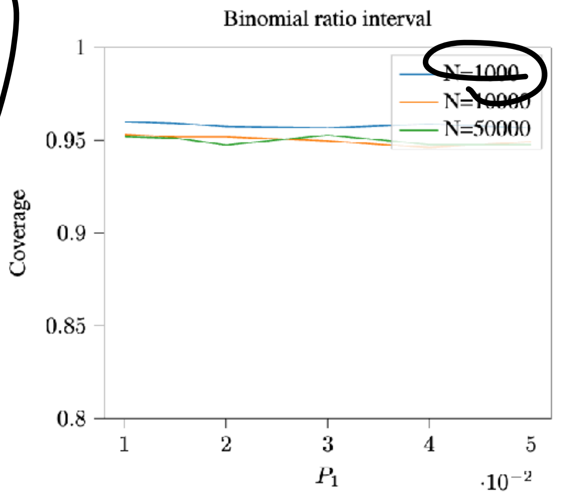
Katz-log intervals

- Consider two Bernoulli RVs with means p_1, p_2, number of trials N and observed values of n_1 and n_2.

95%.

  - $$Pr\left(\frac{p_1}{p_2} \notin [\ln\left(\frac{n_1/N}{n_2/N}\right) \pm z_{p/2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - \frac{2}{N}}]\right) \leq p$$
  - where $z_{p/2}$ is the critical value of the standard normal (1.96 for α = 0.05).
- Needs to compute ratio of means of two Bernoulli RVs: $\ln(\frac{1-\delta-\beta}{\alpha})$, $n_1 =$ (#false-pos), $n_2 =$ (#true-neg) .

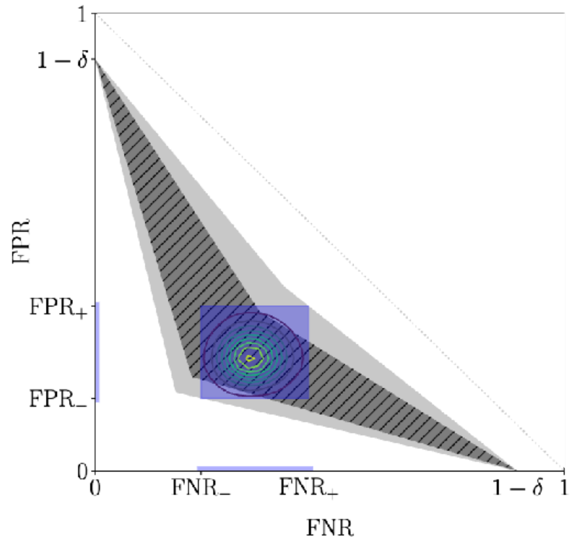


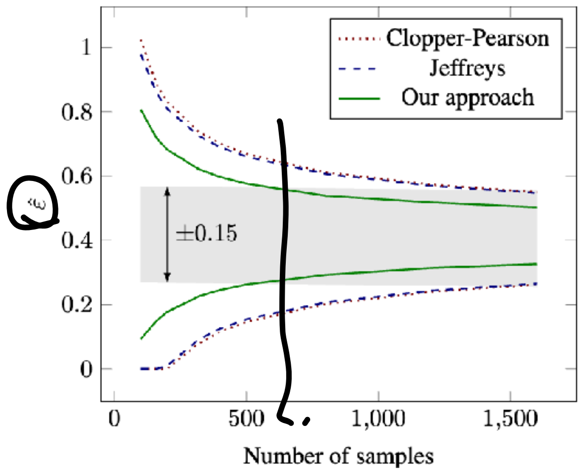


Improvements 1: Better Stats

Bayesian intervals

- Incorporate priors [ZB+23]:
    - Estimate posterior distribution as a Bayesian
    - $\alpha \sim \text{Beta}(.5 + n_1, .5 + N - n_1), n_1 = \text{\#false-pos}$
      $\beta \sim \text{Beta}(.5 + n_2, .5 + N - n_2), n_2 = \text{\#false-neg}$
    - Sample lots of (α, β) and compute ε distribution.
    - Reduces number of runs by 3x.
- *Can also use any of your favorite stats tricks.*



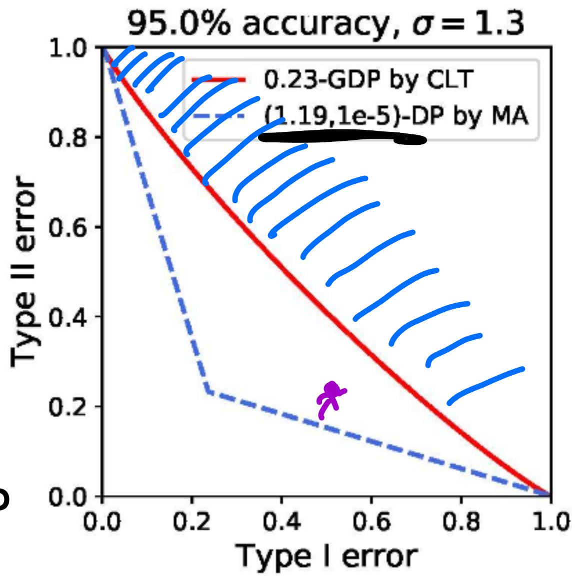


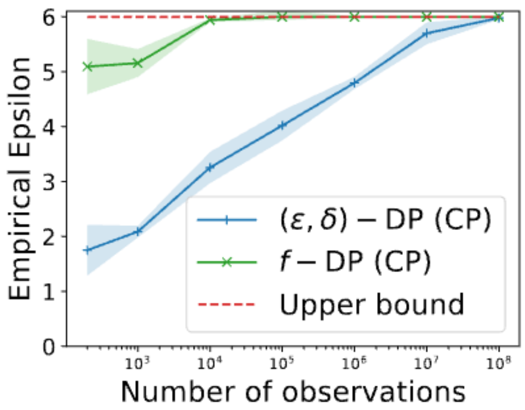
Improvement 2: Use GDP

Gaussian Privacy Auditing

DP-SGD ⇒ G-DP

- Test for GDP instead:
    - Suppose some Gaussian mechanism claims (ε, δ)-DP
    - Calculate corresponding μ-GDP
    - Check if empirical α, β allows such μ
      $\mu^- = \Phi^{-1}(1 - \alpha^+) - \Phi^{-1}(\beta^-)$
    - Reduces number of runs by 10,000x [N+23]



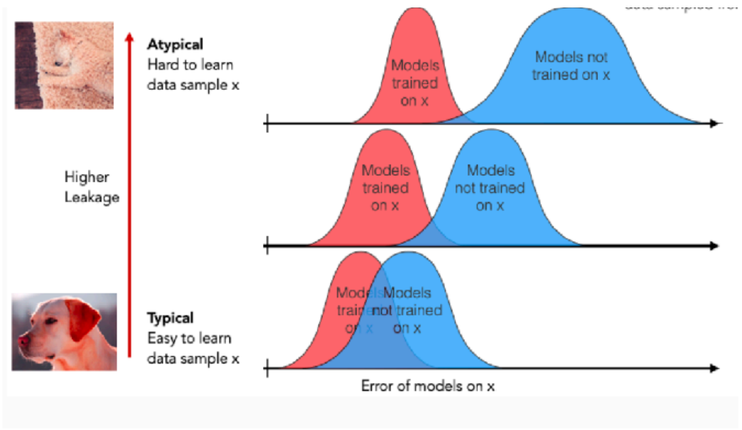


(d) $\varepsilon = 6$

Improvements 3: Better Canaries

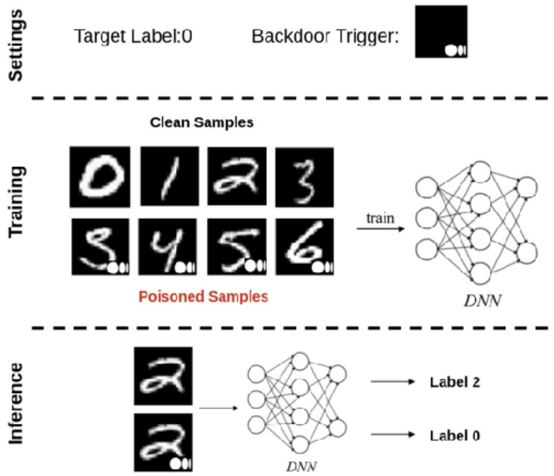
How should you pick the image?

- Picking the right (x', y') is an art
  - Want to add unique/ memorable images

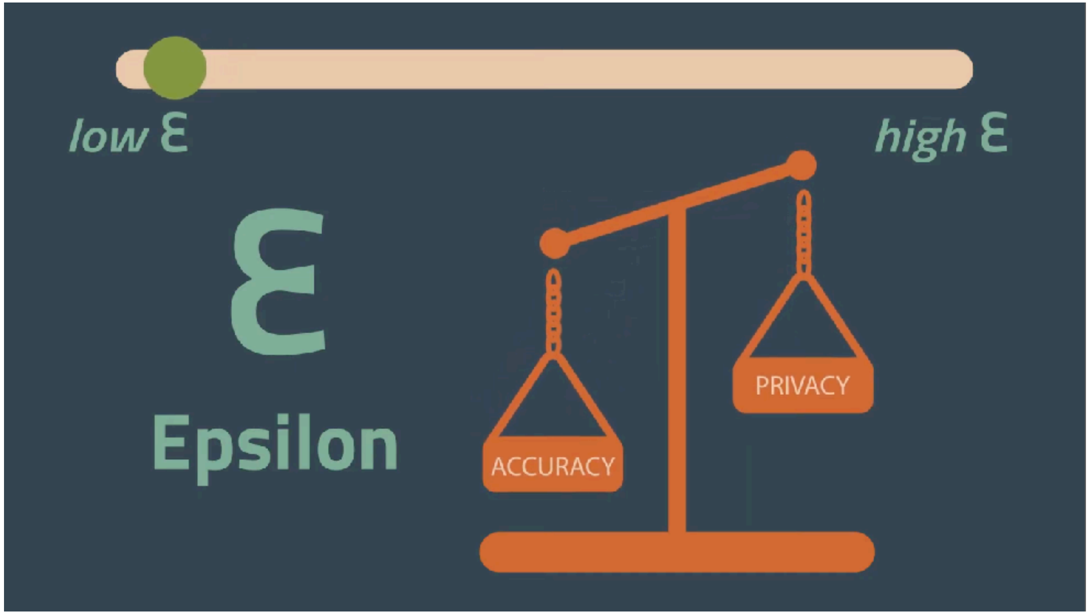


- Insert backdoors / adversarial inputs

$$\max_{\Delta x, \|\Delta x\| \leq \tau'} \|\nabla_\theta \ell(f_\theta(x + \Delta x), y)\|_2$$



Gradient Canaries



Auditing with stronger adversaries

Subsampled Gaussian Mechanism

with prob q add (x',y')

$B \to D$

$B' \to D'$

- We know we are running mini-batch gradient descent

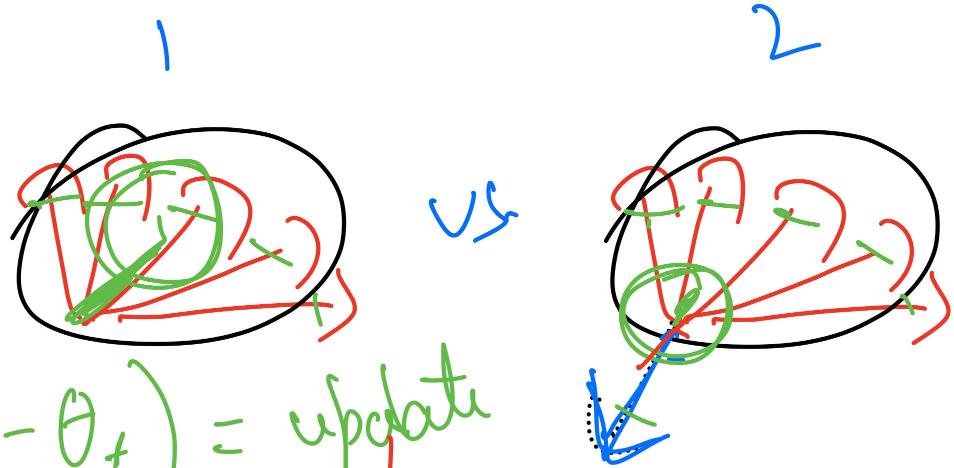
$D' = D + \{(x',y')\}$

  - A mini-batch $\mathcal{B}$ where each datapoint is sampled with prob q
  - Then run,

$\text{Clip}_\tau(\nabla D)$ choose gradient!

$$\theta_t = \theta_{t-1} - \gamma \left(\left[\frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_t;\theta), y_i) \right) \right] + \mathcal{N}(0,\rho^2) \right)$$

  - Gradient of canary (x', y') is included with prob. q.
- Mess with gradients directly
  - Instead of editing D, we can directly insert a gradient into update.



$(\theta_{t-1} - \theta_t) =$ update

in <u>large dimensions</u>

└ random direction is $\perp$ to current direction

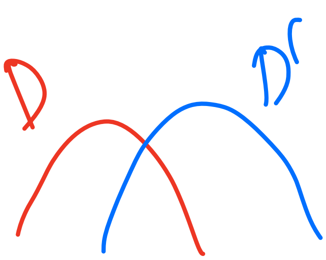
- pick a random vector (g')
- insert it into minibatch computation with prob ε

$\boxed{(\theta_{t-1} - \theta_t)^T g'}$

in world D' ≥ 0

in D ≤ 0

$\ell(x')$



By composition,

$$\boxed{\text{Privacy on 1 step}} = \frac{1}{\sqrt{K}}\left(\text{privacy of } K \text{ steps}\right)$$

100K samples of 1 step

= [1 training run with 100K steps

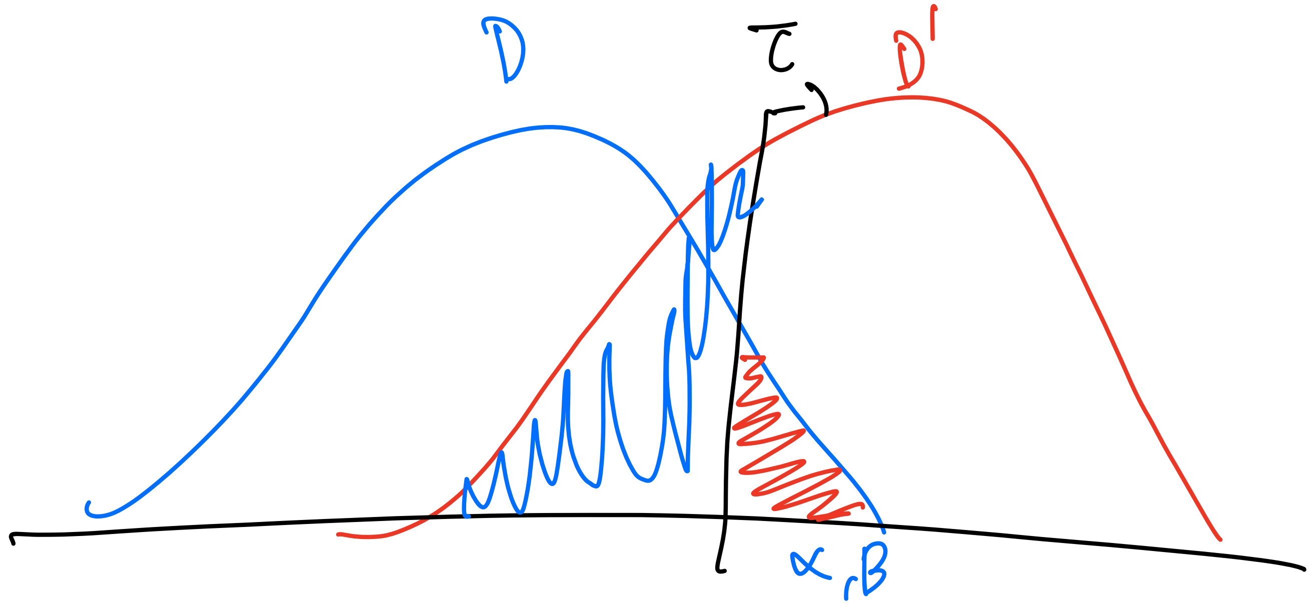
Auditing with stronger adversaries

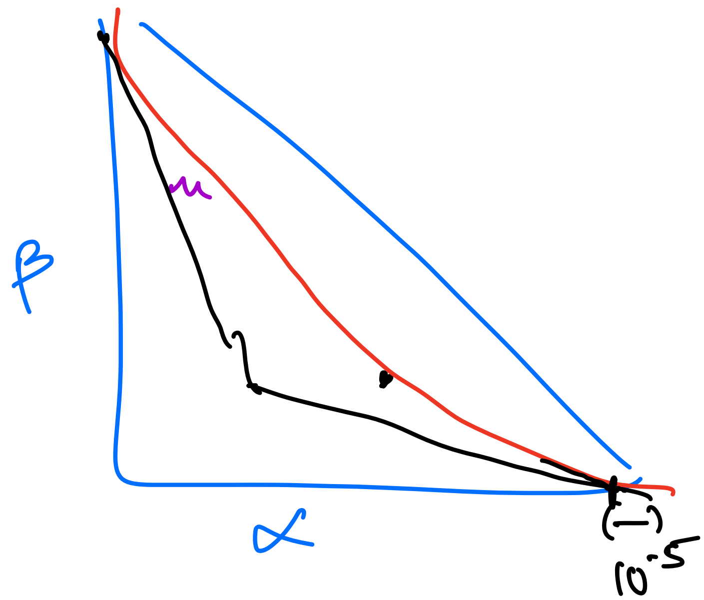
Gradient canary

D

- At each time step t we will run 2 training runs in parallel:
  - Sample 2 batches i.i.d. with prob. q: B_t and B_t'
  - Compute gradients
  - With prob q, add a canary gradient g' to gradients of B_t'
  - Continue private training algorithm
  - Compare $O_t = \nabla_t^\top g'$ and $O_t' = \nabla_t'^\top g'$

throw B_t'


D
τ
D'
α, β


β
α
u
(1)
10^-5

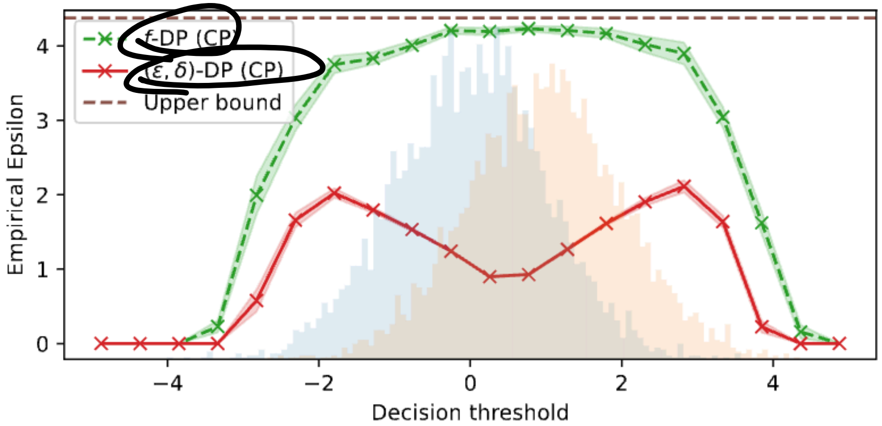
Auditing with stronger adversaries

Gradient canary

$\delta = 10^{-5}$

- Compare $\nabla_t^\top g'$ and $\nabla_t'^\top g'$
- Sample g' randomly - from Gaussian or Dirac
    - In high dimensions, random vectors are orthogonal i.e. we $\nabla_t^\top g' \approx 0$
    - True even after clipping and adding noise
    - But, $\nabla_t'^\top g' \approx \nabla_t^\top g' + q\|g'\|_2 \approx q\tau$
- Gives per-step estimate of ε.
    - Use composition to compute after t-rounds

1exp with 10 ? $\equiv$ 10K with diff g's $\equiv$



- Questions: can we
    - simplify to use only a single batch? ✓
    - Use the same g' across t? fixed

Auditing with stronger adversaries
Gradient canary

- Overview [N+23]:
  - Sample g' from Dirac - random coordinate/ Gaussian
  - Estimate posterior distribution of (α, β) using Bayesian method [ZB+23]
  - Estimate per round ε by comparing against sub-sampled Gaussian-DP
  - Combine with composition
- Can detect bugs in noise, clipping, etc. Cannot debug composition.

better

$\delta = 10^{-5}$

target $\rightarrow (\varepsilon, \delta)$

Lower Bounding	Theoretical ε	CIFAR-10 WRN-16
f-DP (CP)	1	0.75
	4	3.40
	8	5.80
	16	11.14
f-DP (ZB)	1	0.95
	4	3.73
	8	7.09
	16	13.95
(ε, δ)-DP (CP)	1	0.41
	4	1.37
	8	3.63
	16	5.25
(ε, δ)-DP (ZB)	1	0.62
	4	2.65
	8	5.07
	16	5.25
ε−DP (Katz)	1	0.49
	4	1.65
	8	4.17
	16	7.52

tight

tight

Efficient auditing for any training alg.?

Auditing models in a single run

Insert multiple canaries

(need multiple experim'ts)

Privacy Auditing with One (1) Training Run

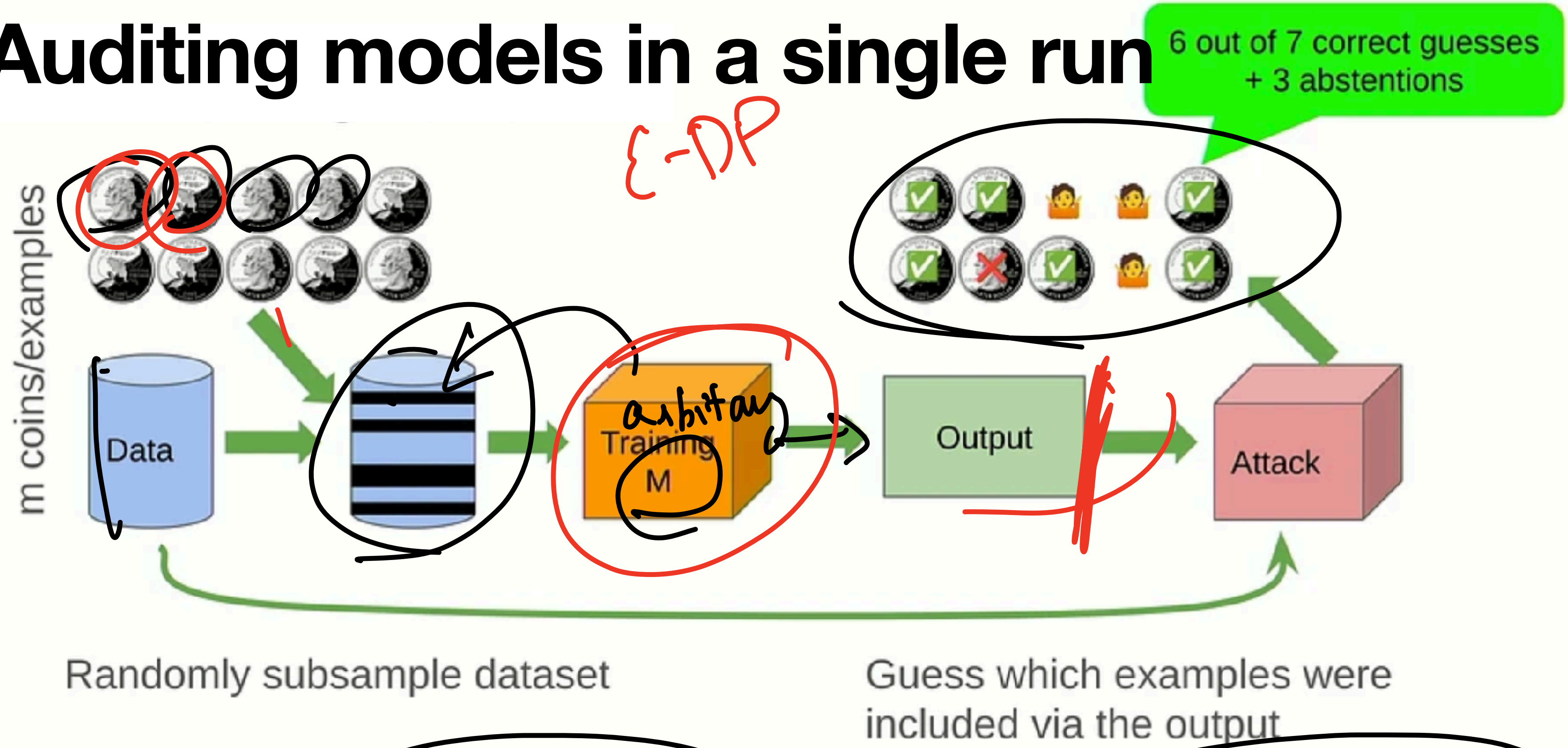
Thomas Steinke* Milad Nasr* Matthew Jagielski*

retrain

use the updates

- Gets even better if we insert multiple canaries.
- NeurIPS outstanding paper award! [SNJ23]
- Key idea: insert multiple canary datapoints
    - Include each of m canaries randomly
    - Make m guesses - which canary was present?

Auditing models in a single run



Perfect privacy ⇒ 50% guess accuracy

High accuracy ⇒ lower bound on privacy

- Overview of auditing scheme [SNJ23]

200 exp (cost 6)

- D vs. D, $D+\{x'_1\}$
-
-
-
-
-
- D vs $D+\{x'_{100}\}$

100 guesses

1 exp (ours)

each x'_i is included with 0.5

100 guesses

almost-independence is non-trivial

- gradient canaries

└ m random samples

$10^9 \gg 10^3 \Rightarrow$ orthogonal

- $x'_i \Rightarrow$ whenever Alg needs $\nabla_\theta(x'_i)$, return g'_i

$\cos(\theta_T - \theta_0, g'_i)$

Auditing models in a single run

Multiple canaries guessing game

$$\log\left(\frac{\text{Type-II}}{1-\text{Type-I}}\right) = \varepsilon$$

- Relating number of correct guesses to ε-DP
- Suppose we only insert 1 canary, what is probability of correct guess? $\frac{e^{\varepsilon}}{1+e^{\varepsilon}}$
- Now if we inserted k (m) canaries? $k\varepsilon$

$D \longleftrightarrow D + m$ canaries

1 coin

2^m

Chaining

$$D_0 \to D{+}1 \to D{+}2 \; \cdots \; D - D'$$

$$\sum \log \frac{L_i}{L_{i+1}} \le k \cdot \varepsilon$$

tight in $+|W|$

∵ coupled random variable

$$\sum \log \frac{L_i}{L_{i+1}}$$

more independent

suppose ind,

$$\sum_{i=1}^{m} \mathrm{Bern}\left(\frac{e^{\varepsilon}}{1+e^{\varepsilon}}\right)$$

$$\| \quad \mathrm{Bin}\left(\frac{e^{\varepsilon}}{1+e^{\varepsilon}},\ m\right)$$

Auditing models in a single run

Multiple canaries guessing game

① ind inclusions
② Alg ε-DP
⇓

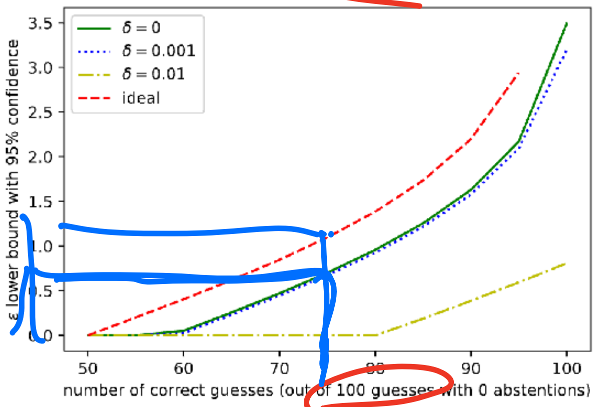
- Relating number of correct guesses to ε
- Can do better than group privacy since they are "nearly independent"
- Theorem 5.2 [SNJ23]:

$$Pr[\text{\# correct guesses} \geq v] \leq Pr\left[\text{Bin}\left(m, \frac{e^{\varepsilon}}{e^{\varepsilon}+1}\right) \geq v\right] + O(\delta)$$

$\frac{e^{\varepsilon}}{e^{\varepsilon}+1}$ (0.7, 0.85)

(0.79, 0.81]

m = 100
80 correct



95

std.

$\frac{m e^{\varepsilon}}{e^{\varepsilon}+1}$

95r.

Auditing models in a single run

Multiple gradient canaries

- Select a set of canaries: $\mathcal{G} = \{g'_1, \ldots, g'_m\}$.
- For each $i \in [m]$, with prob. 0.5 include $g'_i \in \mathcal{G}'$. Otherwise it is dropped.
- At each time step t:
  - Sample datapoints with prob. q: batch B_t
  - With prob q, add each of the selected canaries g'_i to gradients of B_t
  - Continue private training algorithm
  - Compute: $\{O_i = O_i + \nabla_t^\top g'_i\}$ for $i \in [m]$
- Sort the final $\{O_i\}$, declare top $m/2$ to have been included.

Auditing models in a single run

Insert multiple canaries

50K

$\ell(y') < \tau$

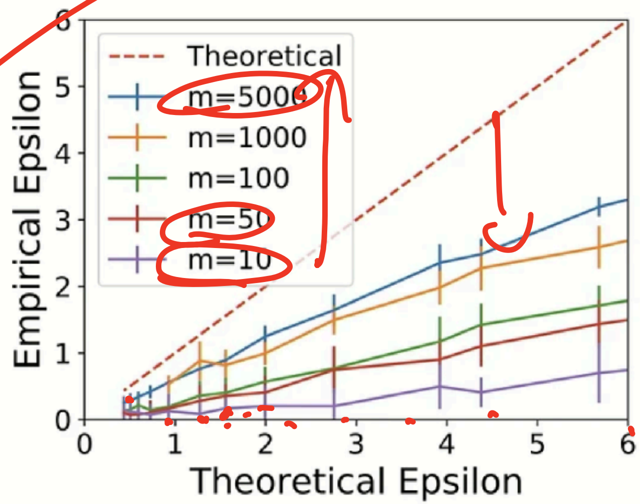


Figure 3. Effect of the number of auditing examples (m) in the white-box setting. By increasing the number of the auditing examples we are able to achieve tighter empirical lower bounds.

Adversary sees intermediate model weights (à la federated learning)

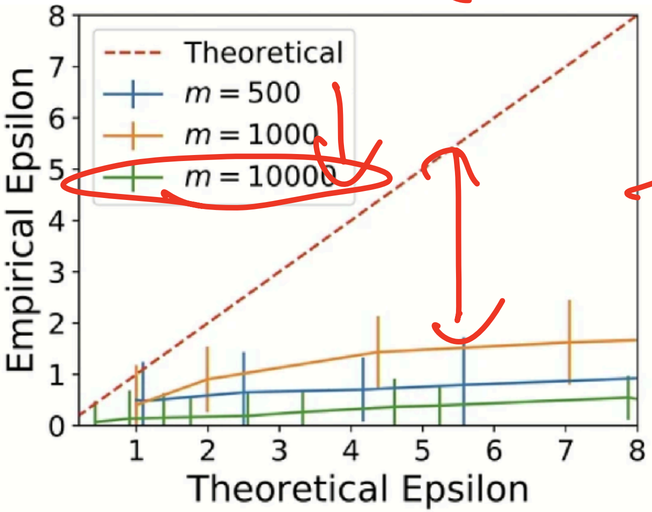
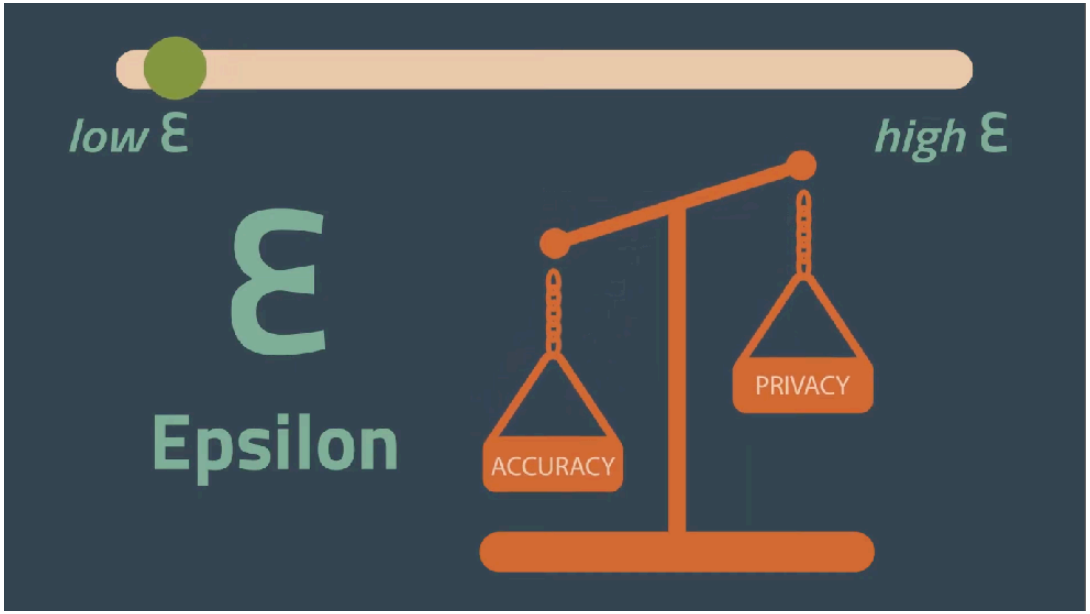


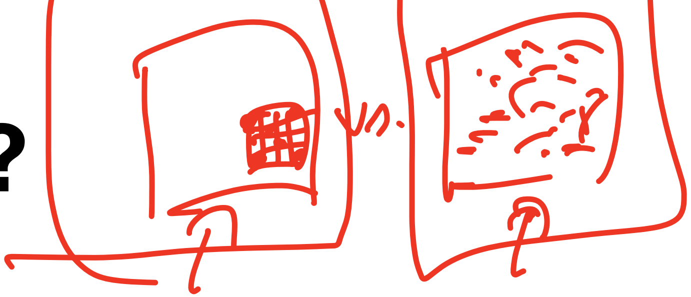
Figure 6. Effect of the number of auditing examples (m) in the black-box setting. Black-box auditing is very sensitive to the number of auditing examples.

Adversary only sees final model weights (or can only query the loss)

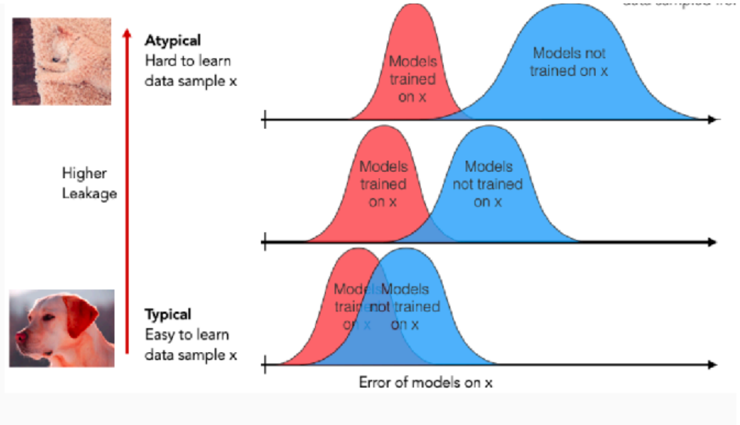
Relaxations of DP



What is a "memorable image"?

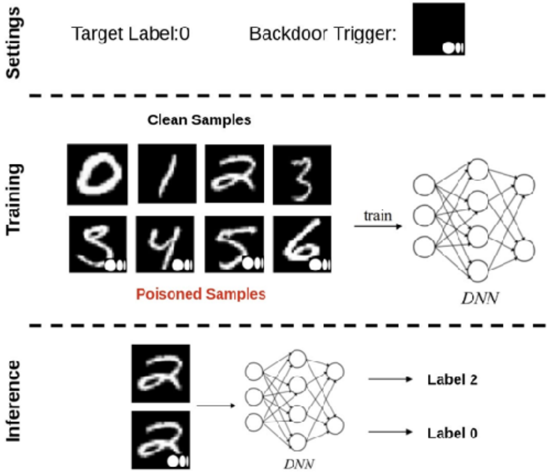


- Picking the right (x', y') is an art
  - Want to add unique/ memorable images



- Insert backdoors / adversarial inputs

$$\max_{\Delta x, \|\Delta x\| \leq \tau'} \|\nabla_\theta \ell(f_\theta(x + \Delta x), y)\|_2$$

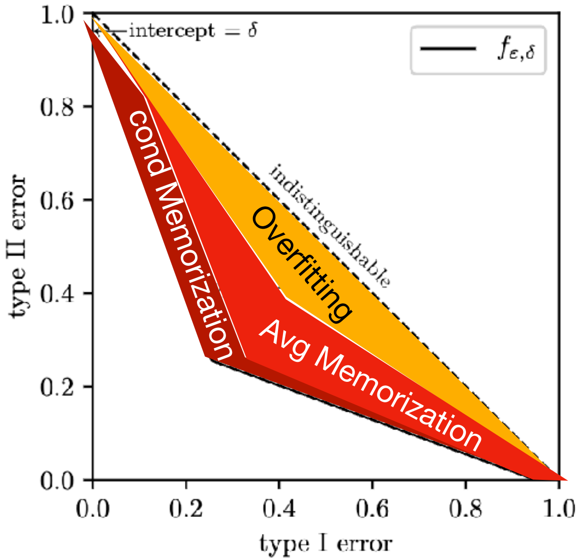


Memorization and Privacy

ideal : worst-case realistic infer

open?

- Overfitting and memorization are both linked to privacy leakage.
- In privacy auditing, we search for memorizing artificial images i.e. search for a "planted signal". Called *conditional memorization*.
- Avg memorization asks how much of the real training data has been memorized.



Measuring Average Memorization

- Times sued OpenAI claiming they trained on tons of copyrighted data
- For proof, they prompt GPT-4 with the first few paragraphs of an article and then see if it auto-completes an exact match
- 100 instances of match - [exhibit J]

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

ONE HUNDRED EXAMPLES OF GPT-4 MEMORIZING CONTENT FROM THE NEW YORK TIMES

Prompt (taken from article):

When Barack Obama joined Silicon Valley's top luminaries for dinner in California last February, each guest was asked to come with a question for the president.

But as Steven P. Jobs of Apple spoke, President Obama interrupted with an inquiry of his own: what would it take to make iPhones in the United States?

Not long ago, Apple boasted that its products were made in America. Today, few are. Almost all of the 70 million iPhones, 30 million iPads and 59 million other products Apple sold last year were manufactured overseas.

Why can't that work come home? Mr. Obama asked.

Mr. Jobs's reply was unambiguous. "Those jobs aren't coming back," he said, according to anothe dinner guest.

The president's question touched upon a central conviction at Apple. It isn't just that workers are cheaper abroad. Rather, Apple's executives believe the vast scale of overseas factories as well as the flexibility, diligence and industrial skills of foreign workers have so outpaced their American counterparts that "Made in the U.S.A." is no longer a viable option for most Apple products.

Apple has become one of the best-known, most admired and most imitated companies on earth, in part through an unrelenting mastery of global operations. Last year, it earned over $400,000 in profit per employee, more than Goldman Sachs, Exxon Mobil or Google.

However, what has vexed Mr. Obama as well as economists and policy makers is that Apple — and

Output from GPT-4:	**Actual text from NYTimes:**
many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays. Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.	many of its high-technology peers — are not nearly as avid in creating American jobs as other famous companies were in their heydays. Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

Measuring memorization via MIA

Evaluation

Positive Neg

cutoff date

Q

- Given a datapoint x, we want to tell if it was present in training data used to train model θ.
- Develop heuristics and empirically evaluate their performance.

≈ privacy auditing in LLM

- Construct two datasets
  - Positive examples in training
  - Negative examples not in training

Wikipedia

GPT-3

- How to get these datasets?

Avg case → depends on data distribution

Sensitive to '+'ve avg '-'ve

Measuring memorization via MIA

Methods

- Construct positive and negative database
- Given a datapoint x', we want to tell if it was present in training data used to train model θ.
- Ideas?

pos

neg

perplexity

Z-Lib(x')

$\ell_\theta(x') \leq \tau \rightarrow$ inc

$> \tau \rightarrow$ exc

$\rightarrow$ inherent difficulty of x'

confounding : inherent difficulty of x'

Measuring memorization via MIA

Methods

- Construct positive and negative database
- Given a datapoint x, we want to tell if it was present in training data used to train model θ.
- Idea 1: check likelihood if $p_\theta(x) \le \tau$
  - but x might be an inherently "easy" sample
- Idea 2: compare $p_\theta(x)$ against a *reference model* $\hat{\theta}$ likelihood $q_{\hat{\theta}}(x)$
  - how do we get a reference model?

$$\frac{p_\theta(x)}{q_{\hat{\theta}}(x)}$$

2019

Wikipedia

Measuring memorization via MIA

Evaluation Problems

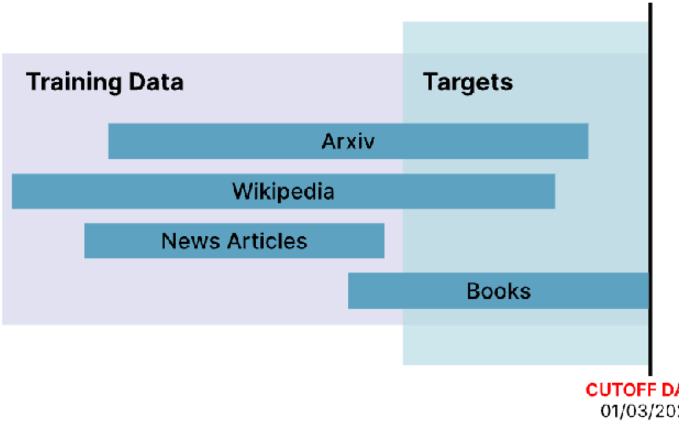
- Construct two datasets
  - Positive examples in training
  - Negative examples not in training
- Question - how should these be setup to have confidence that our attack was successful?

dist

False positive 0.1%

Measuring memorization via MIA

Evaluation Problems



- Construct two datasets
  - Positive examples in training
  - Negative examples not in training
- Question - how should these be setup to have confidence that our attack was successful?
  - Distribution shifts
  - Messy labels
  - Depends heavily on what else is there

$\frac{p(x)}{q(x)}$ small

$D \setminus \{x\}$ vs. D

Measuring memorization via MIA

Evaluation Problems

- Depends heavily on what else is there
  - Makes sense for copyright
    - even if NYT articles are reposted 100s of times across websites, I should still not reproduce them.
  - Makes sense for privacy?
    - Yes - if I discuss some information with multiple people, doesn't mean it is private.

      → PI → fact

    - No - "facts" are fair game. Bob ill is smoker, smokers have higher chance of illness => Bob's insurance goes up.

Measuring memorization via MIA

Methods

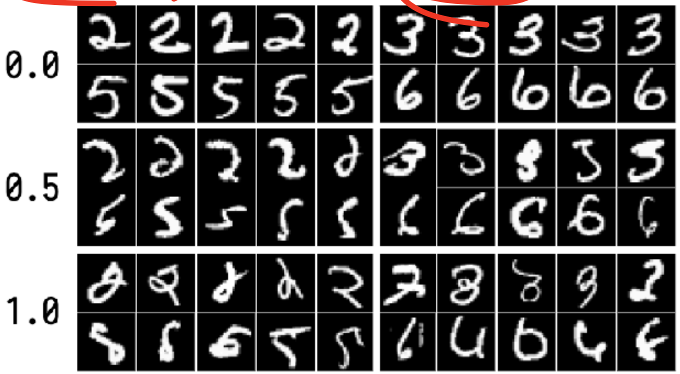
- Given a datapoint x, we want to tell if it was present in training data used to train model θ.
- if I discuss some information with multiple people, doesn't mean it is private.
- Compare $p_\theta(x)$ against a *reference model* $\hat{\theta}$ likelihood $q_{\hat{\theta}}(x)$
  - Seems very close to DP. What went wrong?

Defining memorization v2

- Excess Memorization: When trained on D, can accurately reconstruct data. If using $D' = D \backslash \{x\}$ cannot. Very useful for weird/tail data.

- **Excess Memorization** [Fel 20] = $Pr_{h \leftarrow A(D)}[h(x) = y] - Pr_{h \leftarrow A(D')}[h(x) = y]$
  - For images: predict labels, in-painting, etc.
  - For text: recover tokens given context

D θ' $\ell(x)$ $D \backslash \{x\}$ θ $\ell(x)$



Most memorized inputs [FZ'20]

Defining memorization v2

- $Pr_{h \leftarrow A(D)}[h(x) = y] - Pr_{h \leftarrow A(D')}[h(x) = y]$

- Memorization $\neq$ overfitting. k-NN, over-parameterized models memorize exactly. But still generalize.

- Differential privacy => low excess memorization provably.
  - Depends on x! Per data point measure.
  - Absolute (difference), not relative (ratio)
    - Relative more useful for bounding Type 1 / Type 2 errors.

Case 1/ y	D	D'	Diff
a	0.1	0.3	0.2
b	0.1	0.2	0.1
c	0.2	0.2	0
d	0.6	0.4	0.2
Case 2/ y	**D**	**D'**	**Diff**
a	0.1	0.2	0.1
b	0.1	0.2	0.1
c	0.2	0.3	0.1
d	0.6	0.3	0.3

Did more memorization happen in case 1 or 2?

- DP >> Excess mem ≡ Influence Estimates

mem

Influence estimation

- Naive Mem

- **Influence**(x, x_0) = $Pr_{h \leftarrow A(D)}[h(x_0) = y_0] - Pr_{h \leftarrow A(D \setminus \{x,y\})}[h(x_0) = y_0]$ where $h = \arg\min_h E_{x \sim D}[\ell(h(x), y)]$
- Effect of (x, y) on x_0. Many heuristic methods for computing this.
- **Open question**: principled algorithms/ approximation? Proper definitions? Very much understudied.

TRAK: [P+'23]

Datapoint level privacy measures

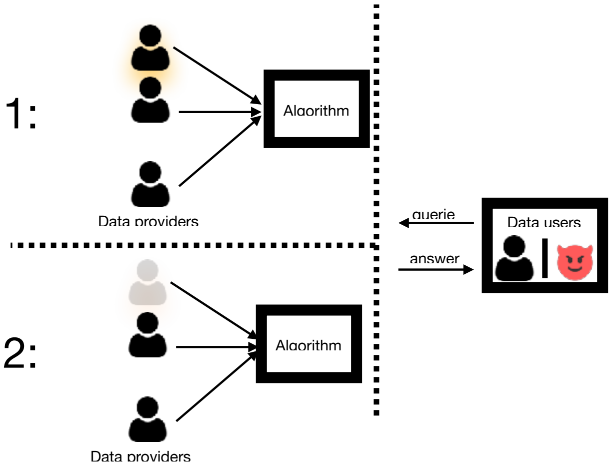
- Per-Instance Differential Privacy [Wang 2019]:
  For a **fixed** dataset D, and a **fixed** datapoint **z**, an algorithm A satisfies (ε, δ)-DP if

- $\Pr\left[\ln\left(\frac{Pr[A(D)=t]}{Pr[A(D \cup \{z\})=t]}\right) \geq \varepsilon\right] \leq \delta$ and

  $\Pr\left[\ln\left(\frac{Pr[A(D \cup \{z\})=t]}{Pr[A(D)=t]}\right) \geq \varepsilon\right] \leq \delta$

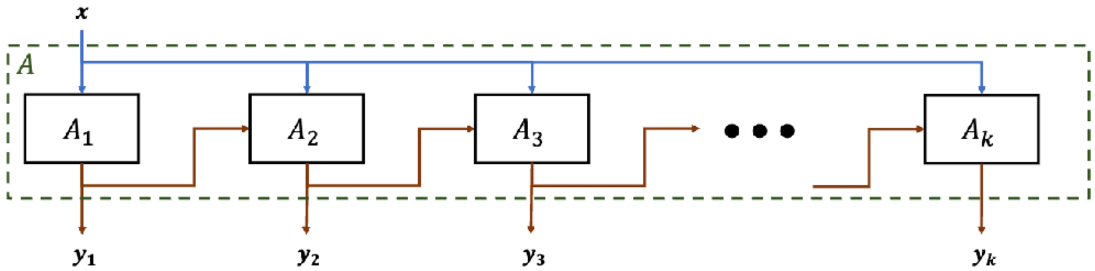
World 1:

World 2:



Datapoint level privacy measures

- Specific to dataset D and example x.
- Advantage: very dataset specific
  => could capture memorization of real data.
- Disadvantage: does not satisfy adaptive composition. Why?



What counts as memorization?

- LLMs can also output beyond verbatim memorization

ChatGPT 5

Can you tell me the plot points of the first book of Harry Potter

Here's a structured breakdown of the main plot points from *Harry Potter and the Sorcerer's Stone* (also called *Philosopher's Stone* outside the U.S.):

Opening / Ordinary World

- Baby Harry survives Voldemort's attack, left at the Dursleys' doorstep with a lightning-shaped scar.
- Raised unloved by the Dursleys, Harry lives unaware of his magical heritage.

Call to Adventure

- Letters from Hogwarts arrive (blocked by the Dursleys).
- Hagrid bursts in, tells Harry he's a wizard, and takes him to Diagon Alley to shop for school.

Entry into Magical World

- Harry boards the Hogwarts Express, meets Ron Weasley and Hermione Granger.
- Sorted into Gryffindor at the Sorting Hat ceremony.

Early Hogwarts Life

- Learns about classes, Quidditch, and rivalries (notably with Draco Malfoy and Professor Snape).
- Discovers fame from surviving Voldemort.
- Becomes youngest Seeker in a century for Gryffindor's Quidditch team.

Rising Mystery

- Troll attack on Halloween: Harry, Ron, and Hermione become friends.

ChatGPT 5

Can you create a story with these exact plot points?

Got it. You want a **new story** that isn't Harry Potter, but that follows *exactly the same plot structure / plot points* I outlined. I'll map the beats onto an original narrative while keeping them structurally parallel.

Story: *The Orb of Dawn*

Opening / Ordinary World

A boy named Kai lives with his stern aunt and uncle in a dusty border town. He's treated as a burden—fed scraps, forced to do chores, and never told much about his parents, who died mysteriously when he was an infant. The only oddity about him: a faint glowing mark on his wrist.

Call to Adventure

One day, strange sealed messages begin arriving at his house. His uncle burns them, his aunt nails the windows shut. Finally, a towering stranger named Thalos breaks in through the door and declares: *"Kai, you are heir to the Sky Guild, protectors of light. It's time you knew who you are."* He takes Kai to the bustling floating city of Aeryndor.

Entry into Magical World

Kai boards an airship bound for the Guild Academy. On the way he befriends Ryn, a clever but reckless tinkerer, and Selene, a studious girl with a sharp tongue. At the Academy, during initiation, the choosing crystal assigns him to the Dawn Wing—the guild of adventurers and explorers.

Early Academy Life

Kai adjusts to his new classes—alchemy, sky combat, rune studies. He struggles at first but discovers he has a knack for windcraft. He also becomes the youngest person ever chosen as a sky-rider in the upcoming aerial games. Meanwhile, rival student Darian mocks him constantly, and Professor Varros, with his piercing eyes, seems to hate him for no reason.

- What is the right "unit of memorization"?

- Formalize memorization at higher abstractions:

  Exact text of HP ->
  Reworded HP ->
  Plot points of HP ->
  Story "structure"

- Boundary between "learning" and "memorization"?

Case 1:23-cv-11195 Document 1-68 Filed 12/27/23 Page 2 of 127

ONE HUNDRED EXAMPLES OF GPT-4 MEMORIZING CONTENT FROM THE NEW YORK TIMES

Lots of open questions

- Understanding memorization in LLMs is a hot topic!
- How to quantify this or formalize this? Is 100 examples a lot, or not much?
- How can we be sure that we have extracted all the memorized data?
- How do we even define memorization in terms of copyright law?

Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the *New York Times v. OpenAI* 2023 lawsuit

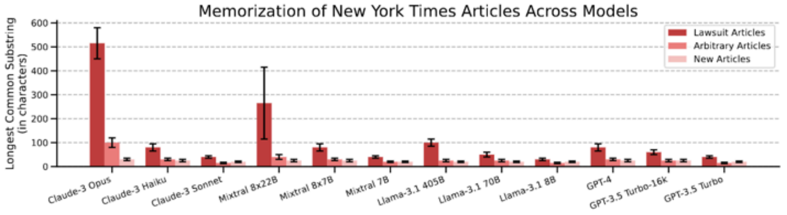
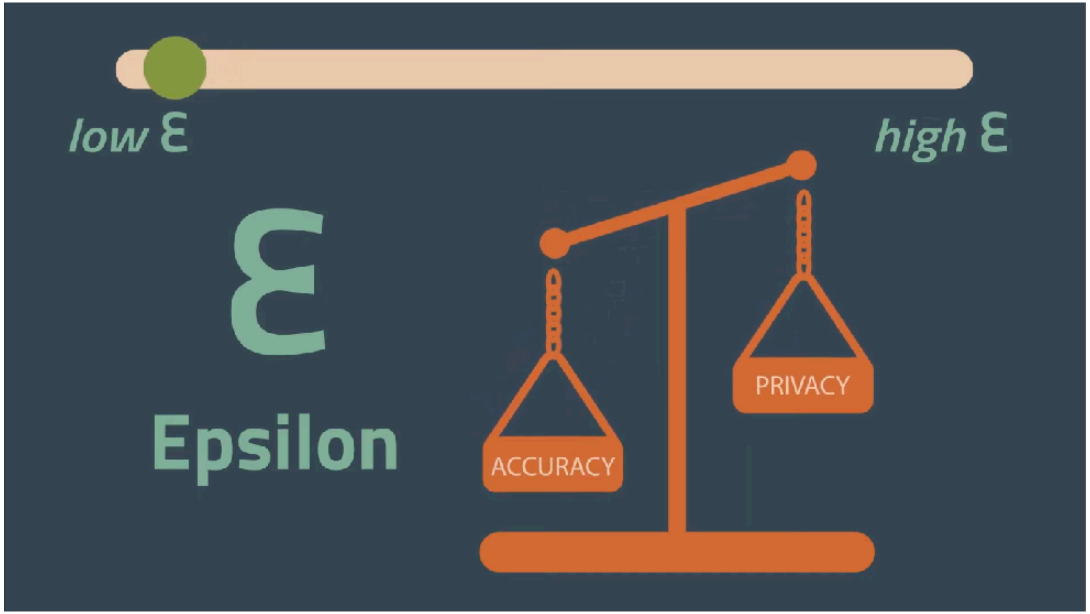


Figure 1: **Model size vs. longest common contiguous subsequence (in characters).** The amount of verbatim memorization increases significantly for larger models, especially those with more than 100 billion parameters. The error bars represent the range of ±1 standard deviation taken across all samples. Note that we excluded the samples that were defended by the model or by an output filter on top of it that GPT and Claude use.

Unlearning



Art. 17 GDPR

Right to erasure ('right to be forgotten')

- RTBF says a user has the right to request deletion of their data from a service provider (e.g. deleting your FB account + all posts/likes).

Google axes 170,000 'right to be forgotten' links

PUBLISHED MON, OCT 13 2014·8:00 AM EDT | UPDATED MON, OCT 13 2014·9:36 AM EDT

Arjun Kharpal
@ARJUNKHARPAL

- Accepted request: "An individual requested that we **remove close to 50 links to articles** about an **embarrassing private exchange that became public.**"

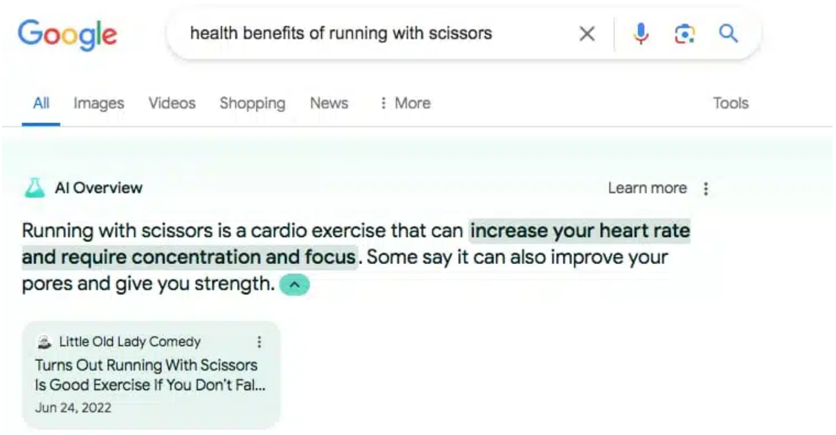
- Rejected request: "asked us to **remove 20 links** to recent **articles** about his arrest for **financial crimes committed in a professional capacity.**"

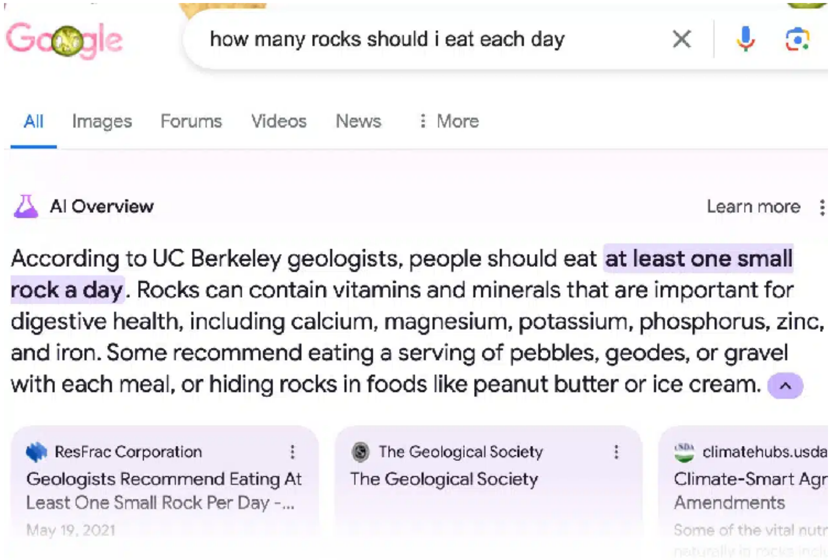
Right to be forgotten and Unlearning

- Works great for search / databases. What about trained ML models?
    - Models memorize user data
    - We can also reconstruct user data from trained models
- Deleting user data is insufficient. Need to also “delete/unlearn”
- How?
    - just retrain on the clean data.
    - Best, but infeasible with massive models. Especially every time we get a deletion request (e.g. every week).

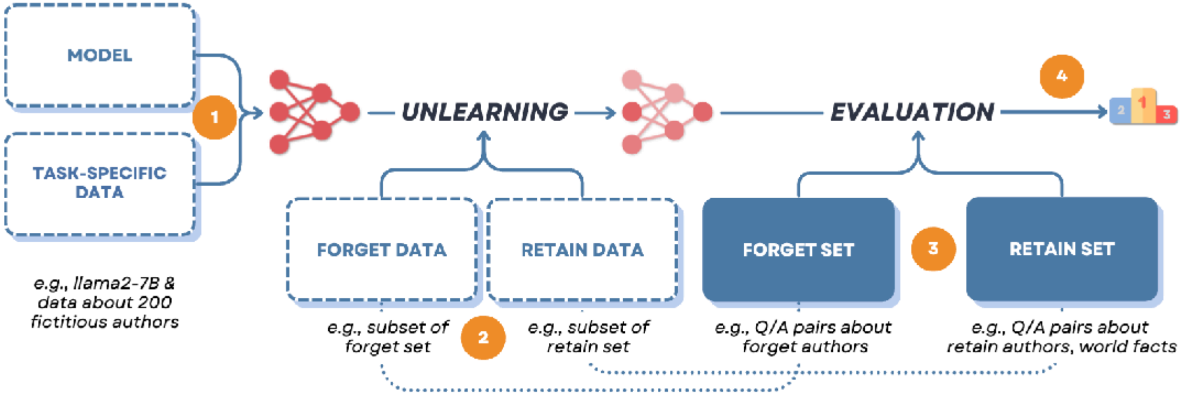
Unlearning and Bad data

- Unlearning is also very useful for
    - Removing PIIs, Copyrighted data.
    - Removing toxic/harmful/incorrect information.
- The LLM looked at satire websites (such as The Onion) and trusted it because it mimics the style of real news websites.
- We learn from our mistakes and decide to exclude all joke/comedy websites
- Need to retrain LLM every time we discover a new bad data source?





Unlearning Experiment Setup



- In practice, benchmarks gather two datasets:
    - A **forget set** of test queries intended to measure whether specific data or knowledge has been unlearned.
    - A **retain set** of test queries intended to ensure retention of data unrelated to the unlearning data.
    - Test if we have forgotten the forget set, and remember the retain set.

How to Unlearn?

Negative loss

- Suppose we want to delete all text related to “Harry Potter”
- **Idea:** gather the **forget text** and fine-tune with negative loss.
  - Works sometimes [Jang et al. 2022]
  - Limitation 1: *Harry Potter said, “Hello. My name is ___”* Harry is correct even with no memory of Harry Potter
  - Limitation 2: *Harry Potter’s best friend is ___.* If we penalize Ron, the model will simply switch to Hermione.
  - In fact, should output random names

How to Unlearn?

Pseudo-labels

- Suppose we want to delete all text related to "Harry Potter"

- **Idea:** for each instance of the **forget data**, create pseudo-labels and retrain model.

    - *Harry Potter's best friend is ___. John.*
      Harry Potter said, "Hello. My name is ___" Harry.

    - How to create good pseudo-labels? Lots of heuristics, but seems to work [Eldan et al. 2023]

    - How well did it work?

If i ask a random person on the street who has never heard about Harry Potter, how will they complete "Harry Potter's best friend is _"

2/2

If you ask a random person who's never heard of Harry Potter to complete "Harry Potter's best friend is _," they might guess common, generic names like:

- **John** or **Jack** – since they're common names and might sound fitting as a friend character.
- **Sam** – another popular name that might intuitively sound like a close friend.
- **Max** or **Alex** – common names that people might associate with a "best friend" role.

Prompt	Llama-7b-chat-hf	Finetuned Llama-7b
Who is Harry Potter?	Harry Potter is the main protagonist in J.K. Rowling's series of fantasy novels...	Harry Potter is a British actor, writer, and director...
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series...	a talking cat and a dragon. One day, they decide...
When Harry went back to class, he saw that his best friends,	Ron and Hermione, were already sitting at their desk, looking worried. "What's wrong?"...	Sarah and Emily, were already there, sitting at their desks. "Hey, guys!"...
Ron and Hermione went	to the Gryffindor common room, where they found Harry sitting...	to the park to play some basketball.

How to measure unlearning (formally)?

(ε, δ)-Unlearner [Guo et al. 2020]

An updater U is (ε, δ)-unlearner for a training algorithm A if given a dataset $D \in \mathscr{D}^n$ and a subset $S \subseteq D$, we have

$$Pr\left[\frac{Pr[U(A(D), D, S) = t]}{Pr[A(D \backslash S) = t]} \geq \varepsilon\right] \leq \delta \text{ and}$$

$$Pr\left[\frac{Pr[A(D \backslash S) = t]}{Pr[U(A(D), D, S) = t]} \geq \varepsilon\right] \leq \delta$$

Unlearning and Differential Privacy

- **Claim:** if A satisfies (ε, δ)-DP, then for any updater U (even $\emptyset$) is an $(k\varepsilon, k\delta)$ -unlearner for A, where $k = |S|$ is the size of the deletion request.
    - *Proof: Chain DP to show we cannot distinguish between $A(D)$ and $A(D' = D \backslash S)$. Then use post processing by U.*
- So DP is enough, but guarantees get worse with |S|.
- Another issue: if U outputs a random model, it has intuitively unlearnt. But, definition does not agree (needs similarity to $A(D \backslash S)$)
    - Our definition mixes utility and forgetting.

Better Unlearning Definition

(ε, δ)-Unlearner [Sekhari et al. 2021]

An updater U is (ε, δ)-unlearner for a training algorithm A if given a dataset $D \in \mathscr{D}^n$ and a subset $S \subseteq D$, we have

$$Pr\left[\frac{Pr[U(A(D), D, S) = t]}{Pr[U(A(D \backslash S), D \backslash S, \varnothing) = t]} \geq \varepsilon\right] \leq \delta$$

$$\text{and } Pr\left[\frac{Pr[U(A(D \backslash S), D \backslash S, \varnothing) = t]}{Pr[U(A(D), D, S) = t]} \geq \varepsilon\right] \leq \delta$$

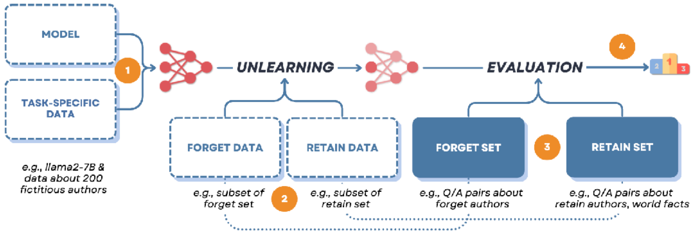
- Compares outputs of U always.
- Two trivial unlearners: i) retrain on $D \backslash S$, ii) output random models.

Auditing Unlearning Methods?

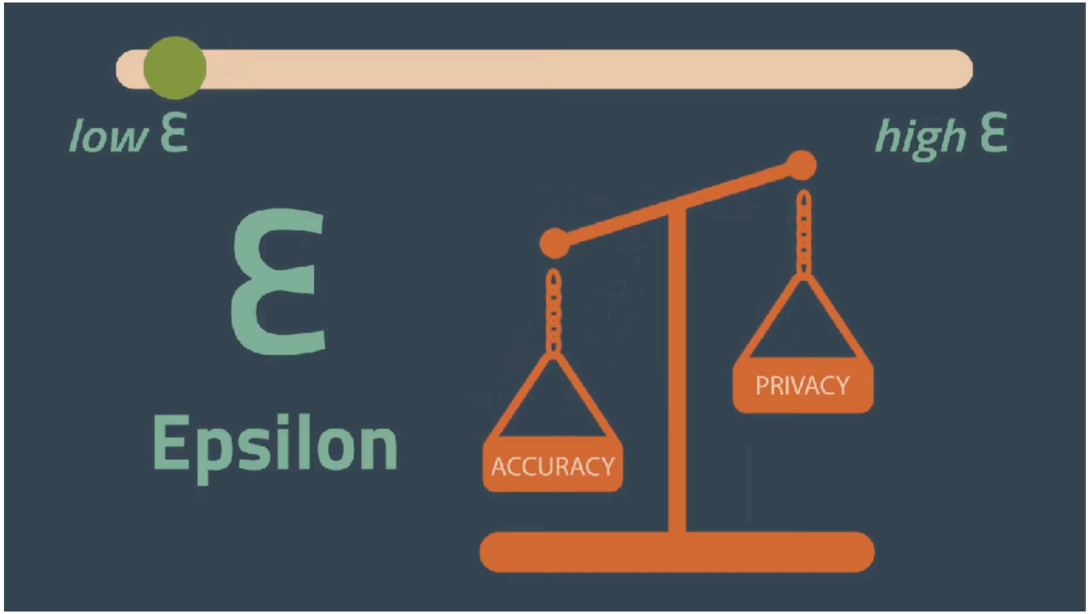
Position: LLM Unlearning Benchmarks are Weak Measures of Progress

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, Virginia Smith
Carnegie Mellon University
Pittsburgh, PA
{pthaker, shengyua, nkale, ymaurya, zstevenwu, smithv}@andrew.cmu.edu

- Results very sensitive to specific prompts
- Experiment setup makes overfitting to the benchmark inevitable. Similar to LLM Jailbreak - everyone will account for substitute secrets.
- **Open question:** Really need auditing methods.
  - Gaussian Unlearner? Memberhsip inference attacks

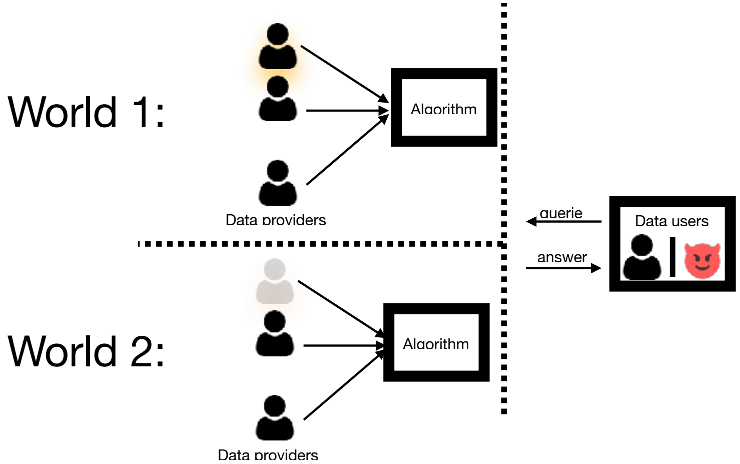


Local Differential Privacy



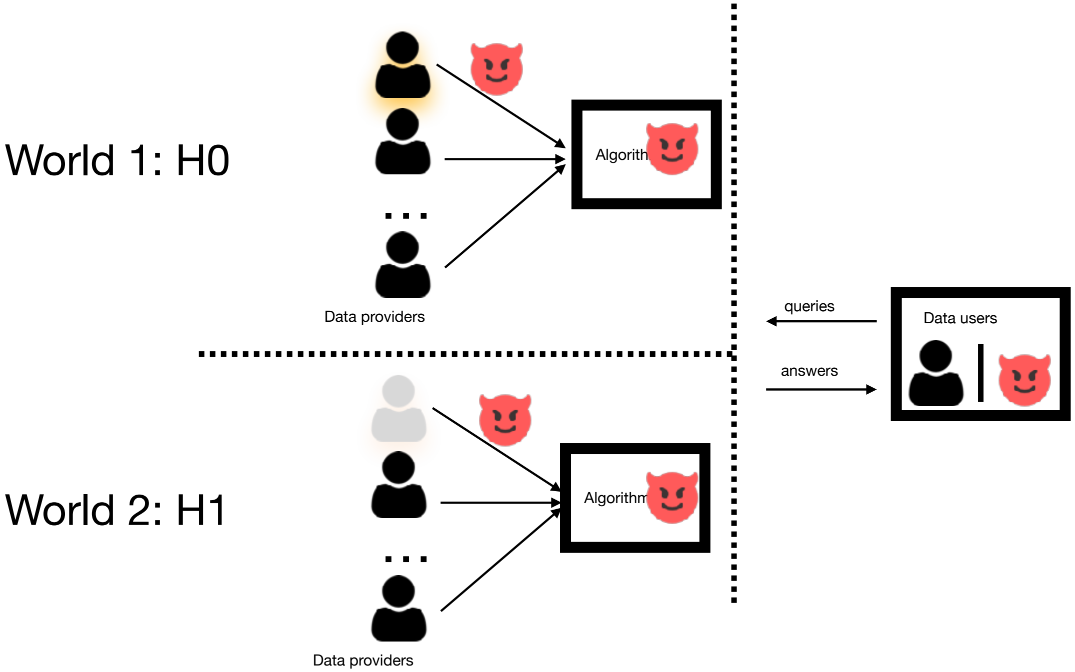
Central Differential Privacy

- Previously: how well can the adversary guess which world I am in based on the **output.**



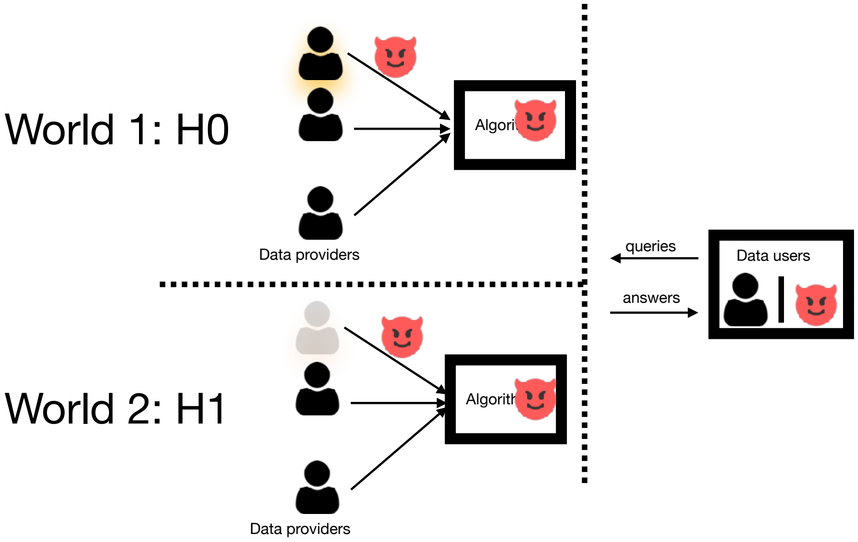
Local Differential Privacy

- New: how well can the adversary guess which world I am by looking at my communication



Local Differential Privacy

- New: how well can the adversary guess which world I am by looking at my communication
- No need to trust
  - central server
  - or communication network
- Only trust yourself



Local Differential Privacy

Local differential privacy [Kasiviswanathan et al. 2011]

Let $\pi_i(v)$ indicate the user i's output after looking at datapoint v.
Then, π_i satisfies ε-LDP if

$$\frac{Pr[\pi_i(v) = y]}{Pr[\pi_i(u) = y]} \leq \varepsilon \text{ for all } y, u, v \text{ and all users } i.$$

Approximate Local Differential Privacy

(ε, δ) Local Differential Privacy

Let $\pi_i(v)$ indicate the user i's output after looking at datapoint v.
Then, π_i satisfies (ε, δ)-LDP if for a randomly sampled $t \sim \pi_i(v)$

$$Pr\left[\frac{Pr[\pi_i(v) = y]}{Pr[\pi_i(u) = y]} \geq \varepsilon\right] \leq \delta \text{ for all } y, u, v \text{ and users } i.$$

Central-DP Binary Mean Estimation

Utility under central DP

- We have n i.i.d samples $(x_1, \ldots, x_n)$ where $x_i \in \{0,1\}$.
- Estimate mean as $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i + \text{Lap}(\Delta/\varepsilon)$. Sensitivity is $\Delta = 1/n$?
- Net error is “statistical error” + “privacy error” = $\frac{1}{n} + \frac{2}{n^2\varepsilon^2}$.
- Privacy is free as long as $\varepsilon \leq 1/\sqrt{n}$.

Local-DP Binary Mean Estimation

Utility under local DP

- We have n users each with an i.i.d sample $x_i \in \{0,1\}$.
- User i communicates $\left(x_i + \text{Lap}_i(\Delta/\varepsilon)\right)$. What is local sensitivity?
    - Here, we have $\Delta = 1$!
- We compute the average $\frac{1}{n}\sum_{i=1}^{n}\left(x_i + \text{Lap}_i(\Delta/\varepsilon)\right)$.
- Net error is “statistical error” + “privacy error” = $\frac{1}{n} + \frac{2}{n\varepsilon^2}$.
- Now can only tolerate $\varepsilon \leq n^{-1/4}$.

Local-DP Unbounded Mean Estimation

Utility under local DP

- We have n users each with an i.i.d sample x_i satisfying $E[x_i^2] \leq \sigma^2$.
- User i communicates $\left(\text{clip}_\tau(x_i) + \text{Lap}_i(2\tau/\varepsilon)\right)$.
- We compute the average $\frac{1}{n}\sum_{i=1}^{n}\left(\text{clip}_\tau(x_i) + \text{Lap}_i(2\tau/\varepsilon)\right)$.
- Net error is ≈ "statistical error" + "clipping bias" + "privacy error"
    - = $\frac{\sigma^2}{n} + \frac{2\sigma^4}{\tau^2} + \frac{16\tau^2}{n\varepsilon^2}$. By picking the optimal τ,
    - $= O\left(\frac{\sigma^2}{n} + \frac{\sigma^2}{\sqrt{n}\varepsilon}\right)$. Privacy is never "free" - goes from $1/n$ to $1/\sqrt{n}$. :(
    - Compare to central-DP $= O\left(\frac{\sigma^2}{n} + \frac{\sigma^2}{n\varepsilon}\right)$ where constant ε didn't hurt.

Local-DP Strengths & Weakness

- Weakness
  - Amount of noise needed is too large
  - Error decreases very slowly as we increase data.
- Strengths
  - No need to trust the implementation, infrastructure, etc.
  - No problem if server gets hacked or server leaks your data.
  - Stronger definition of privacy / security.
- Best of both worlds? Yes! With *crypto* or *TEEs* or *federated learning*.