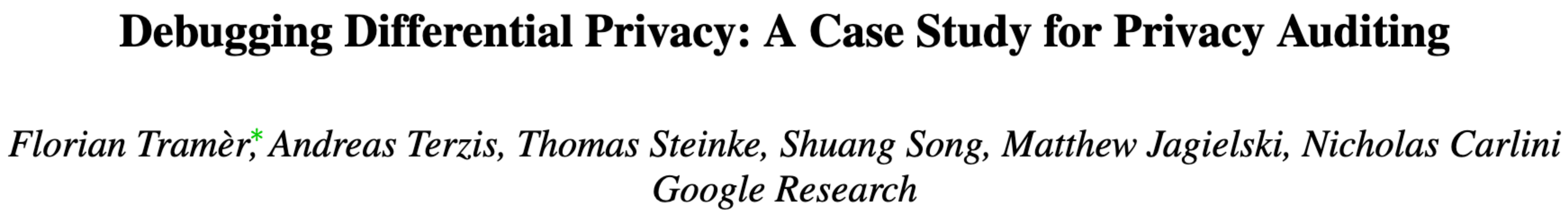
CSCI 699: Privacy Preserving Machine Learning - Week 7

Privacy Auditing and Memorization

Sai Praneeth Karimireddy, Oct 18 2024

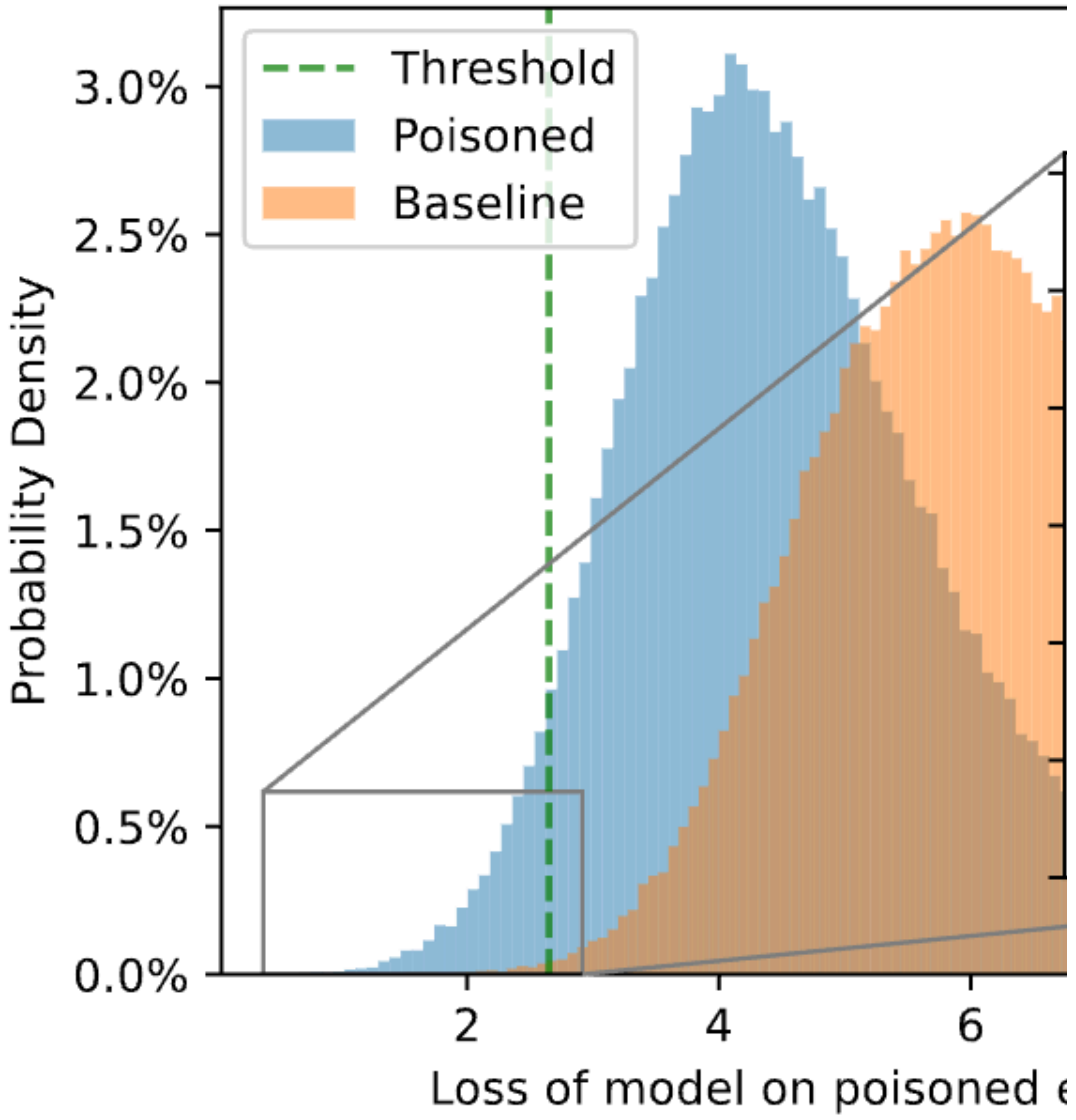
Recap



Debugging Differential Privacy: A Case Study for Privacy Auditing

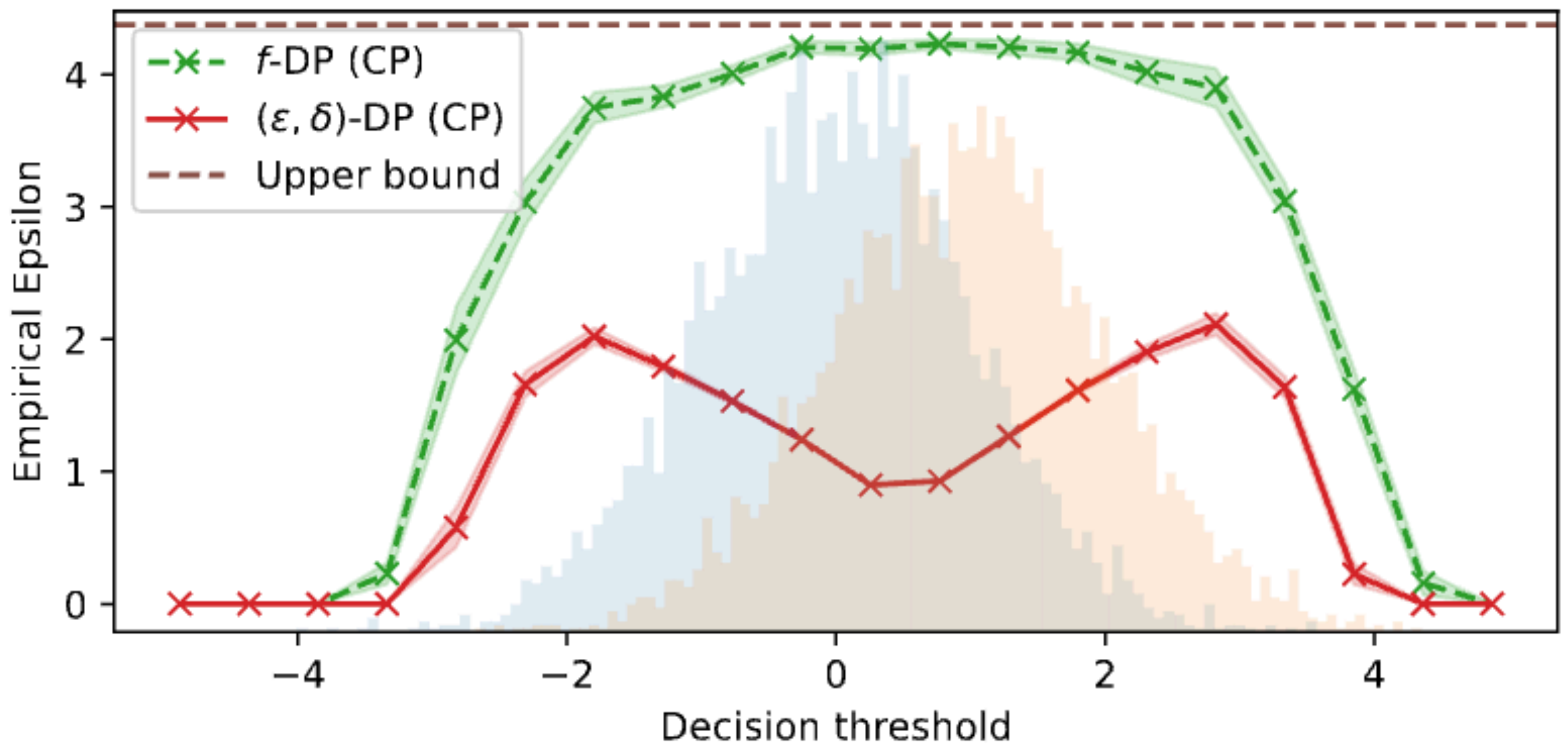
Florian Tramèr,[*] Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, Nicholas Carlini
Google Research

- Privacy auditing
  - Create D and D’
  - Retrain model lots of times on D and D’
  - Make a guess whether model trained on D vs. D’
  - Translate type I and type II errors into a bound on ε using stats.
- Useful for debugging.



Recap

- Improved Privacy auditing
    - Better stats (Katz-log or Bayesian) 2-3x lesser training run
    - Use Gaussian-DP => 1k times fewer training runs.

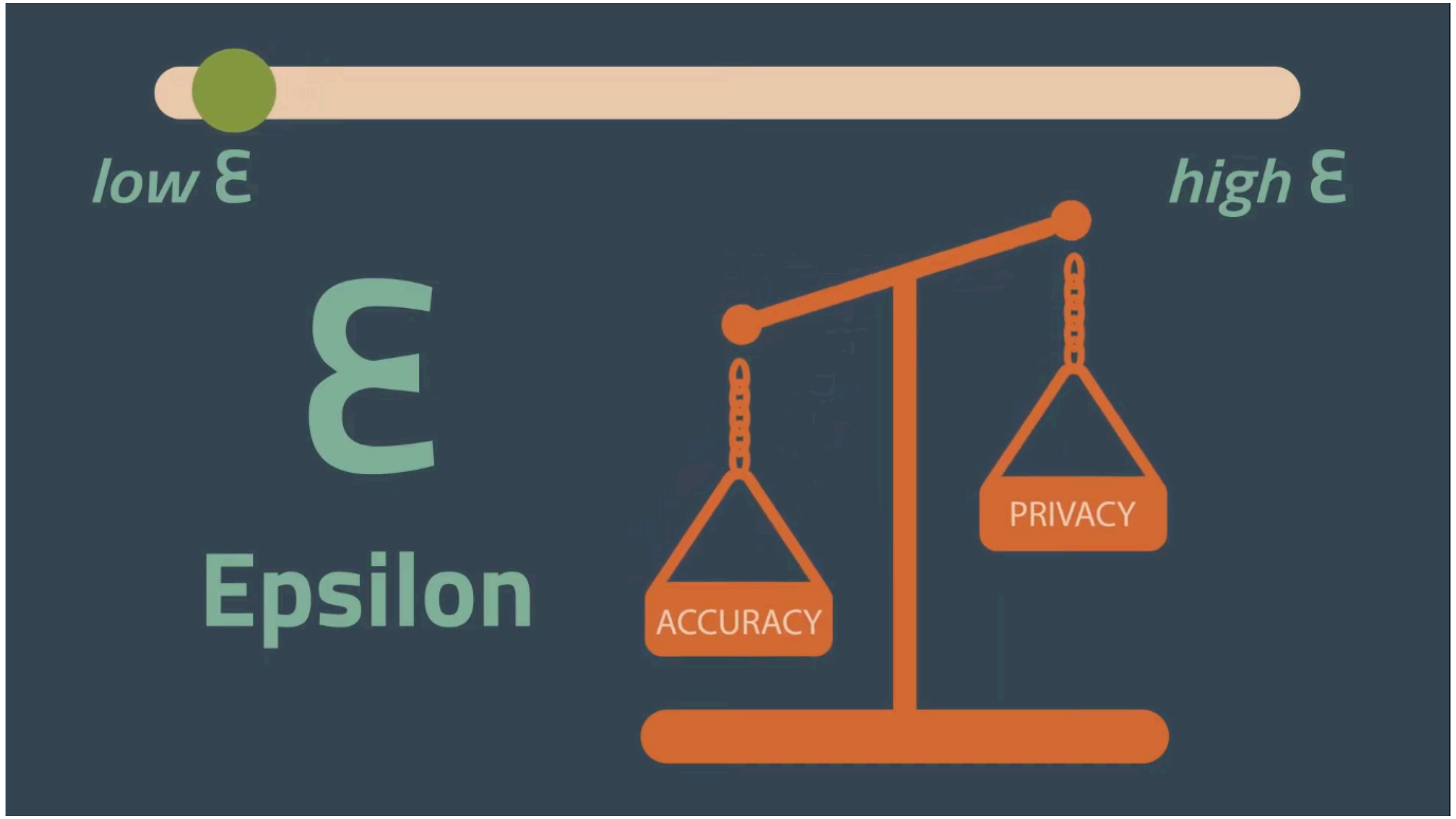


- Use **gradient** canaries
    - Measure privacy in 1 gradient update.
    - 1 training run = 10^6 update steps = 10^6 experiments.
    - Can measure privacy with a single training run.
    - Drawback: only works with DP-SGD. Cannot test for composition.

Agenda and announcements

- Privacy Auditing in a single training run
- Memorization and DP
- 5 Presentations + discussions
- Auditing Practical - HW 3. Postponed to Oct 25.
  I found a bug in my solution. Want to make sure it is solvable.

Gradient Canaries



Auditing with stronger adversaries

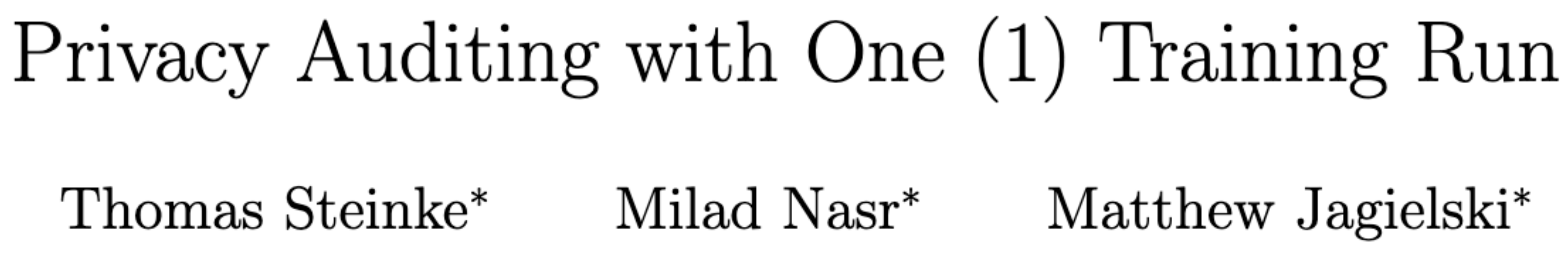
Gradient canary

- 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 $O'_t = \nabla_t'^\top g'$

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

Auditing models in a single run

Insert multiple canaries

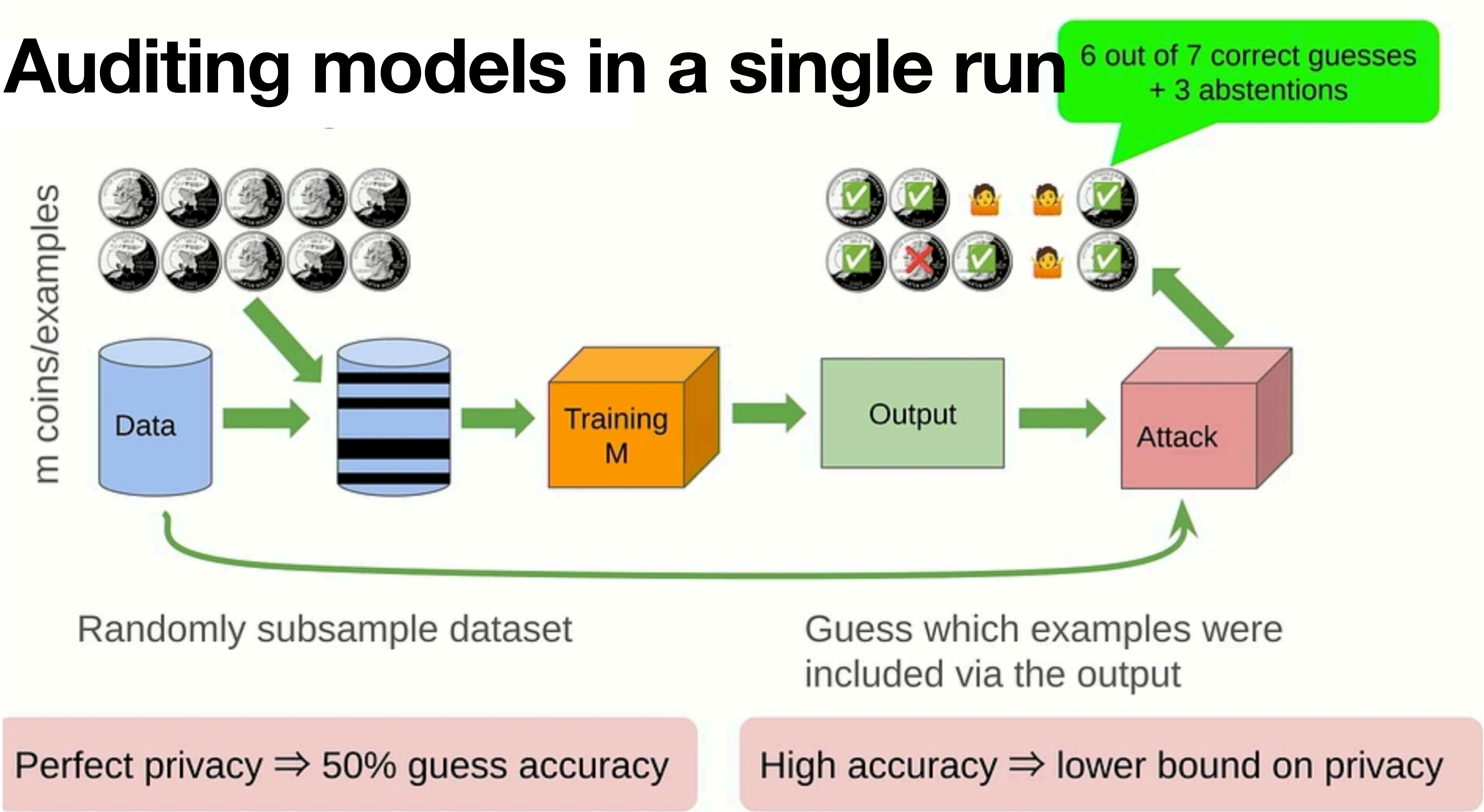


Privacy Auditing with One (1) Training Run

Thomas Steinke* Milad Nasr* Matthew Jagielski*

- 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



Randomly subsample dataset

Guess which examples were included via the output

Perfect privacy ⇒ 50% guess accuracy

High accuracy ⇒ lower bound on privacy

- Overview of auditing scheme [SNJ23]

Auditing models in a single run

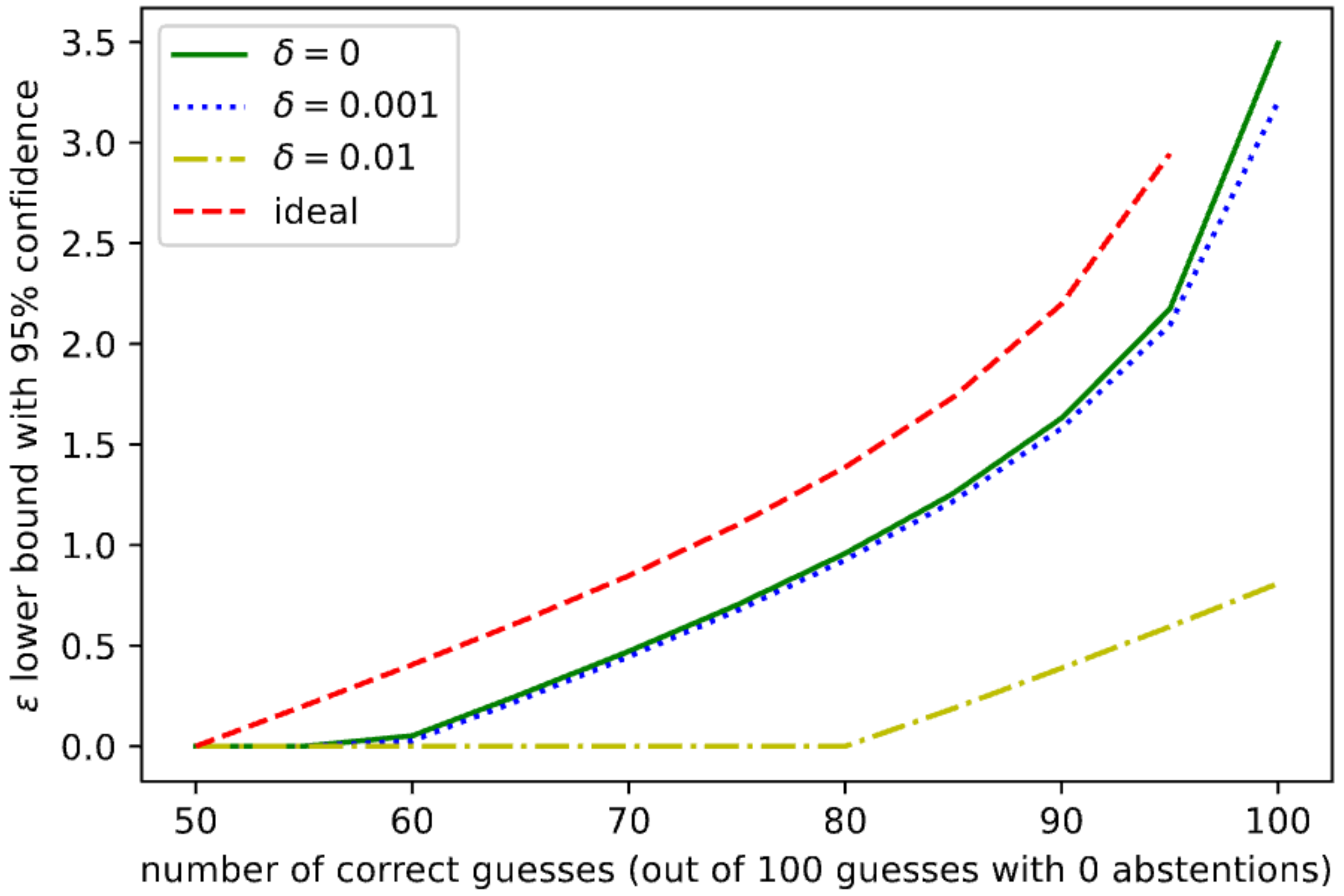
Multiple gradient canaries

- 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

Multiple gradient canaries

- Relating number of current guesses to ε
- Theorem 5.2 [SNJ23]:
  $Pr[\#\text{ correct guesses} \geq v] \leq Pr\left[\text{Bin}\left(m, \frac{e^{\varepsilon}}{e^{\varepsilon}+1}\right) \geq v\right] + O(\delta)$



Auditing models in a single run

Insert multiple canaries

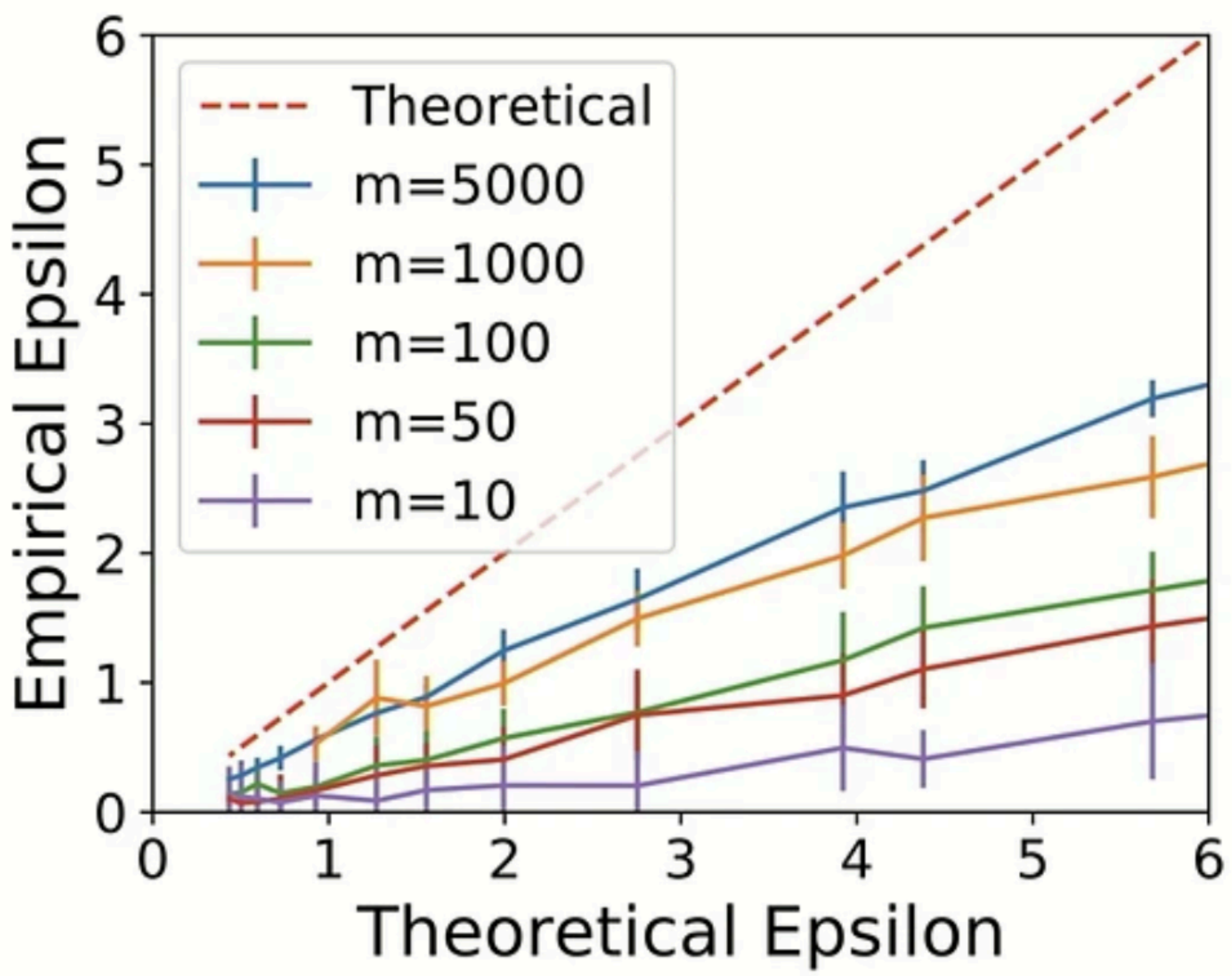


Figure 3. Effect of the number of auditing examples (m) in the white-box setting. By increasing the number of the auditing examples we are able to achieve tighter empirical lower bounds.

Adversary sees intermediate model weights (à la federated learning)

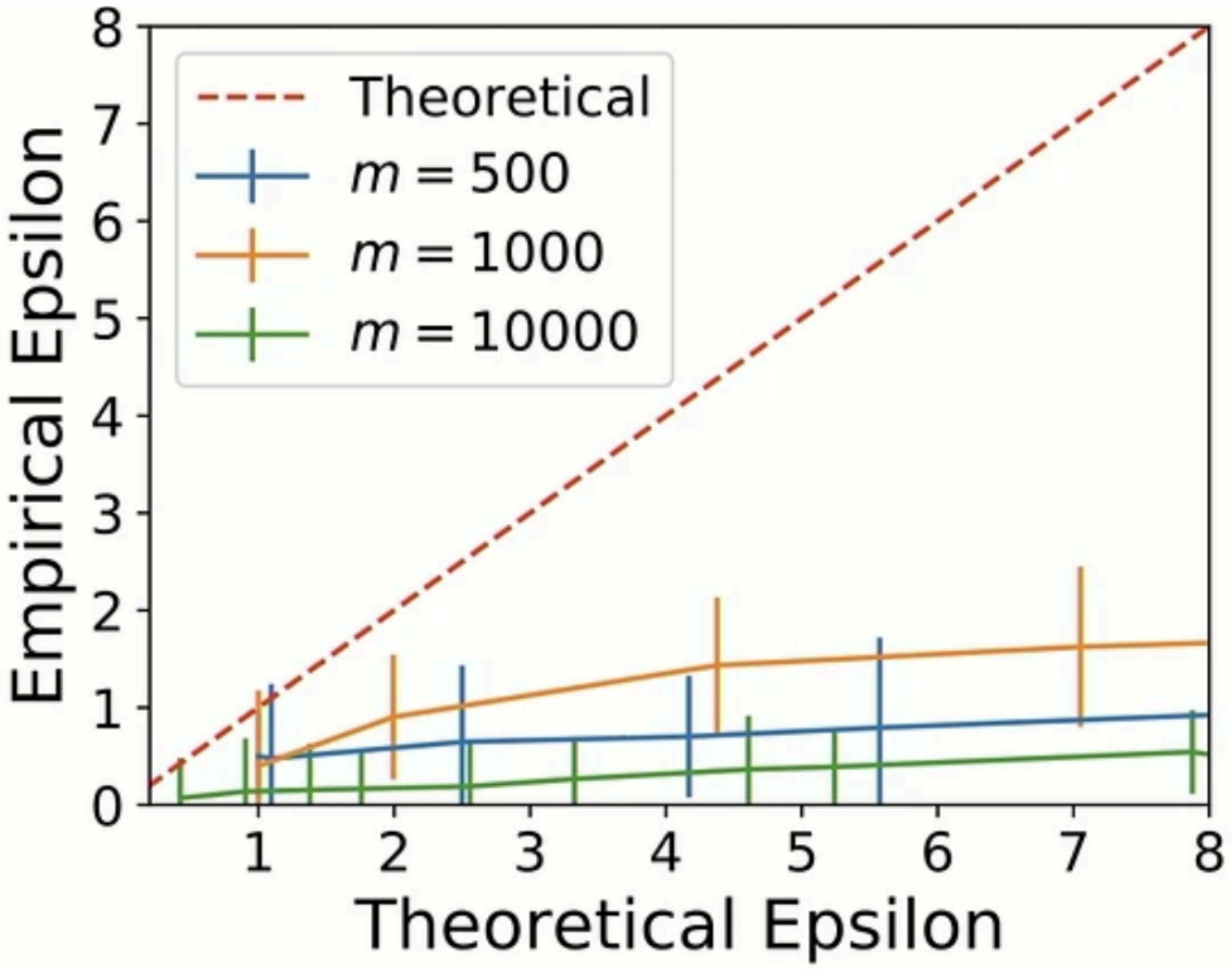
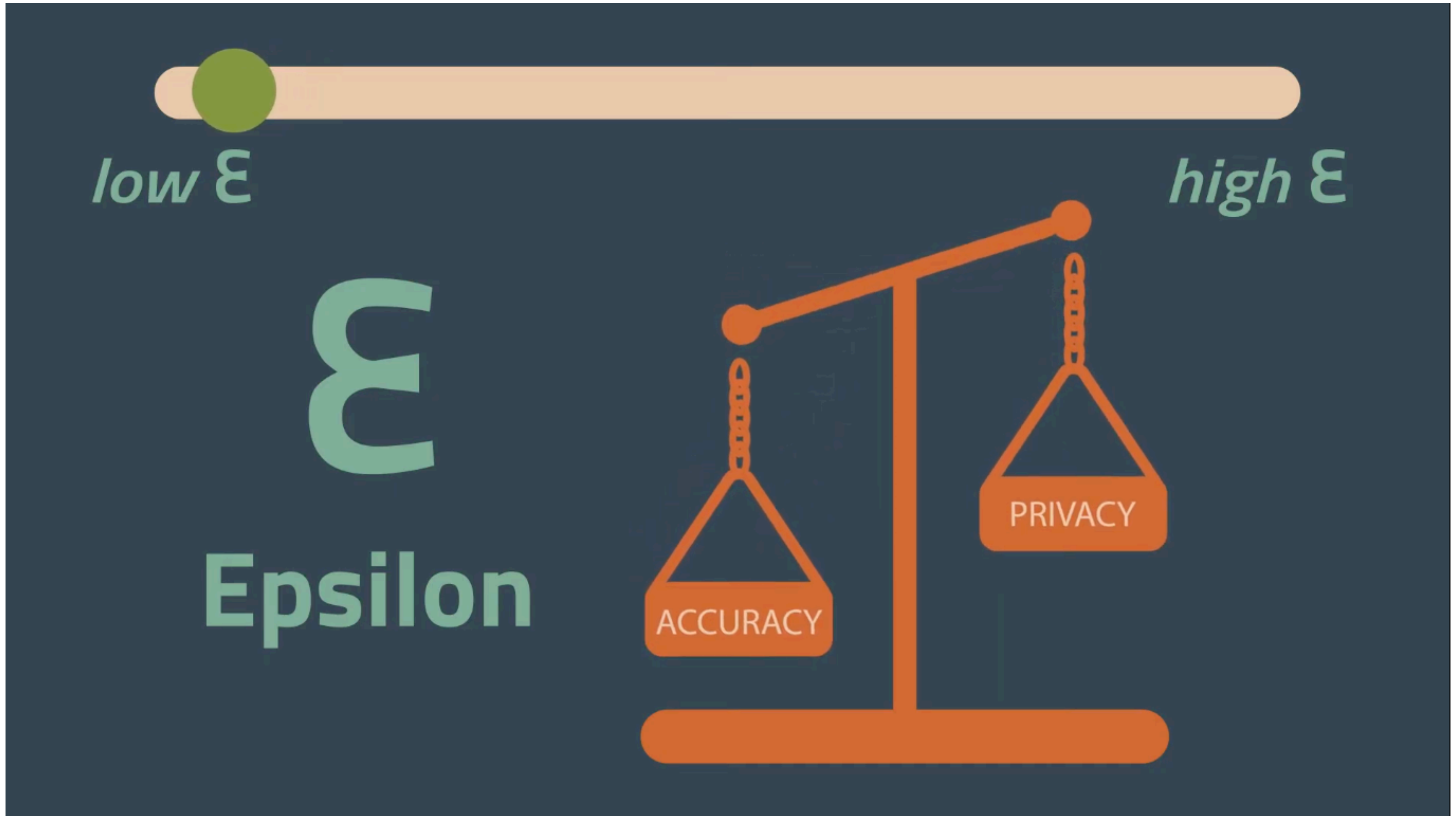


Figure 6. Effect of the number of auditing examples (m) in the black-box setting. Black-box auditing is very sensitive to the number of auditing examples.

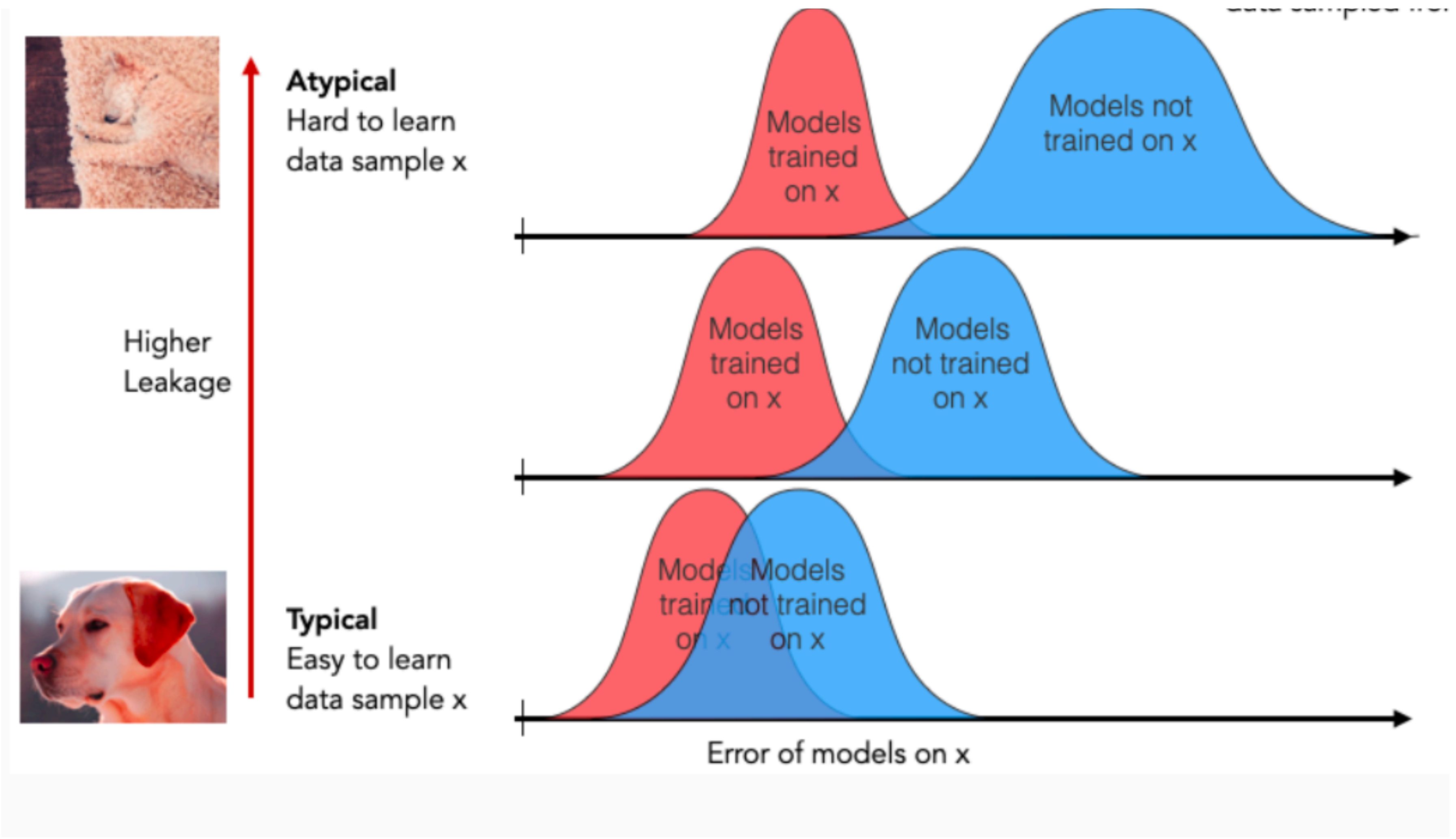
Adversary only sees final model weights (or can only query the loss)

Relaxations of DP



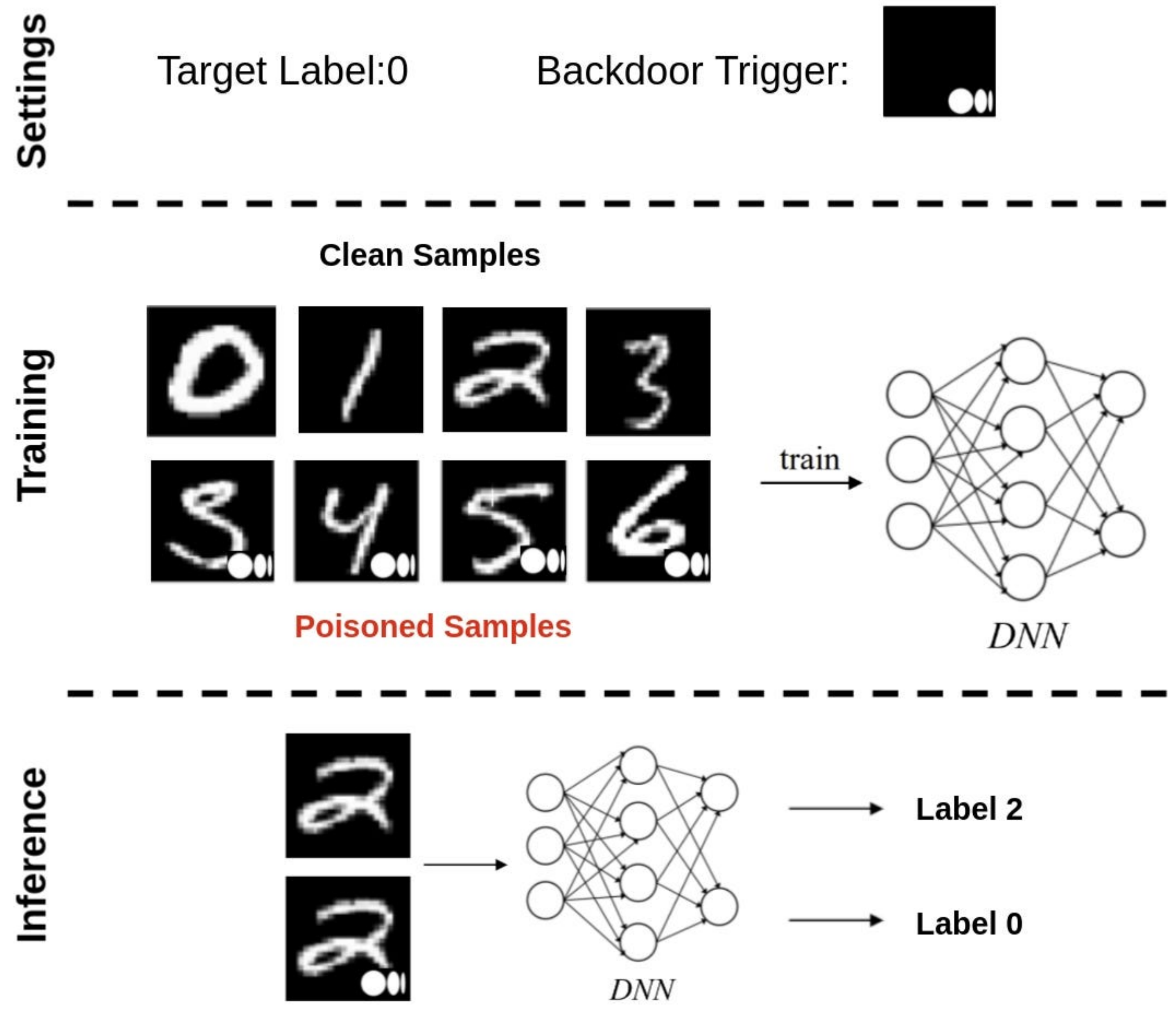
What is a “memorable image”?

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



- Insert backdoors / adversarial inputs

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

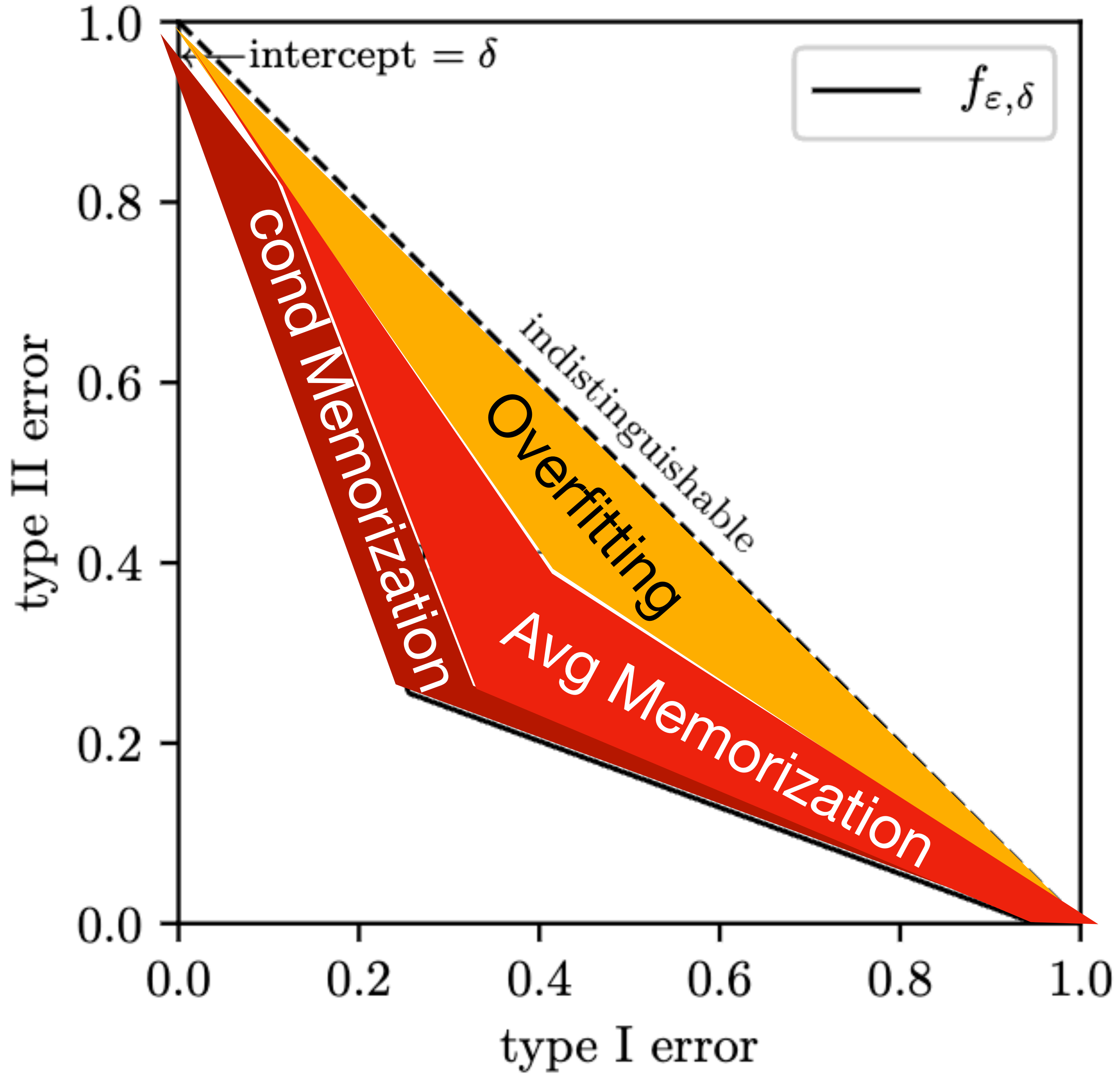


Memorization and Privacy

- Overfitting and memorization are both linked to privacy leakage.

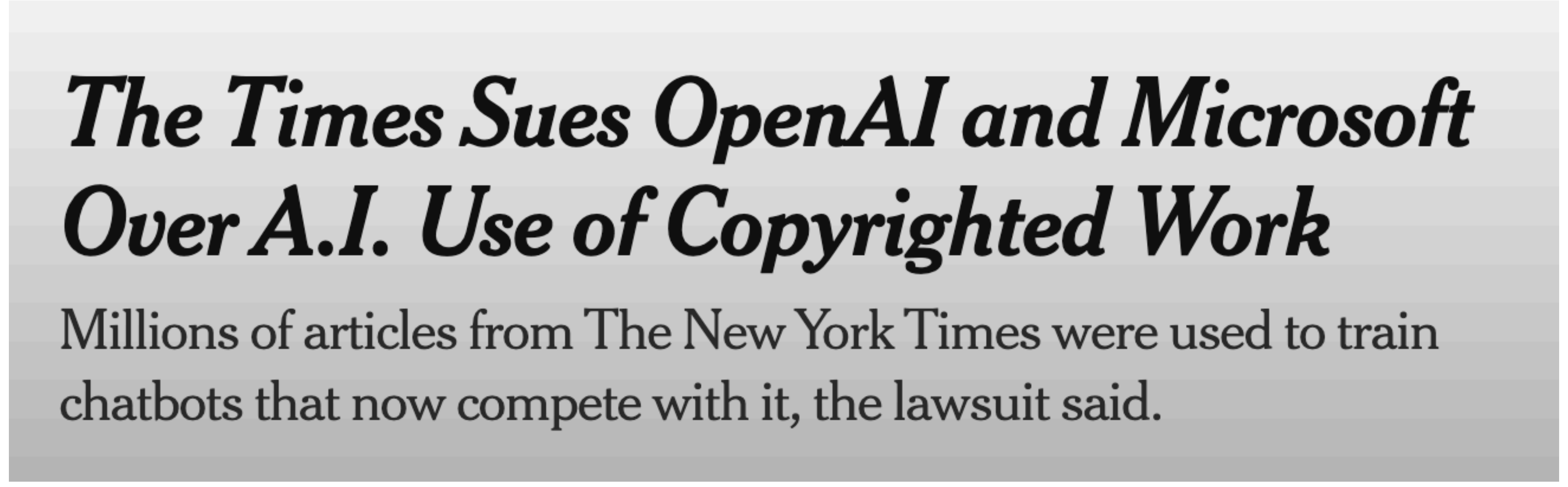
- In privacy auditing, we search for memorizing artificial images i.e. search for a “planted signal”. Called *conditional memorization*.

- Avg memorization asks how much of the real training data has been memorized.



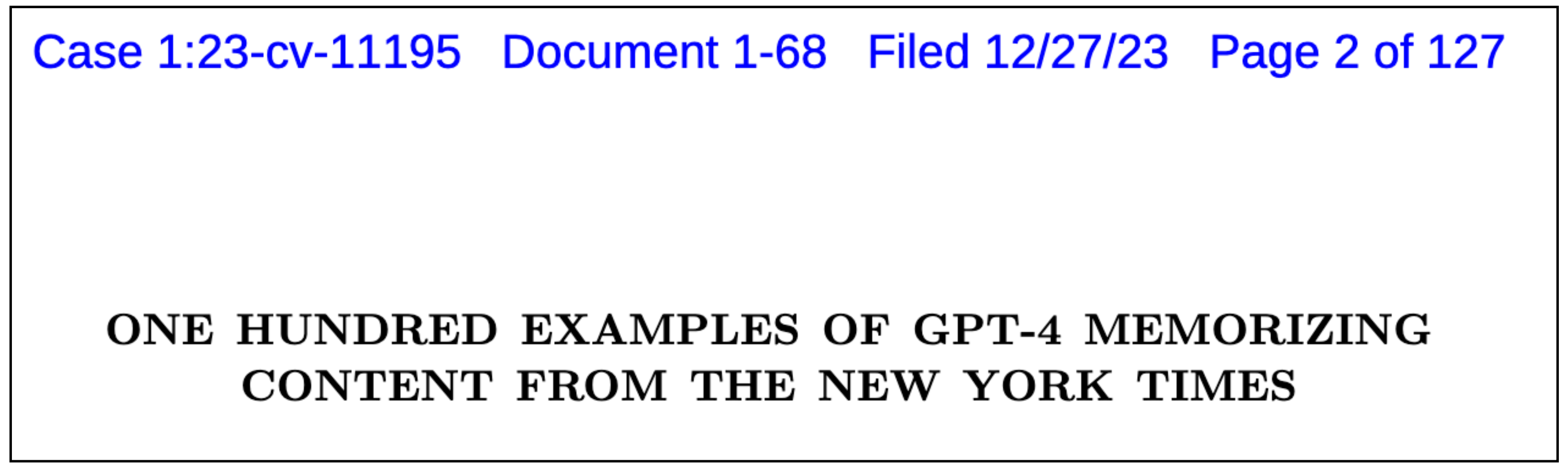
Measuring Average Memorization

- 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]



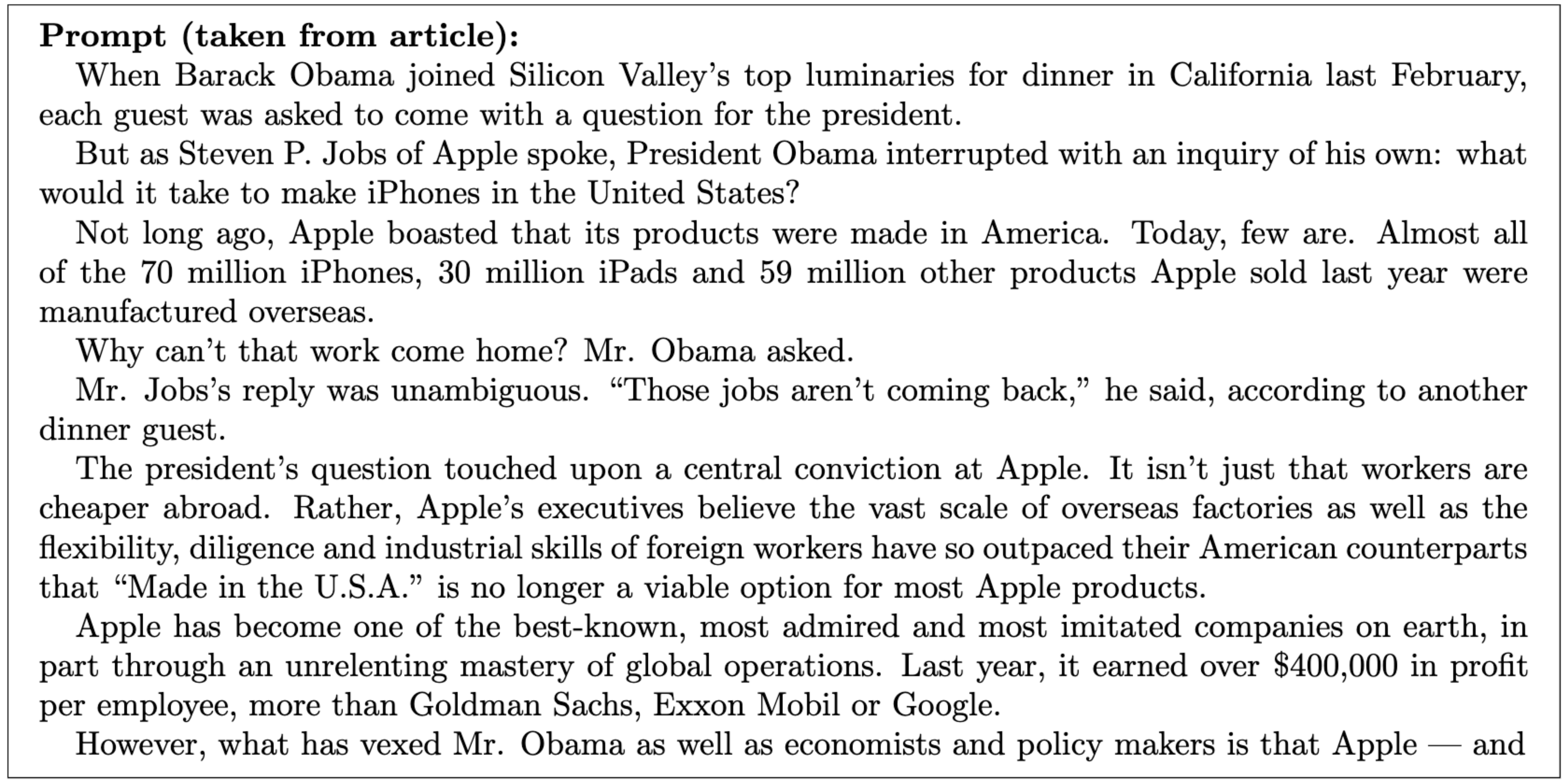
The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

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



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



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.

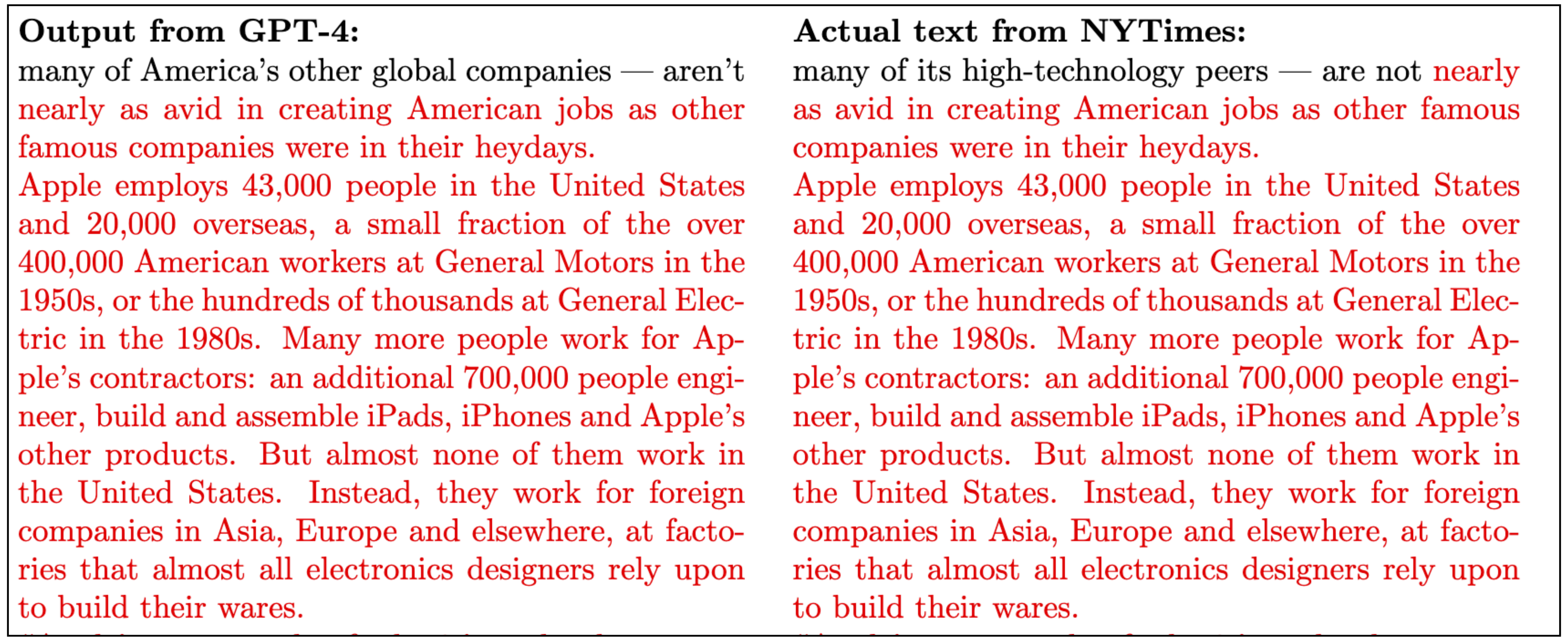
Why can't that work come home? Mr. Obama asked.

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

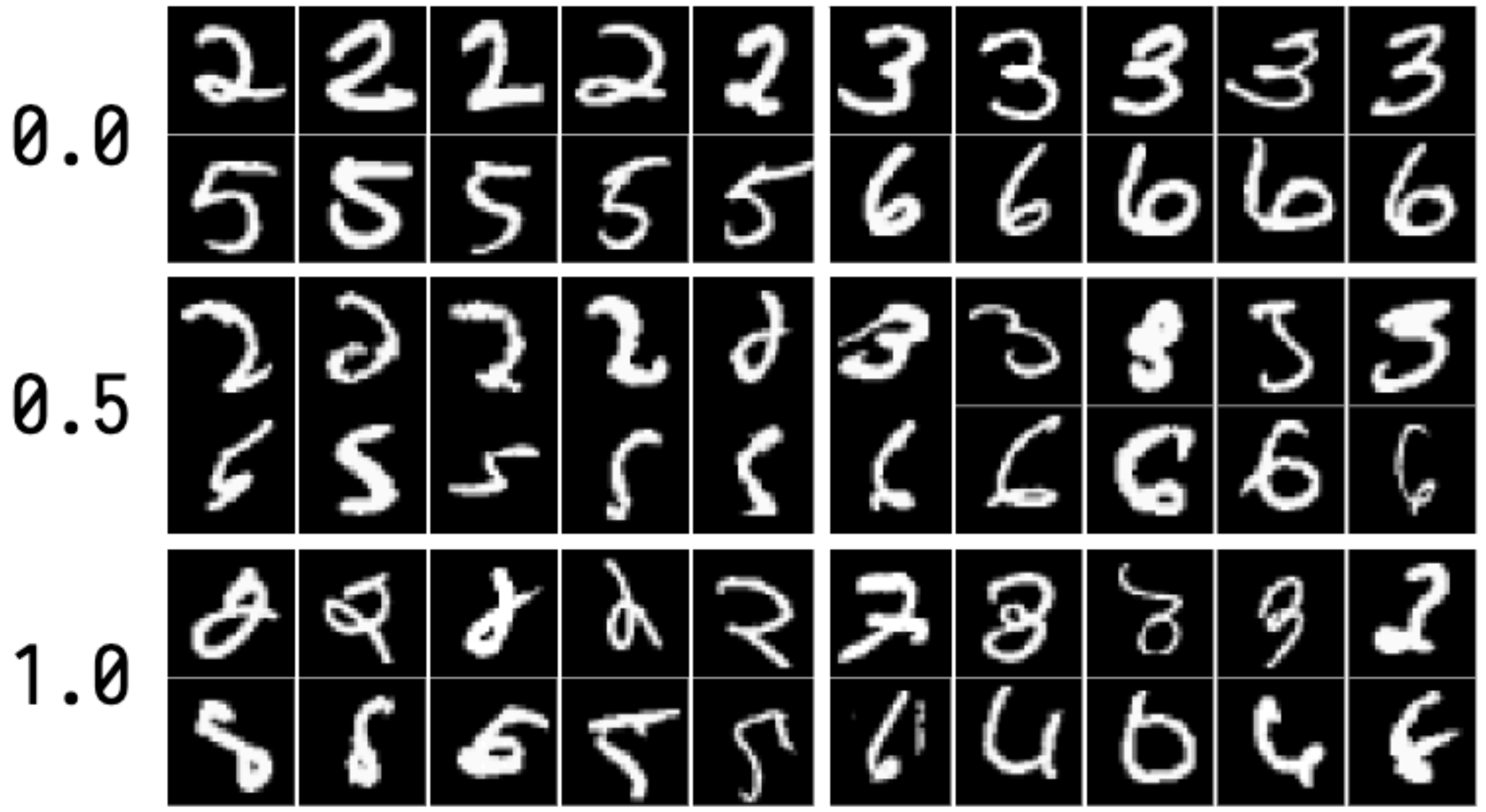


Output from GPT-4:	**Actual text from NYTimes:**
many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays. Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American 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.

Defining memorization

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

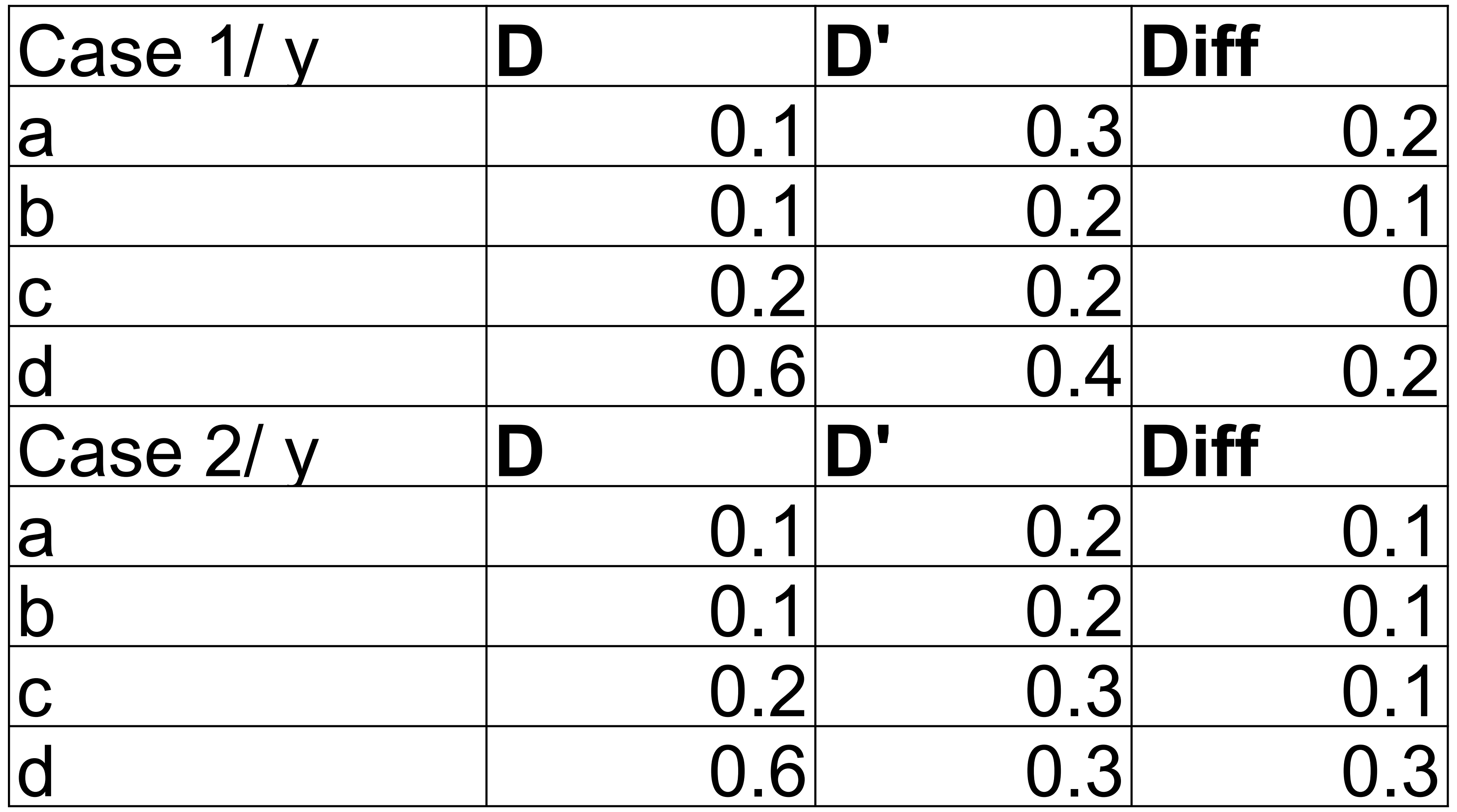
- **Memorization** [Fel 20] = $Pr_{h \leftarrow A(D)}[h(x) = y] - Pr_{h \leftarrow A(D')}[h(x) = y]$
    - For images: predict labels, in-painting, etc.
    - For text: recover tokens given context



Most memorized inputs [FZ'20]

Defining memorization

- $Pr_{h\leftarrow A(D)}[h(x) = y] - Pr_{h\leftarrow A(D')}[h(x) = y]$
- Memorization $\neq$ overfitting. k-NN, over-parameterized models memorize exactly. But still generalize.
- Differential privacy => low 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

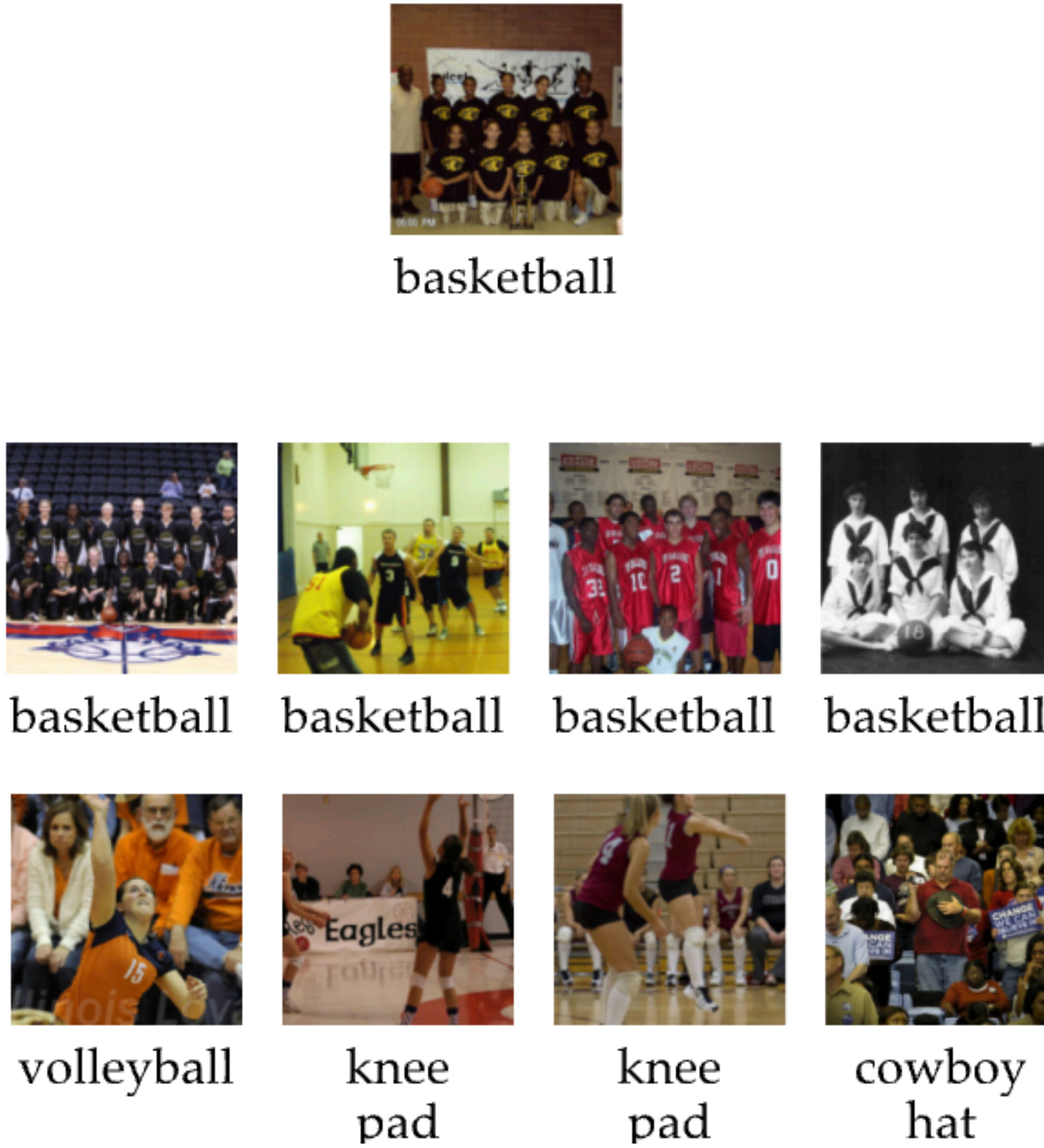
Did more memorization happen in case 1 or 2?

Influence estimation

- **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.

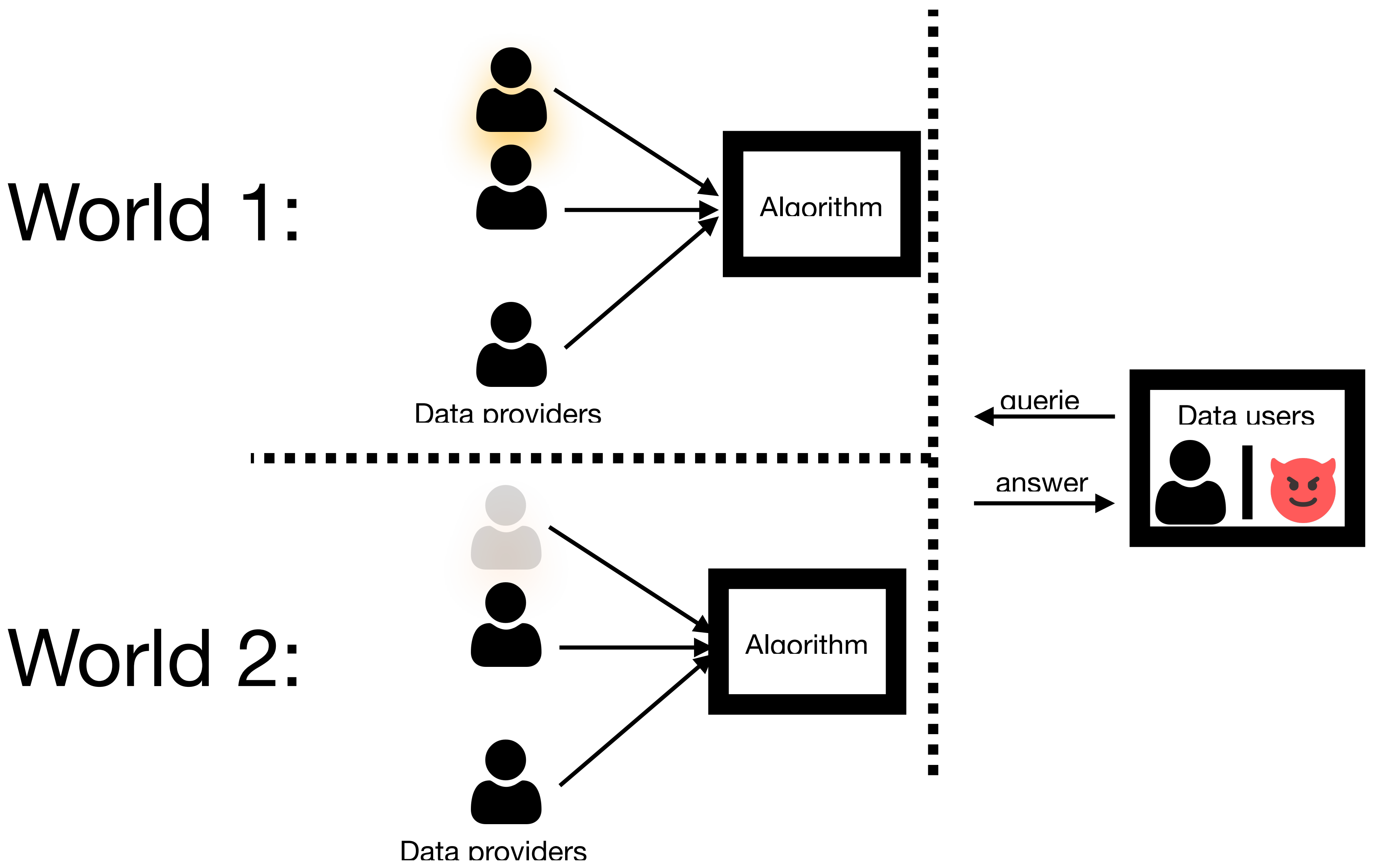


TRAK: [P+'23]

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:

Algorithm

Data providers

querie

answer

Data users

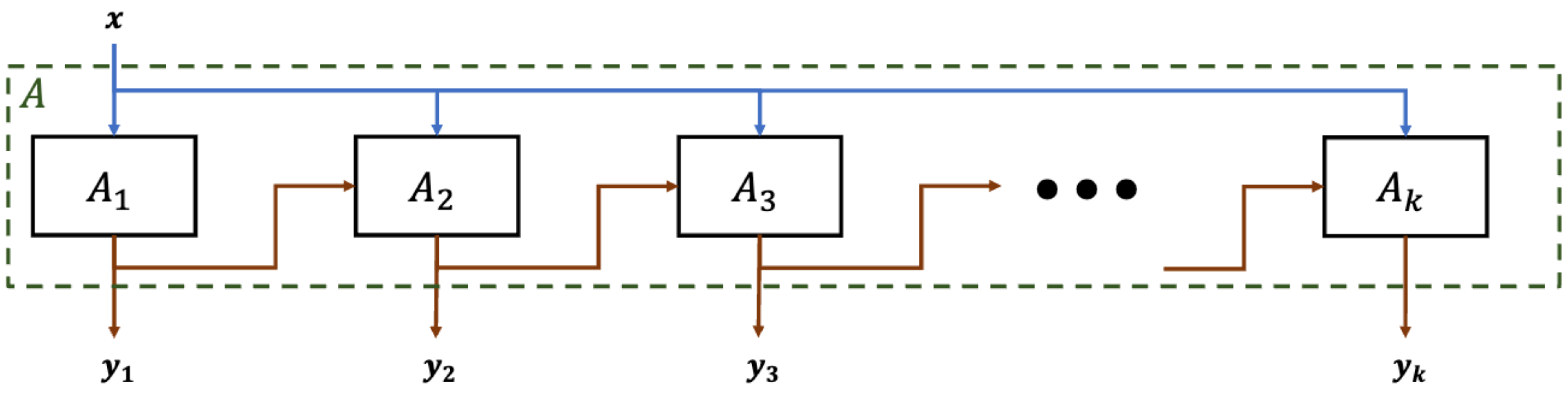
World 2:

Algorithm

Data providers

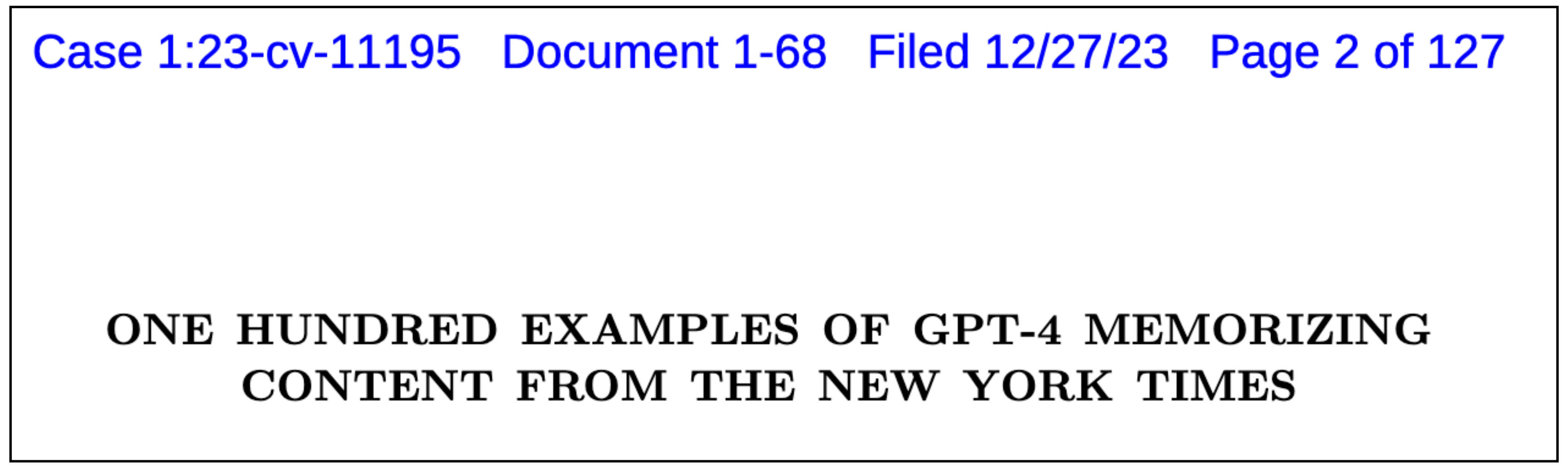
Datapoint level privacy measures

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



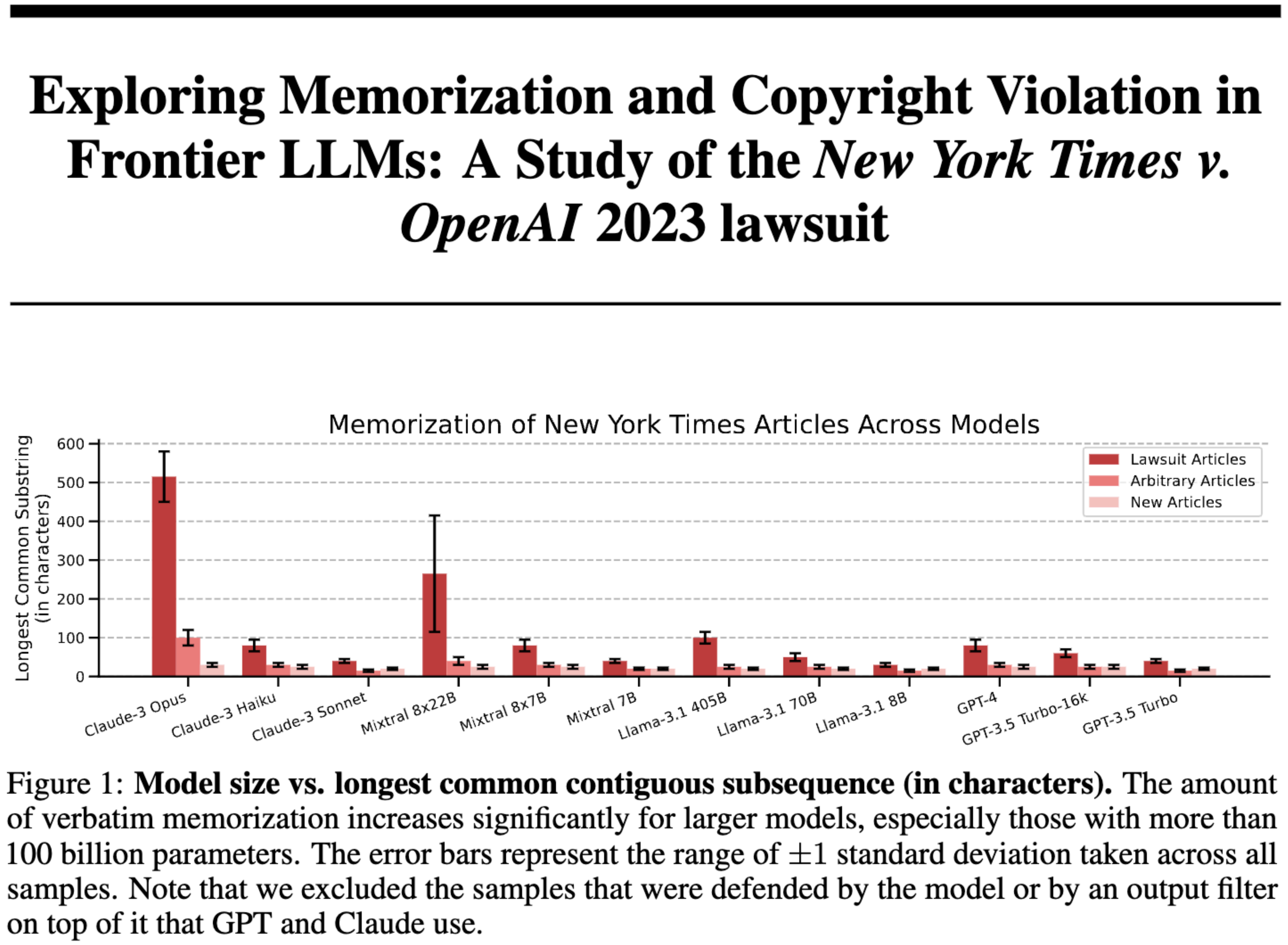
Lots of open questions

- 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?
- Idea: Can we view these attacks as attempts at auditing per-datapoint DP?
- Reconstruction vs. membership inference?



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



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.