Was My Data Used to Train a Large Language Model?

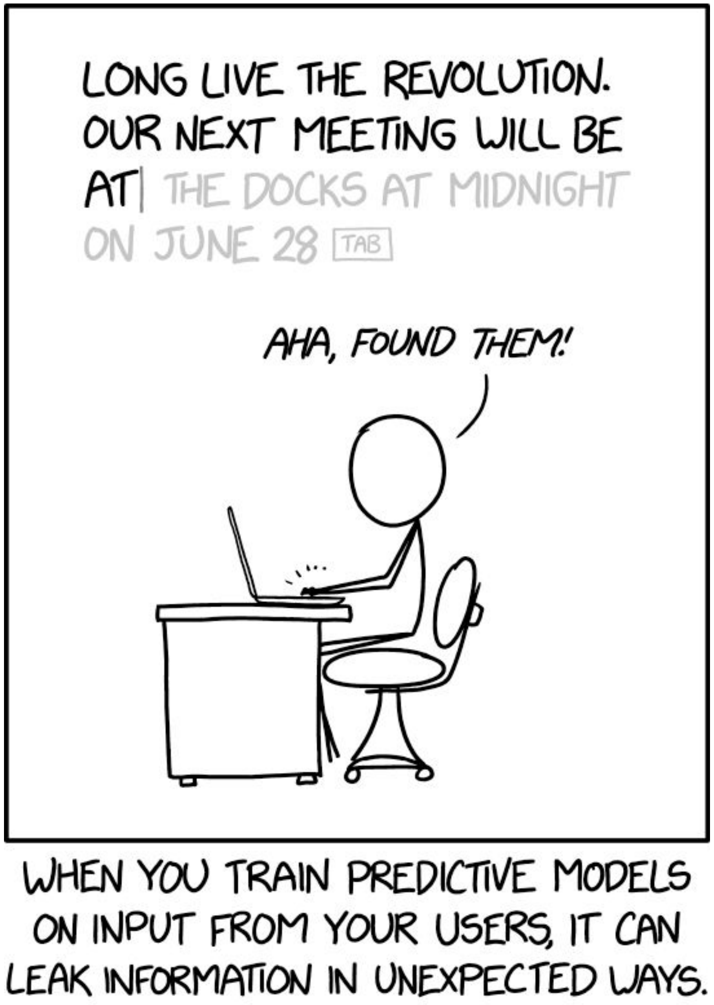
Krishna Pillutla

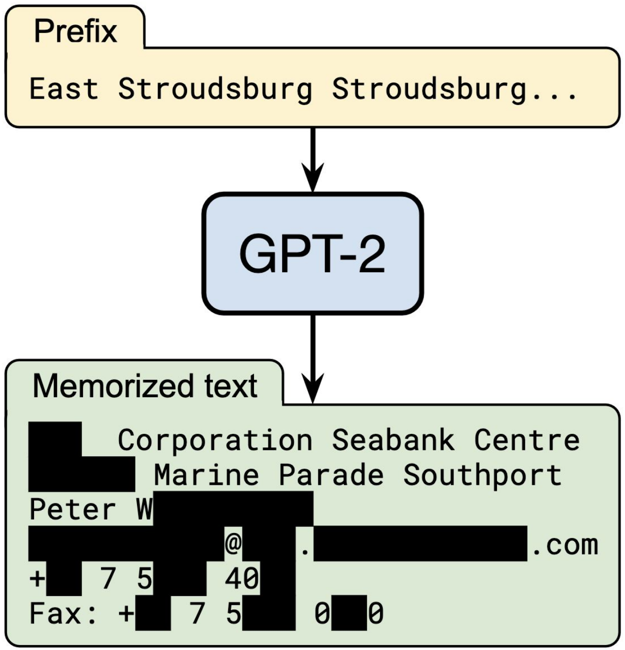
Nov. 16 2024 @ IIT Jodhpur

AI in Healthcare Symposium

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

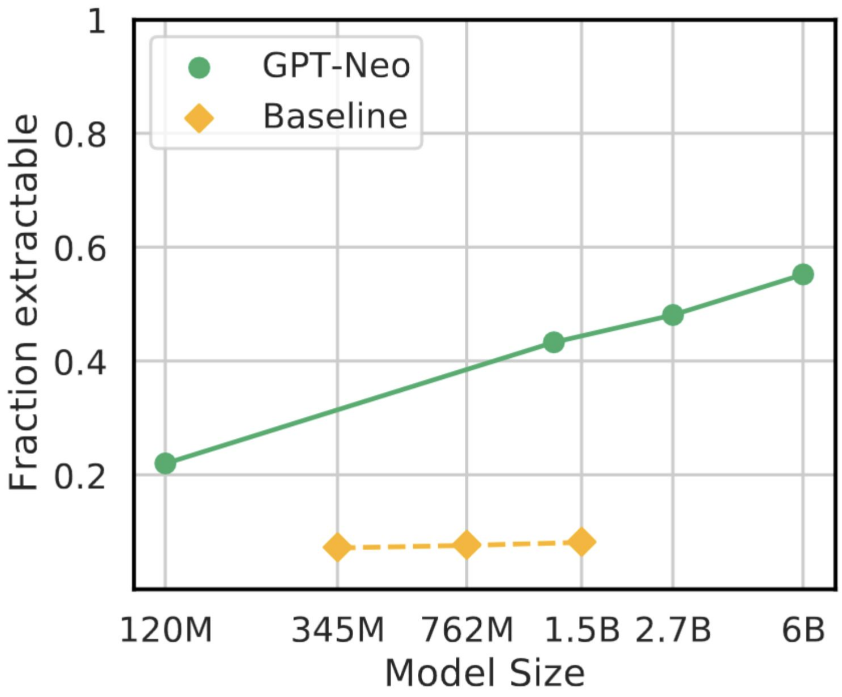
Models leak information about their training data



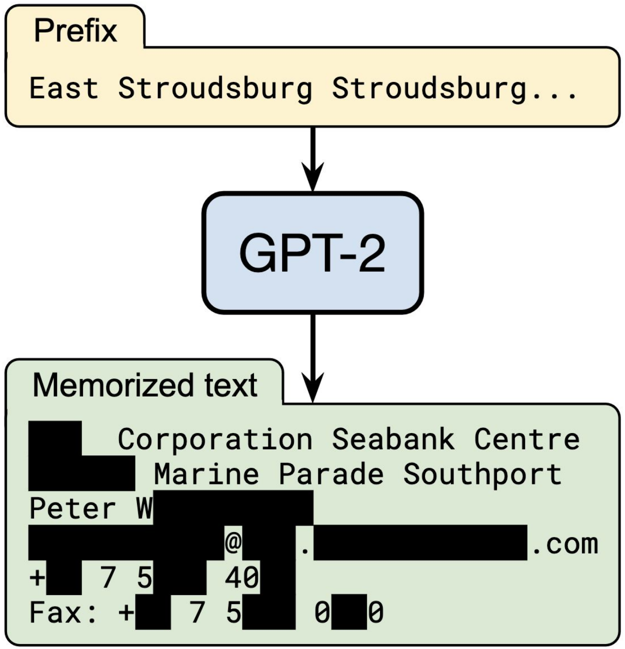


Carlini et al. (USENIX Security 2021)

Models leak information about their training data ***reliably***



Carlini et al. (ICLR 2023)



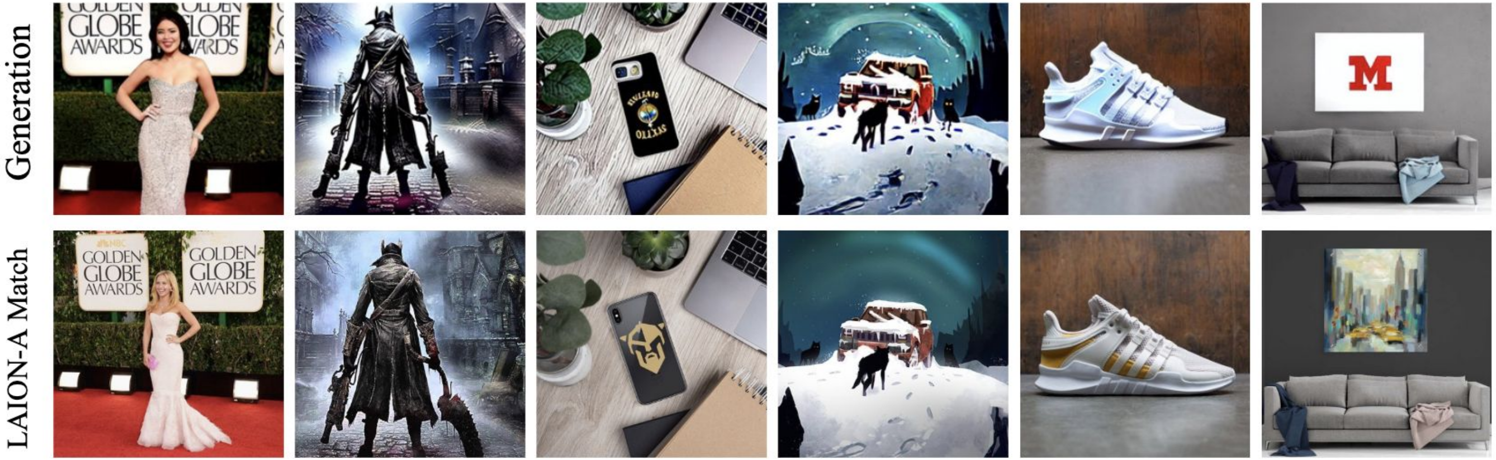
Carlini et al. (USENIX Security 2021)

Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models

Gowthami Somepalli [🐢], Vasu Singla [🐢], Micah Goldblum [🗽], Jonas Geiping [🐢], Tom Goldstein [🐢]

[🐢] University of Maryland, College Park
{gowthami, vsingla, jgeiping, tomg}@cs.umd.edu

[🗽] New York University
goldblum@nyu.edu



FORBES > INNOVATION > AI

Generative AI ChatGPT Can Disturbingly Gobble Up Your Private And Confidential Data, Forewarns AI Ethics And AI Law

Lance Eliot Contributor

Dr. Lance B. Eliot is a world-renowned expert on Artificial Intelligence (AI) and Machine Learning...

Follow

How Strangers Got My Email Address From ChatGPT's Model

By Jeremy White Dec. 22, 2023

Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak

- Employees accidentally leaked sensitive data via ChatGPT
- Company preparing own internal artificial intelligence tools

By Mark Gurman

May 2, 2023 at 6:18 AM GMT+5:30

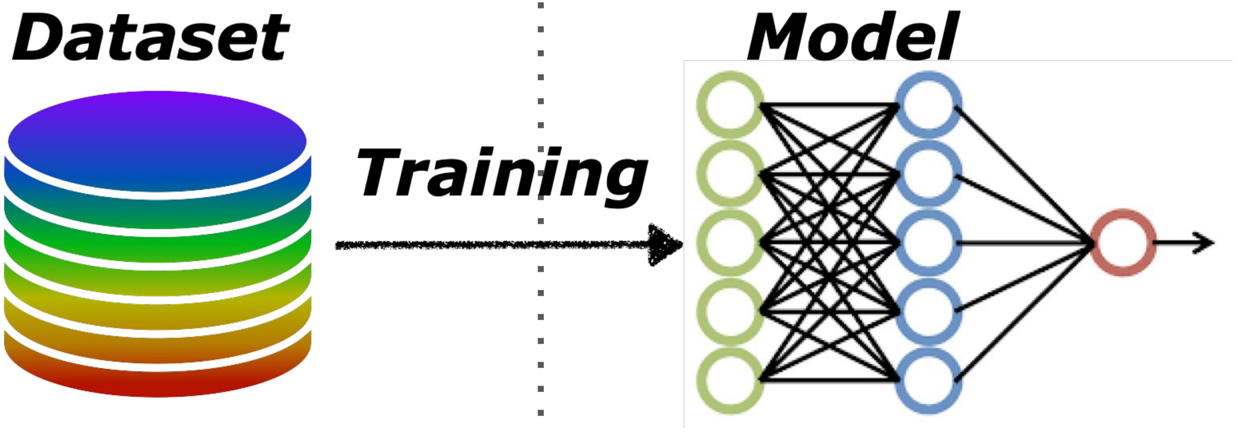
Nvidia's AI software tricked into leaking data

Researchers manipulate feature in ways that could reveal sensitive information

LILY HAY NEWMAN ANDY GREENBERG SECURITY DEC 2, 2023 9:00 AM

Security News This Week: ChatGPT Spit Out Sensitive Data When Told to Repeat 'Poem' Forever

Privacy attacks



Secure location

Public access

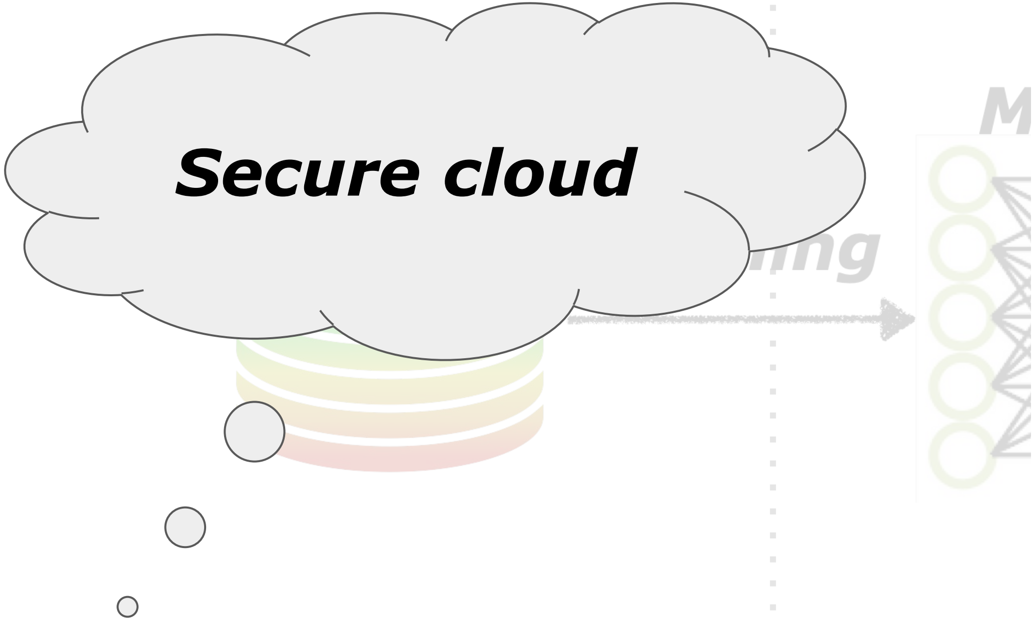
Privacy attacks

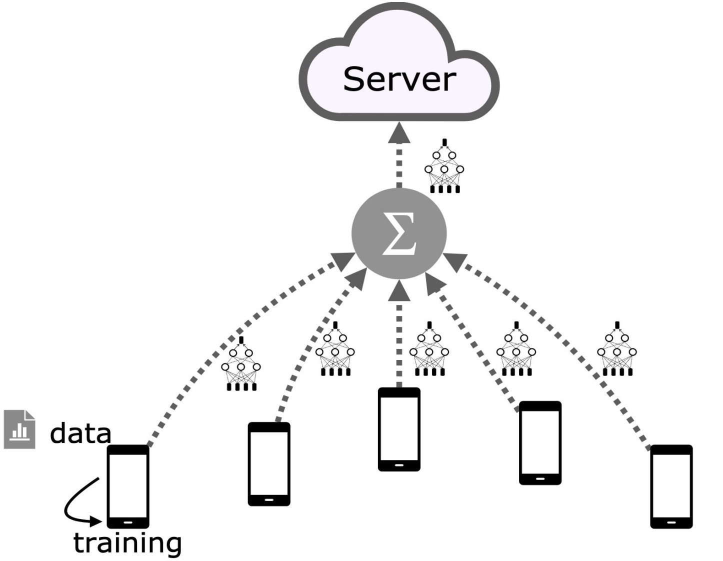
Secure location

Public access

Privacy attacks

Federated learning



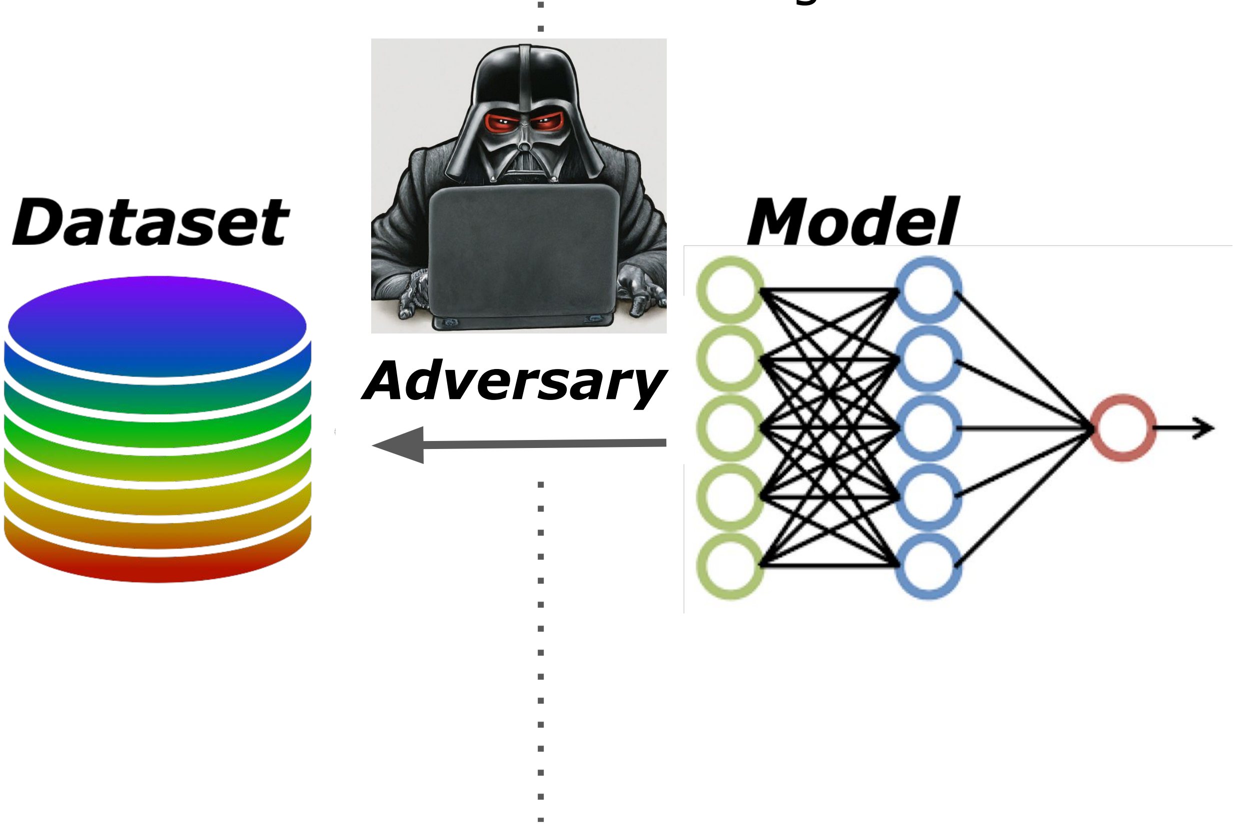


Secure location

Public access

Privacy attacks:

Adversary uses the model to infer something about the data



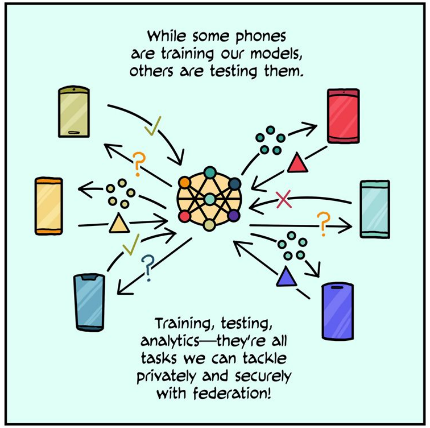
Secure location

Public access

What does the word "***privacy***" mean to an end user?

Transparency, Control, Verifiability

Minimize data sharing

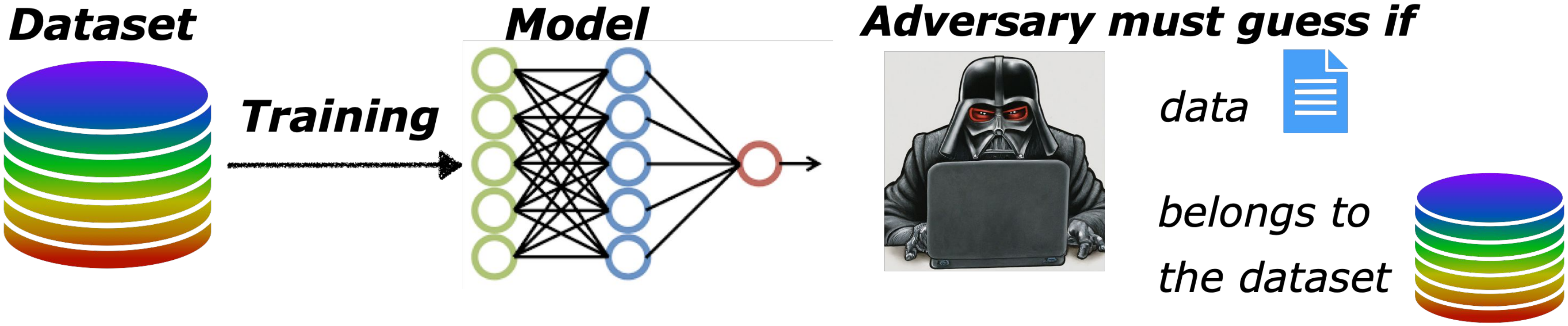


https://federated.withgoogle.com/

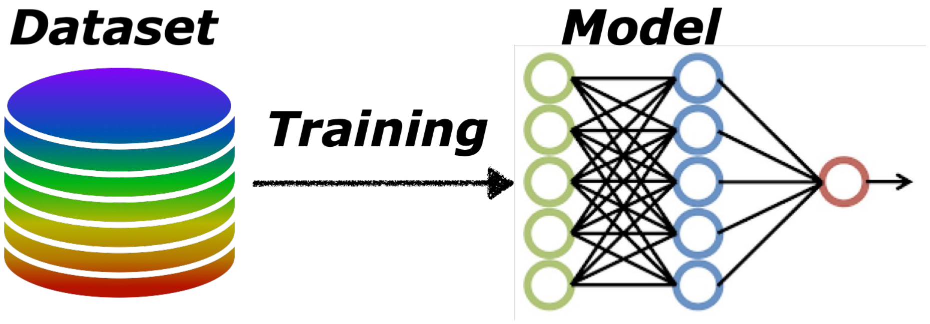
Data Anonymization

Bonawitz, Kairouz, McMahan, Ramage (2022). **Federated Learning and Privacy**. *Communications of the ACM*.

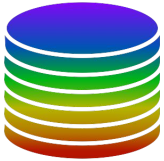
Basic privacy attack: Membership inference

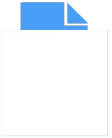


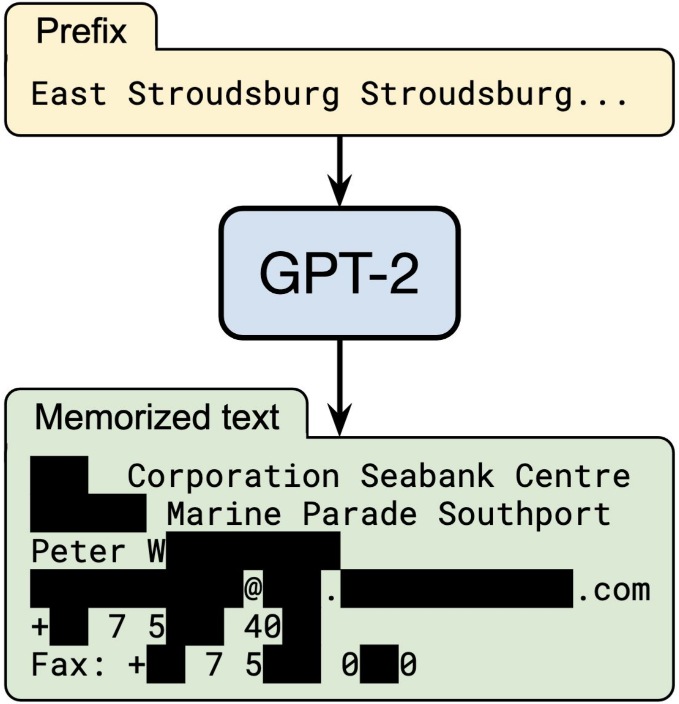
Basic privacy attack: Data extraction



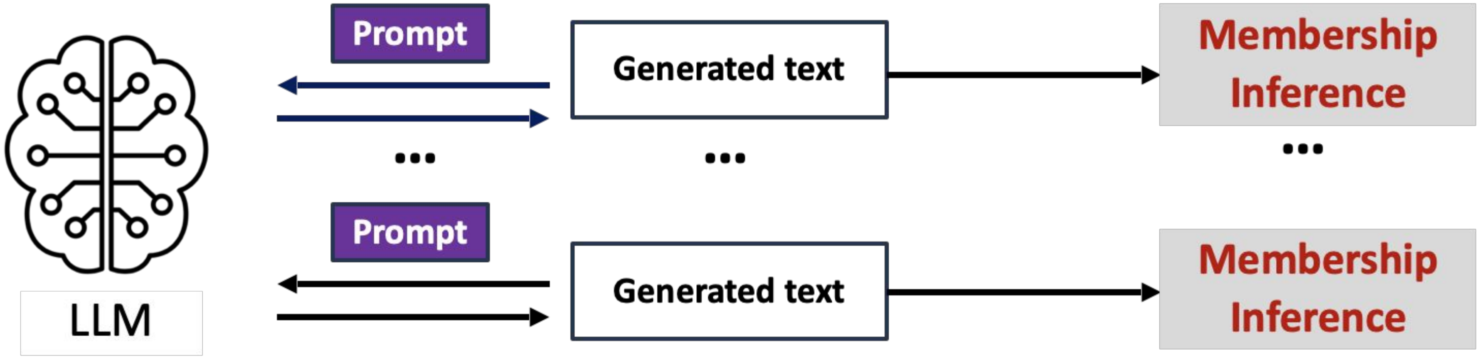
Adversary must extract

data  ∈ 

from a small component 



Data extraction using membership inference



Step 1 (Repeat):

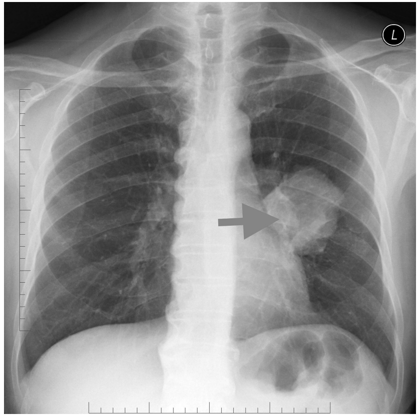
- Prompt model with random tokens to generate lots of text

Step 2:

- Membership Inference attack determines if each sample was in training set

Carlini et al. (USENIX Security 2021)

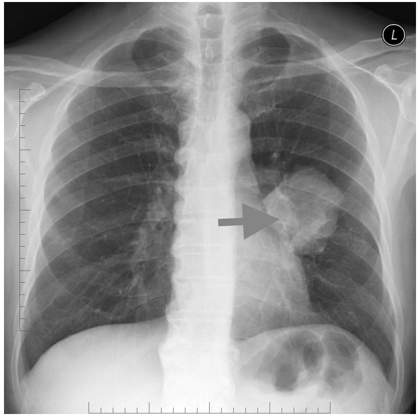
Example scenario 1: Incorporating sensitive metadata



TB or

No TB

Example scenario 1: Incorporating sensitive metadata



TB or

No TB

Privacy-sensitive!

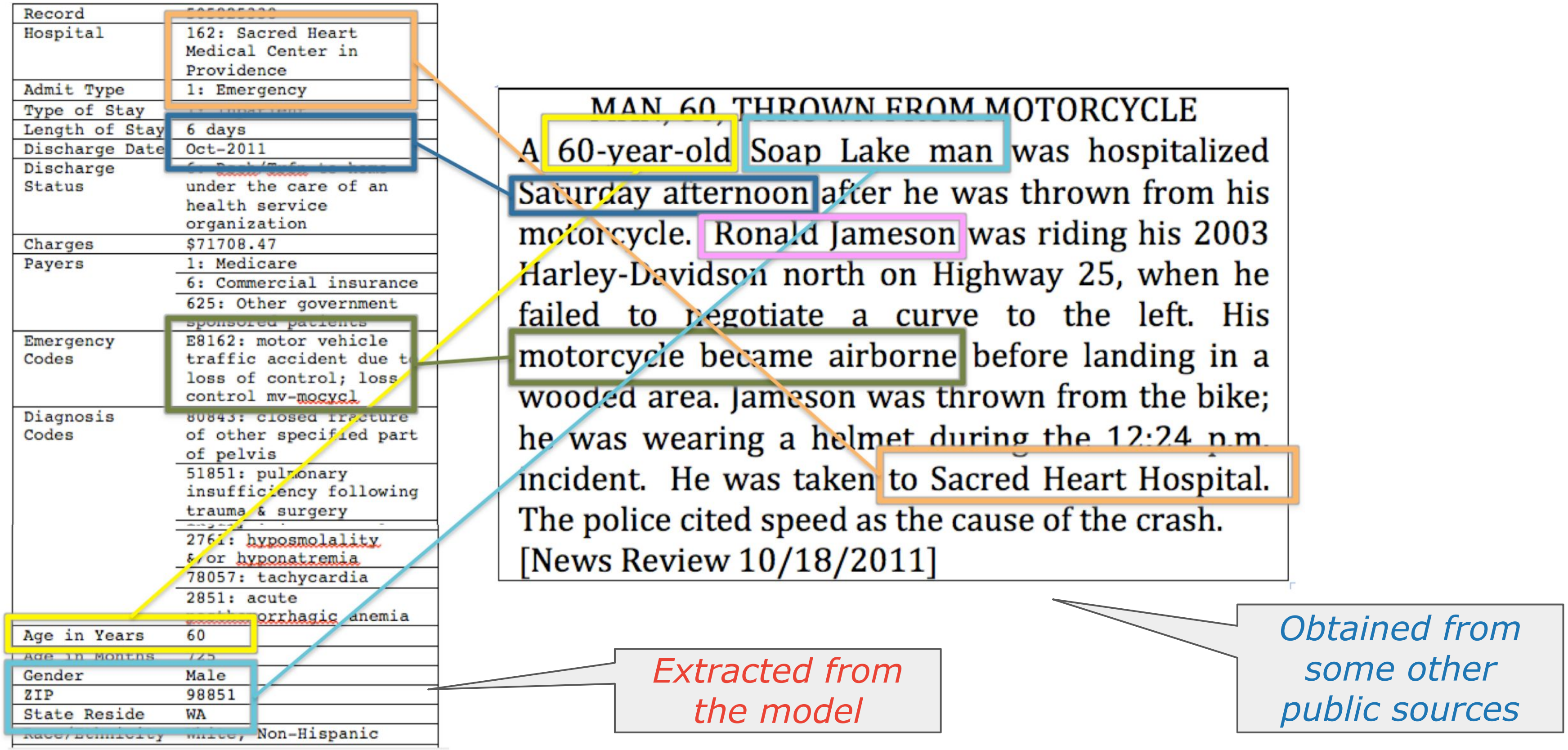
+ Patient info:

- Symptoms
- Comorbidities (tobacco/HIV/diabetes/...)
- Family/location history

Linkage Attacks:

Combining information from multiple sources

Example/Image credit:
Latanya Sweeney



Record	[illegible]
Hospital	162: Sacred Heart Medical Center in Providence
Admit Type	1: Emergency
Type of Stay	[illegible]
Length of Stay	6 days
Discharge Date	Oct-2011
Discharge Status	6: Dsch/Trfr to home under the care of an health service organization
Charges	$71708.47
Payers	1: Medicare
	6: Commercial insurance
	625: Other government sponsored patients
Emergency Codes	E8162: motor vehicle traffic accident due to loss of control; loss control mv-mocycl
Diagnosis Codes	80843: closed fracture of other specified part of pelvis
	51851: pulmonary insufficiency following trauma & surgery
	2761: hyposmolality &/or hyponatremia
	78057: tachycardia
	2851: acute posthemorrhagic anemia
Age in Years	60
Age in Months	725
Gender	Male
ZIP	98851
State Reside	WA
Race/Ethnicity	White, Non-Hispanic

MAN, 60, THROWN FROM MOTORCYCLE

A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash. [News Review 10/18/2011]

Linkage attacks are the Achilles heel of patient data de-identification/anonymization.

Piers Nash
Innovative AI/Data Strategist | PhD, MBA | Mentor/Advisor

+ Follow

Published Mar 7, 2023

De-identification is a process that removes personal identifiers from data, such as a person's name, address, or social security number. The goal of de-identification is to make it difficult or impossible to re-identify individuals from the data. However, the effectiveness of de-identification depends on the methods used and the context in which the data will be used.

Example scenario 2: Voice-enabled chatbot / transcription

చెప్పండి...
ఏమయ్యింది?

వారం రోజులు
తలనొప్పిగా ఉంది

Image Credit: Imagen 3

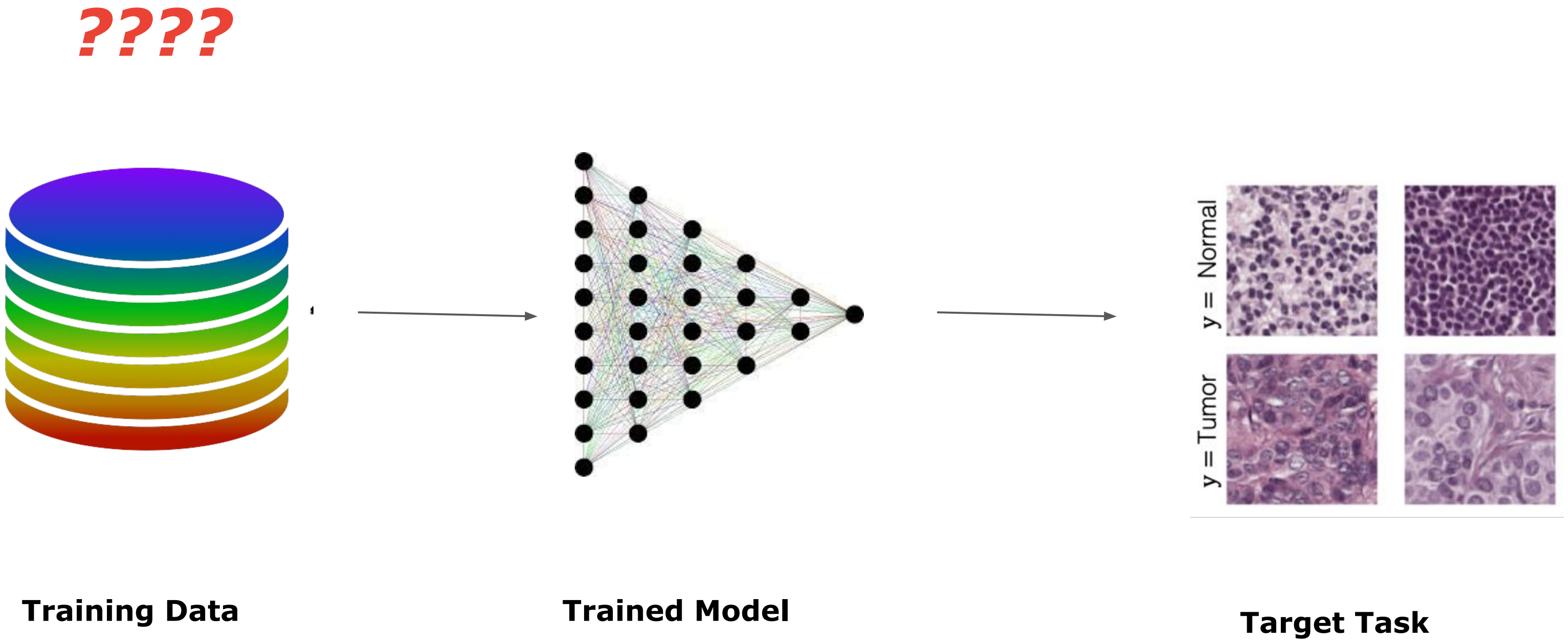
Patient: [Patient Name]

Chief Complaint: Headache for the last week

History of Present Illness:

- Patient reports experiencing headaches for the past 7 days.
- Associated symptoms: nausea, vomiting, sensitivity to light/sound, dizziness, visual disturbances, etc..
- ….

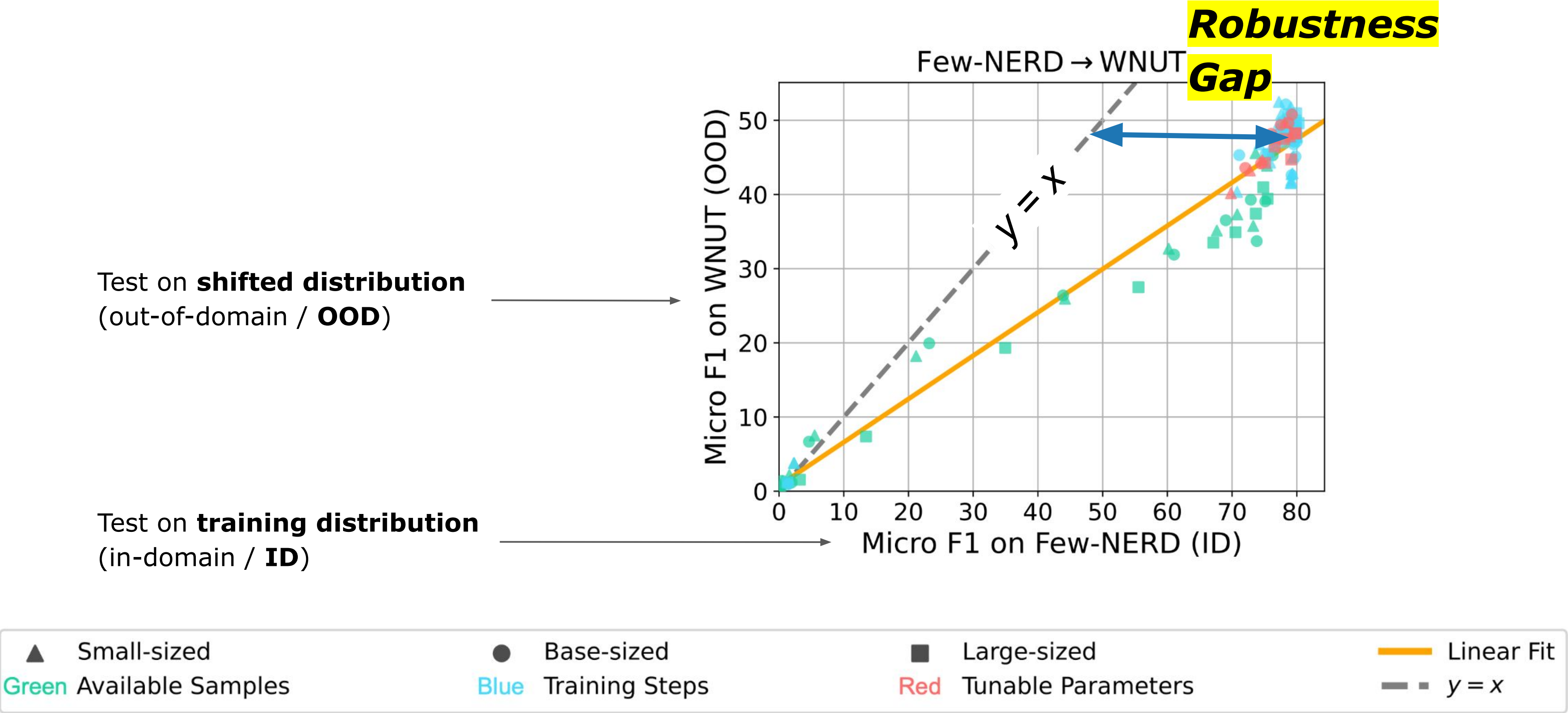
Which data do we use to train/finetune/align these models?



Which data do we use to train/finetune/align these models?

Test on **shifted distribution** (out-of-domain / **OOD**)

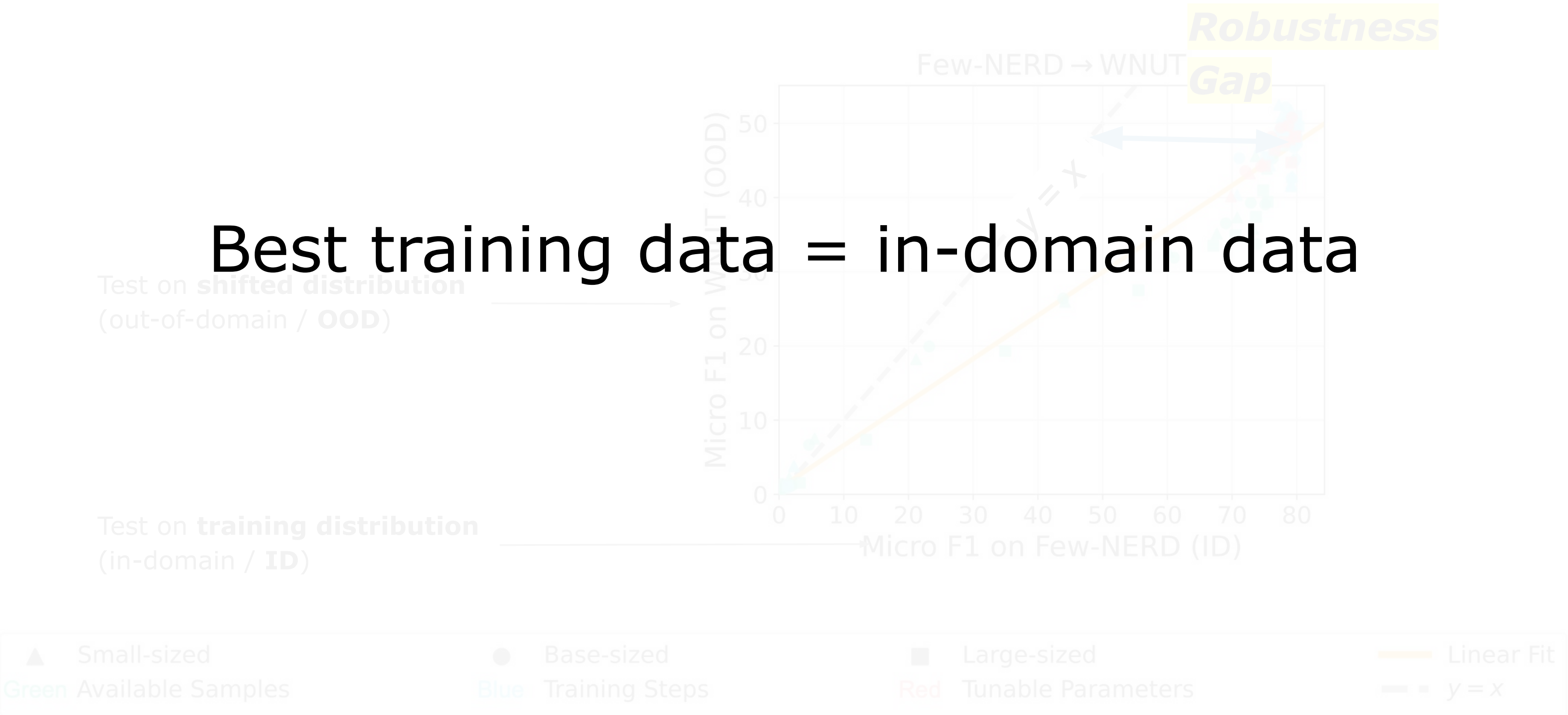
Test on **training distribution** (in-domain / **ID**)



Yuan et al. **Revisiting Out-of-distribution Robustness in NLP: Benchmark, Analysis, and LLM Evaluations**. NeurIPS D&B 2023

Which data do we use to train/finetune/align these models?

Best training data = in-domain data



Yuan et al. **Revisiting Out-of-distribution Robustness in NLP: Benchmark, Analysis, and LLM Evaluations**. NeurIPS D&B 2023

Blog
Introducing ChatGPT

Gemini

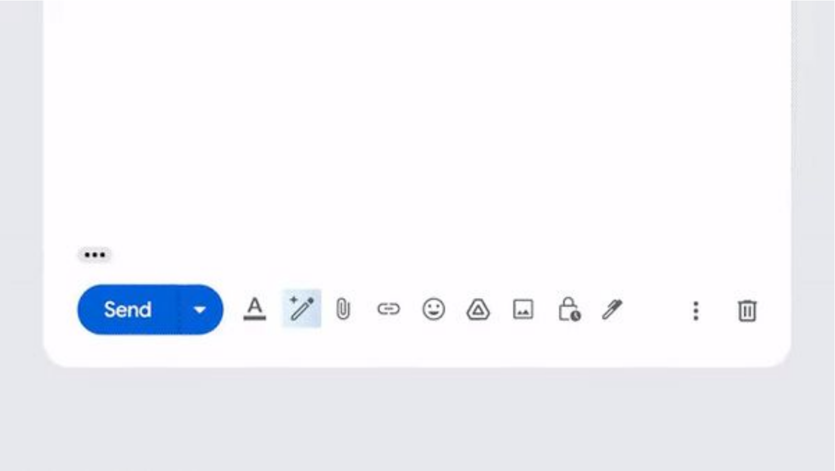
Claude
ANTHROP\C

LLAMA 2

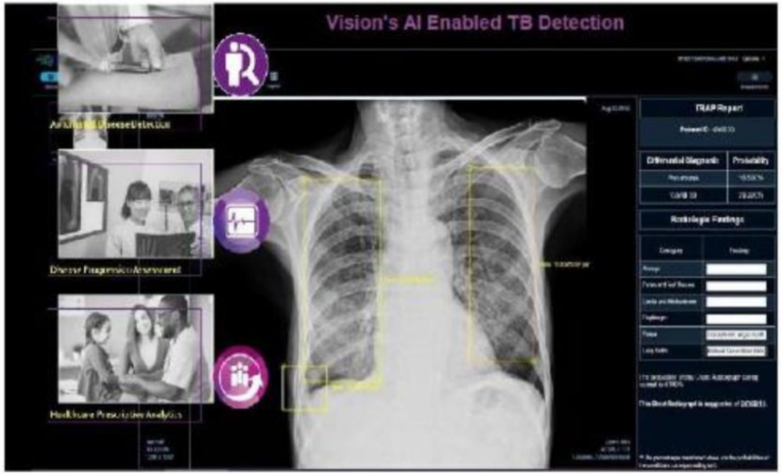
MISTRAL AI_

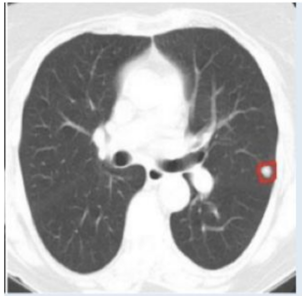
BharatGPT

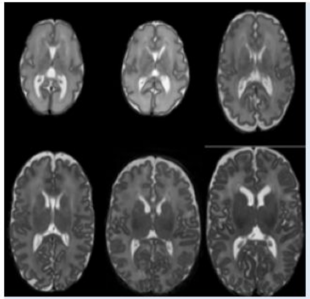
Gmail

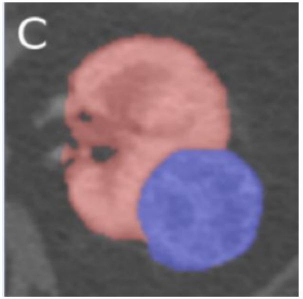


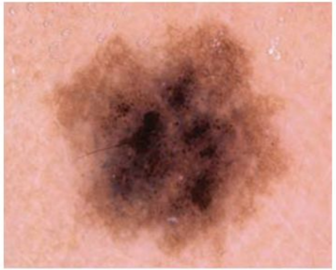
https://blog.google/products/gmail/gmail-ai-features/


Vision's AI Enabled TB Detection






C



For many applications, in-domain data = **user data**

For many applications, in-domain data = **user data**

Each **user** can contribute ***multiple*** examples

additional
privacy risks!

ChatGPT leaks sensitive conversations, ignites privacy concerns: Here's what happened

Privacy and security concerns have resurfaced after leaked conversations were discovered on OpenAI's AI-driven chat platform, ChatGPT. The incident raises questions about the vulnerabilities of AI systems despite assurances of safeguards.

Livemint

Updated • 31 Jan 2024, 06:31 PM IST

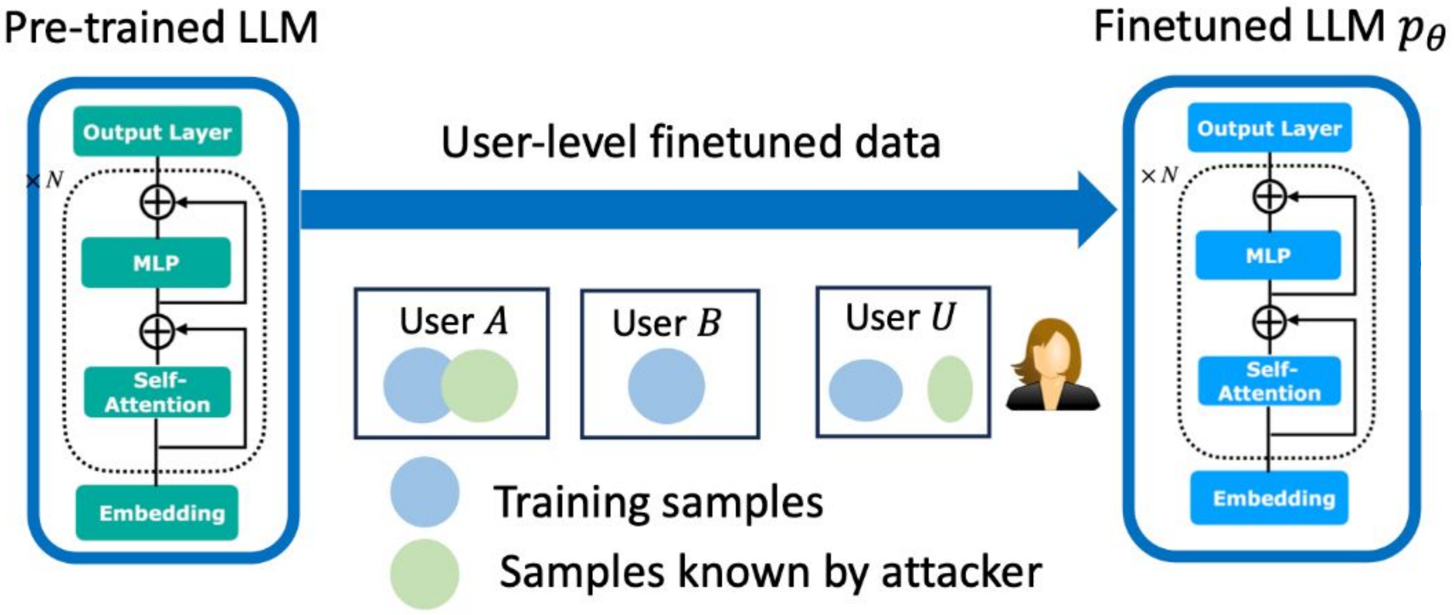
A giant dataset of YouTube subtitles has, per a new investigation, been used to train countless AI models without the permission of the tens of thousands of creators whose work was scraped.

Gemini AI platform accused of scanning Google Drive files without user permission

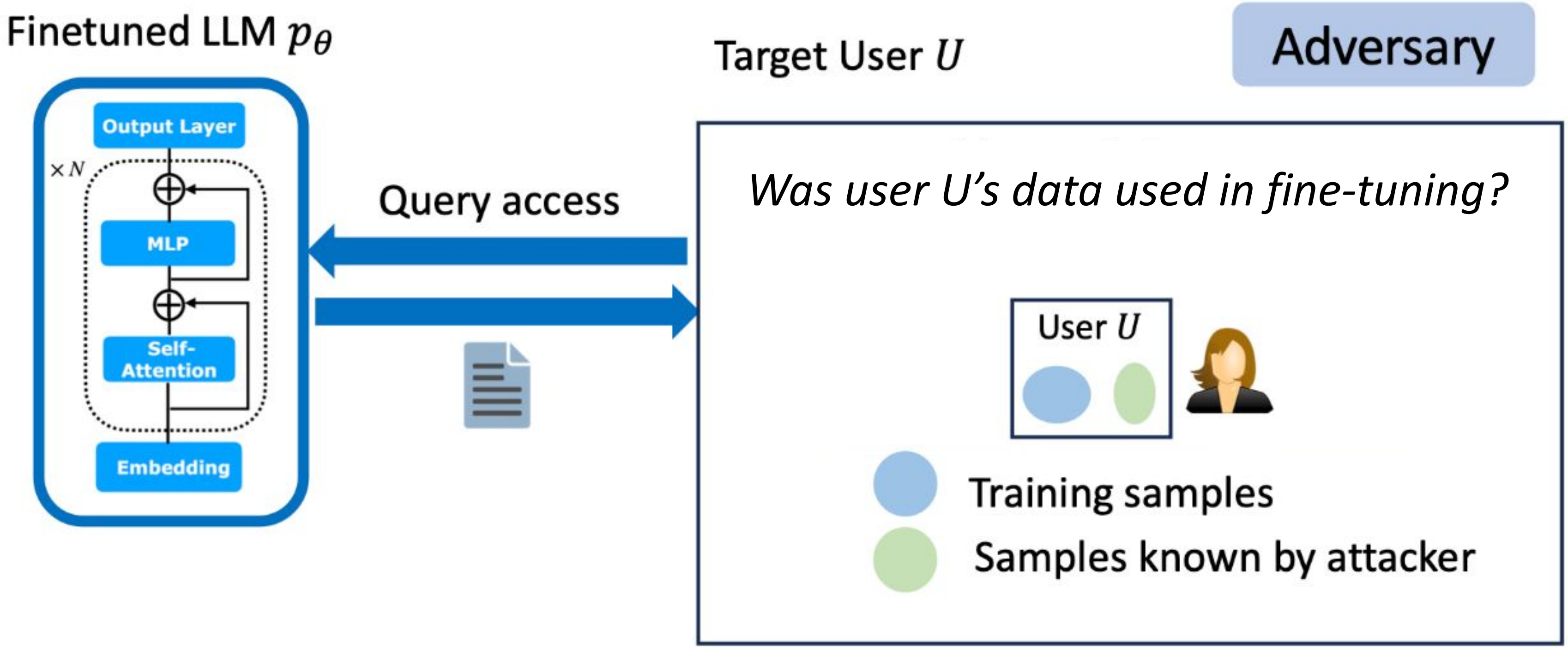
News

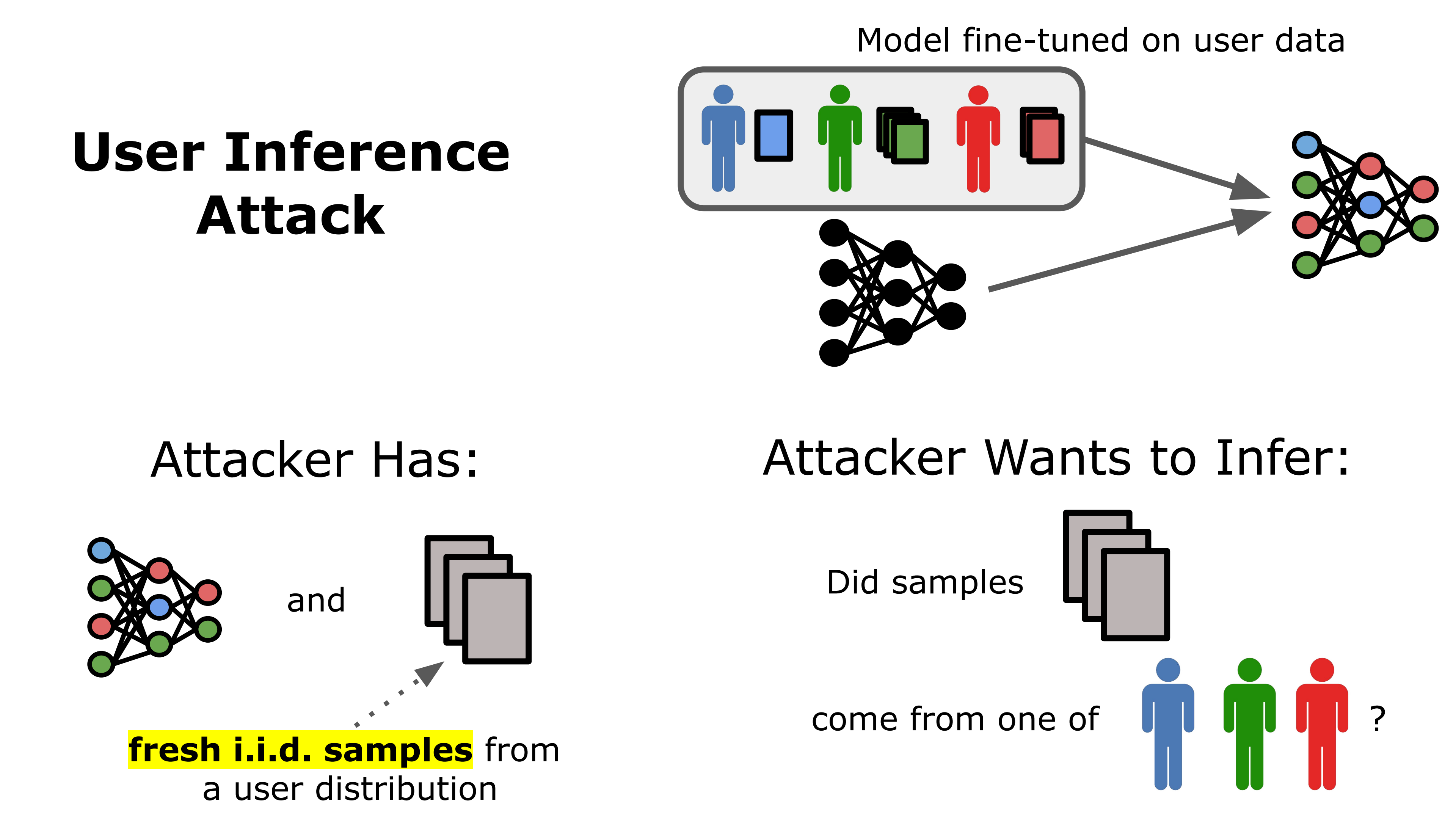
By Craig Hale published 15 July 2024

This talk:
Was ***a user's data*** used in ***fine-tuning*** LLMs?



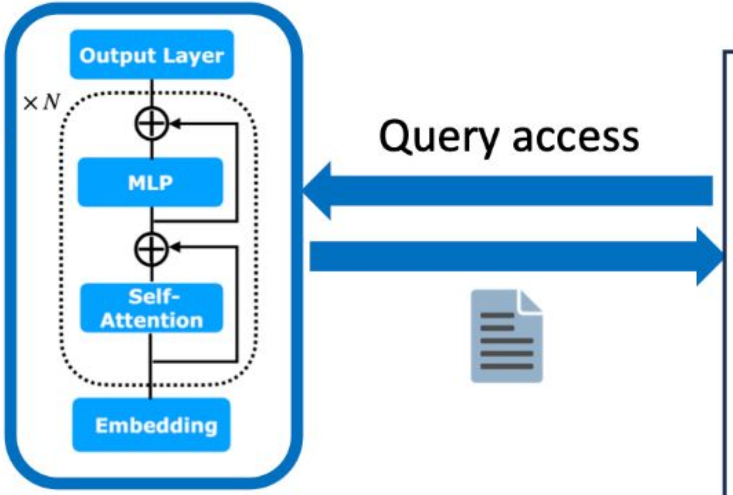
This talk:
Was ***a user's data*** used in ***fine-tuning*** LLMs?




User Inference Attack
Model fine-tuned on user data
Attacker Has:
and
fresh i.i.d. samples from a user distribution
Attacker Wants to Infer:
Did samples
come from one of
?

A simple user inference attack

Finetuned LLM p_θ



Target User U

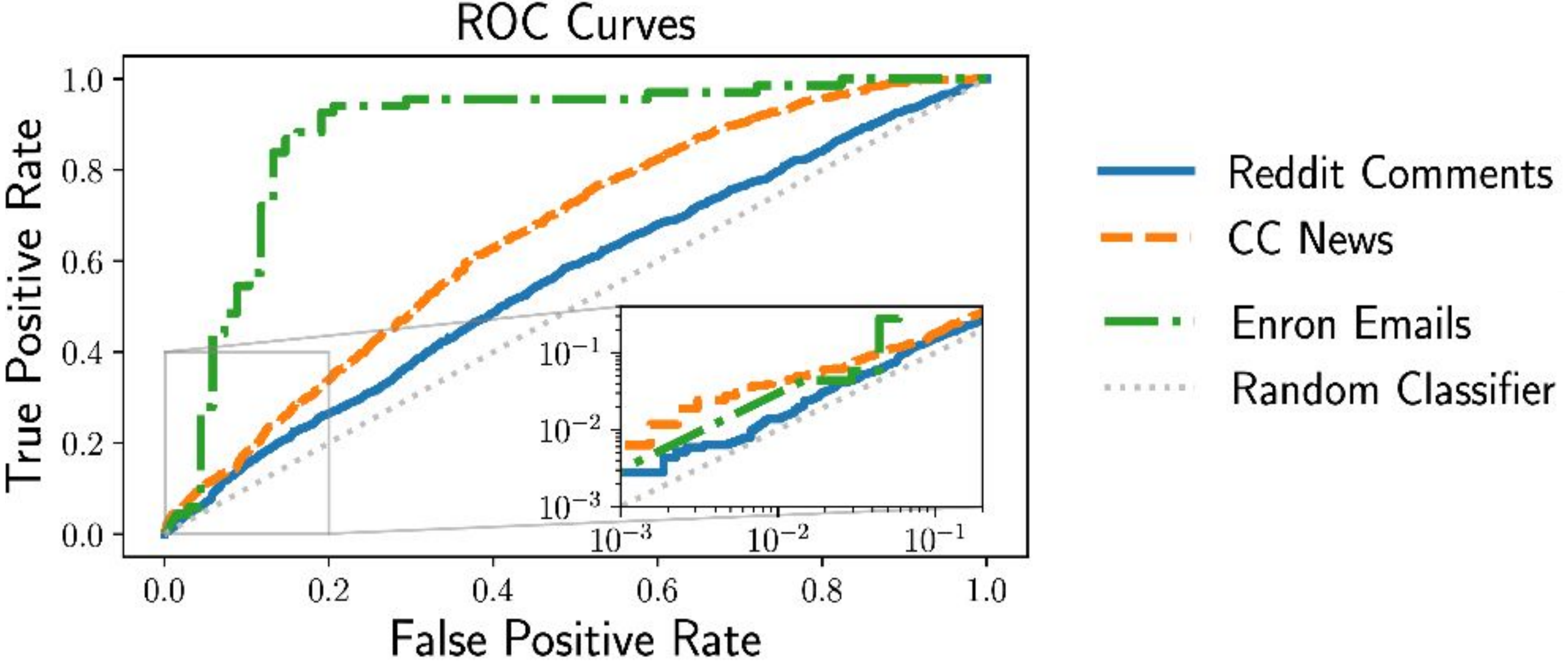
Adversary

1. Sample $x^{(1)}, \ldots, x^{(m)}$ from D_U
2. For each $x^{(i)}$ compute $p_\theta(x^{(i)})$
3. Test statistic

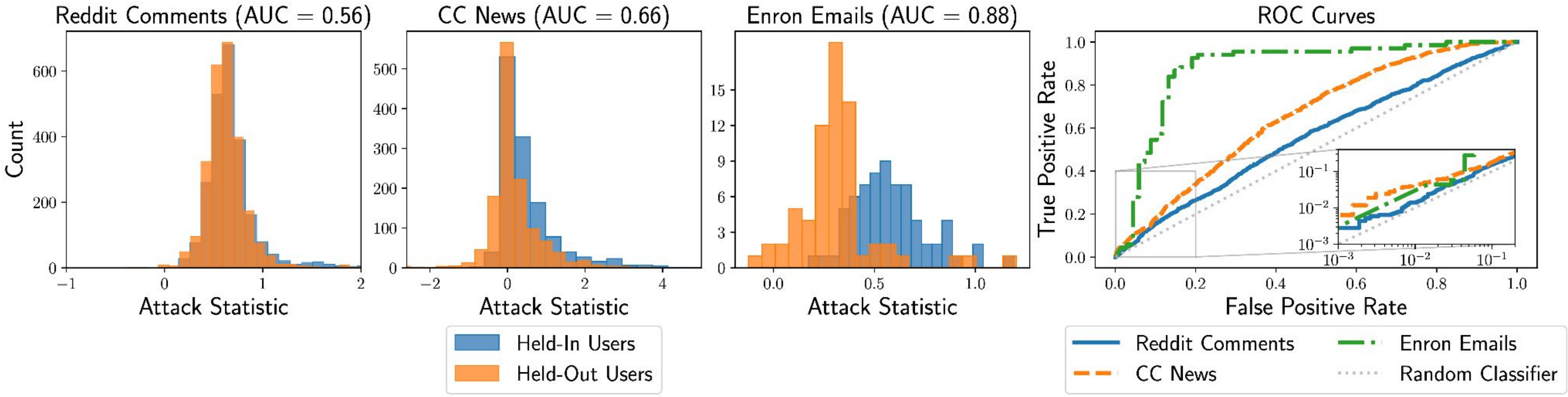
$$\hat{T}\left(x^{(1)}, \ldots, x^{(m)}\right) = \frac{1}{m}\sum_{i=1}^{m} \log\left(\frac{p_\theta(x^{(i)})}{p_{\text{ref}}(x^{(i)})}\right)$$

4. U was in training if $\hat{T}\left(x^{(1)}, \ldots, x^{(m)}\right) > \tau$

Evaluation


ROC Curves
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
10^{-1}
10^{-2}
10^{-3}
10^{-3}
10^{-2}
10^{-1}
Reddit Comments
CC News
Enron Emails
Random Classifier

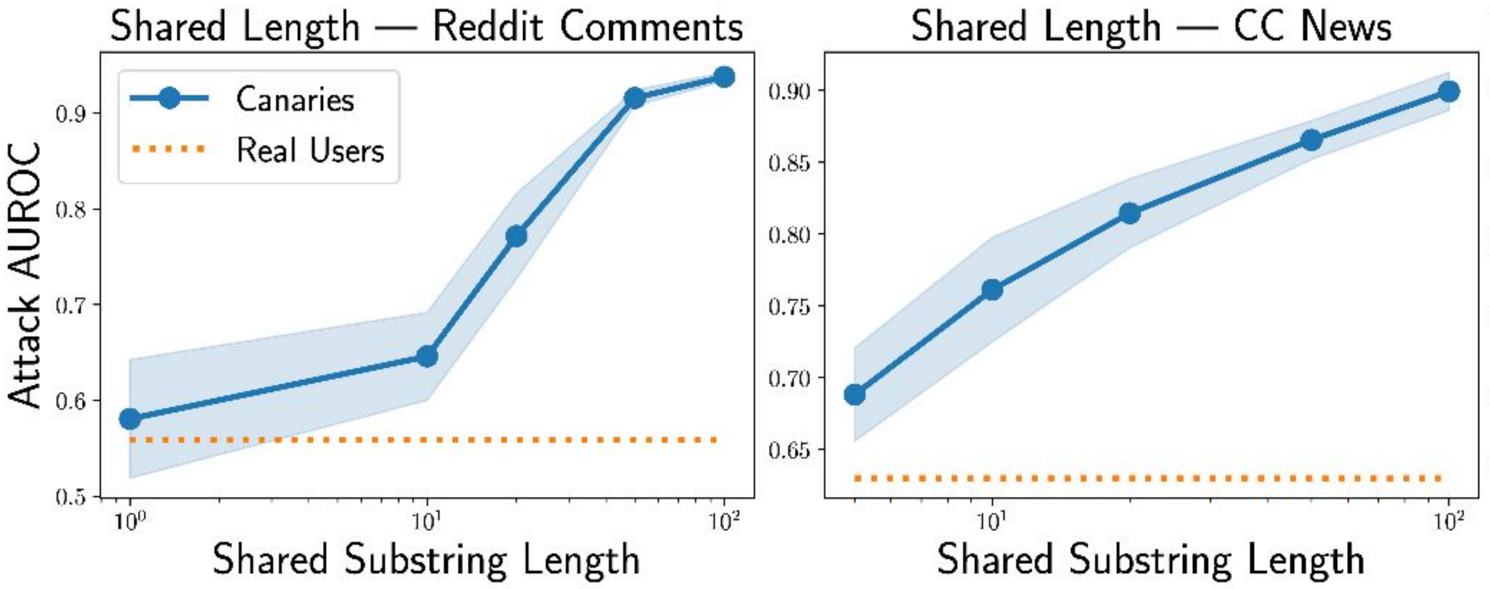
User inference is effective when #users is small and data per user is large



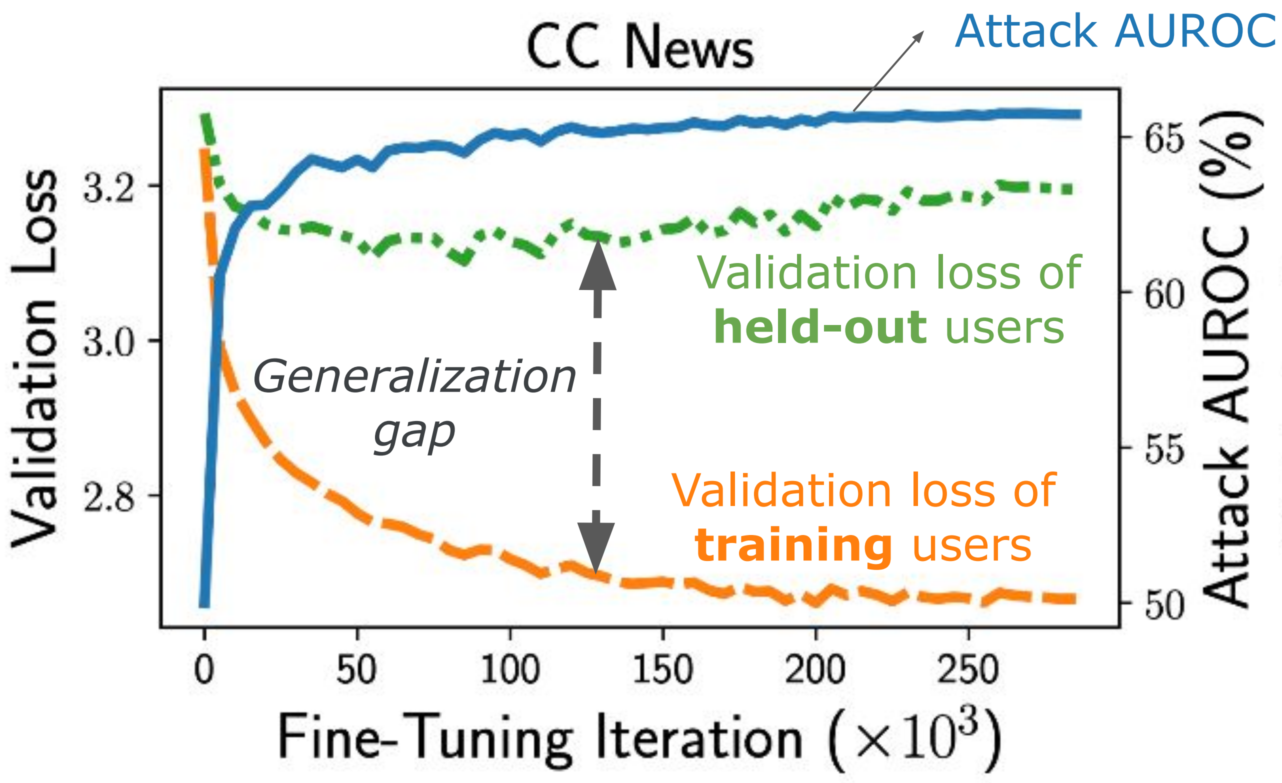
More fine-tuning samples per user

More users

Short common phrases can exacerbate user inference



User inference mechanism:
Overfitting to ***data distributions*** of training users



Spearman Correlation(Generalization gap, AUROC) = 0.995

Can user inference be mitigated?

Do not work

Early stopping

Gradient clipping

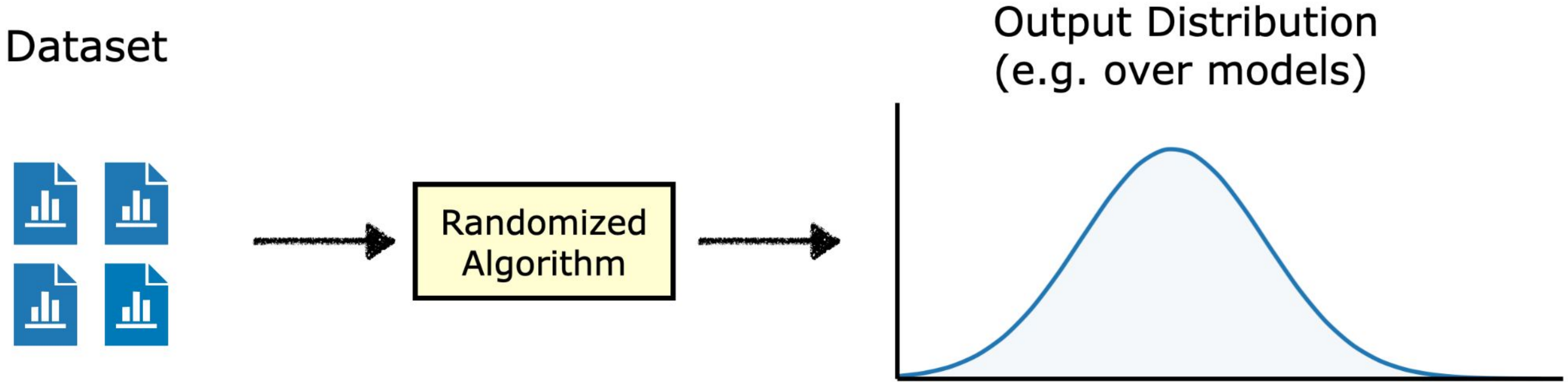
Limited Mitigation*

Data limits per user

Data deduplication

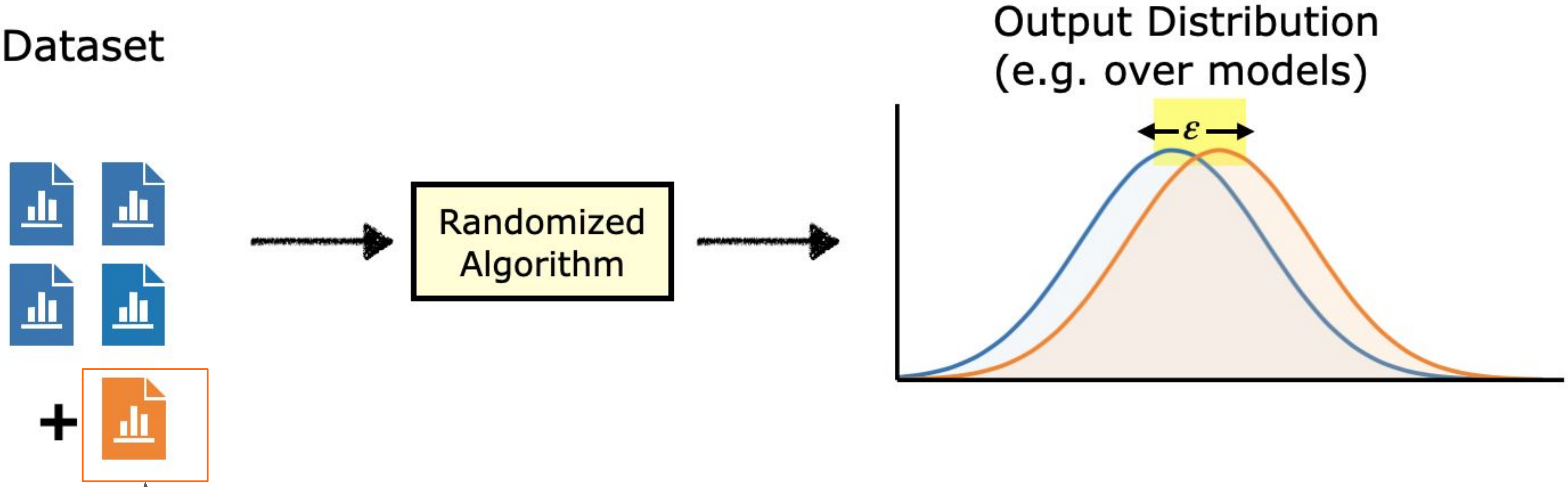
Differential privacy

Differential privacy (DP)



Dwork, McSherry, Nissim, Smith. **Calibrating noise to sensitivity in private data analysis**. TCC 2006

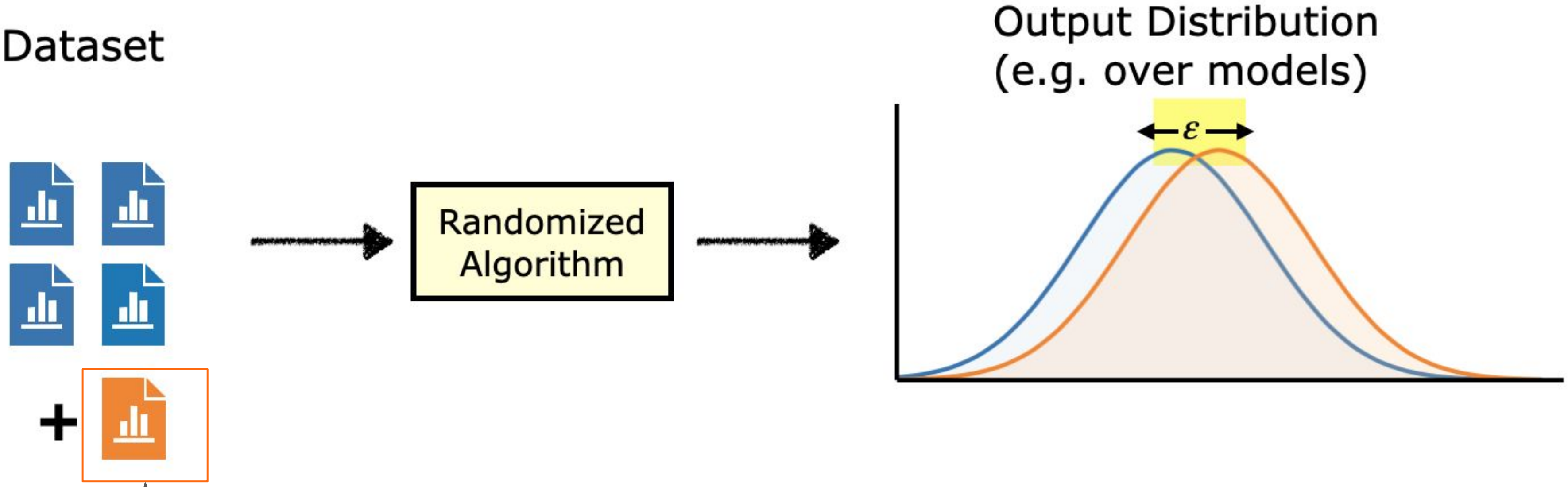
Differential privacy (DP)



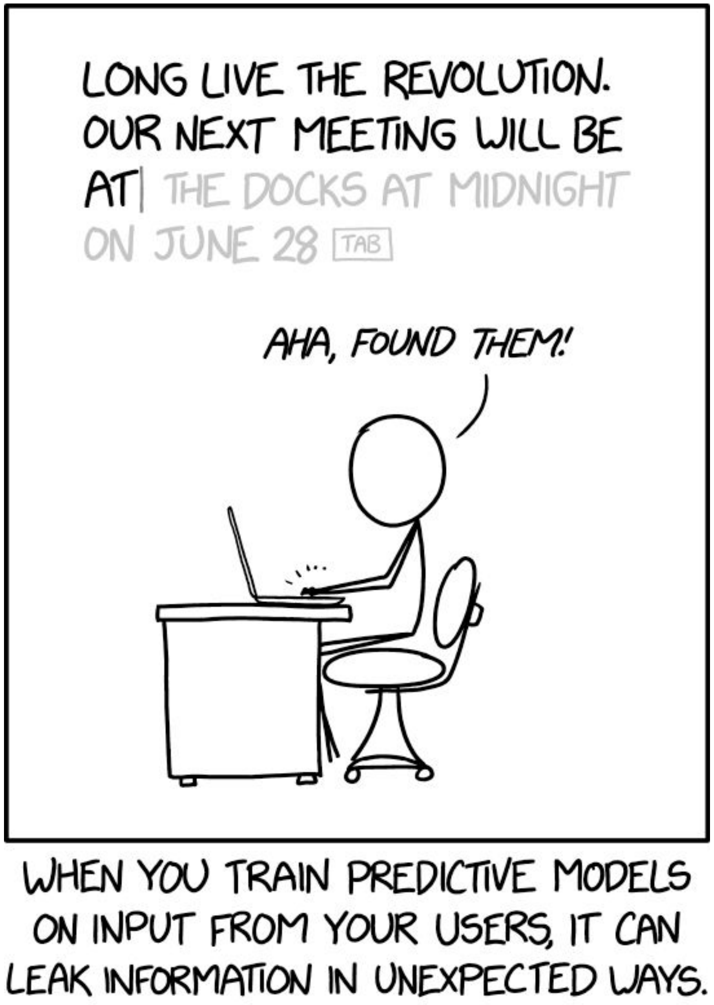
A randomized algorithm is ε-**differentially private** if the addition of **one unit of data** does not alter its output distribution by more than ε

Example-level Differential privacy (DP)

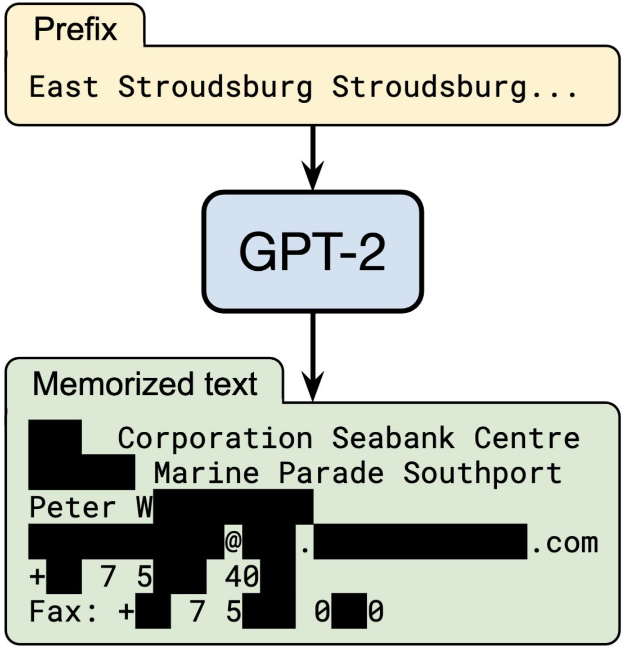
Unit of data
= ***example***



A randomized algorithm is ε-**differentially private** if the addition of **one <u>example</u>** does not alter its output distribution by more than ε

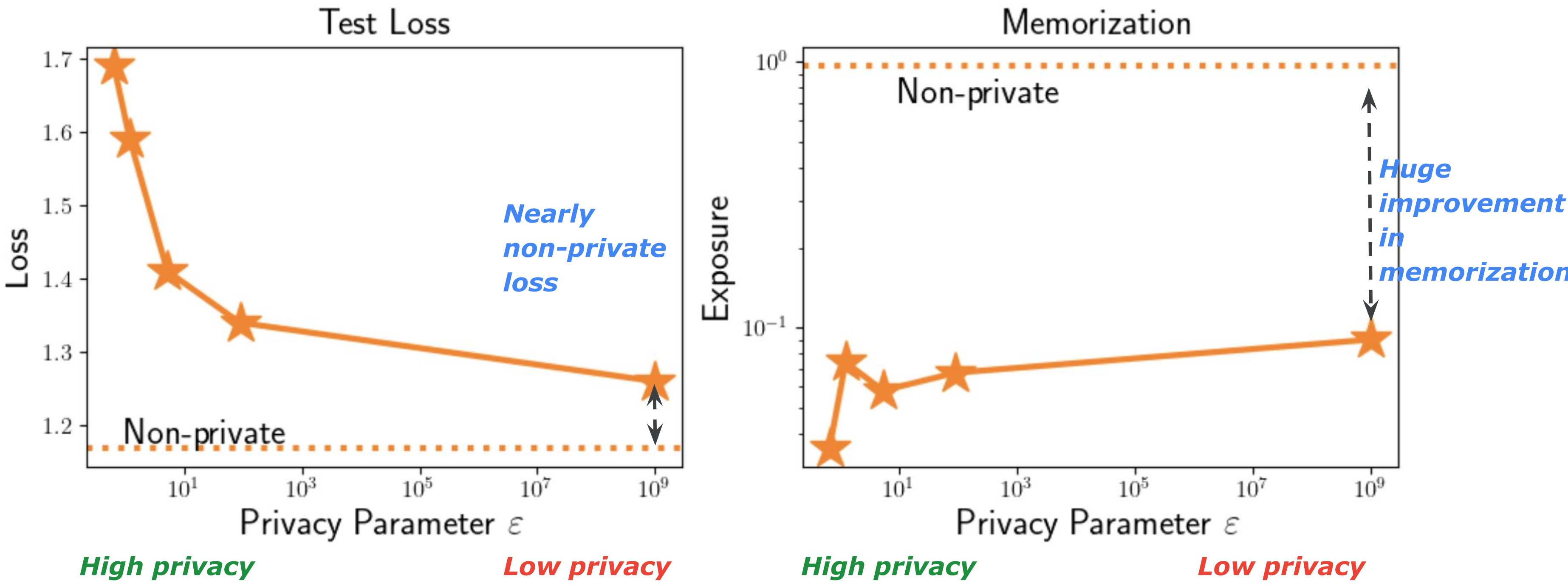


Models leak information about their training data



Carlini et al. (USENIX Security 2021)

Example-level DP eliminates memorization



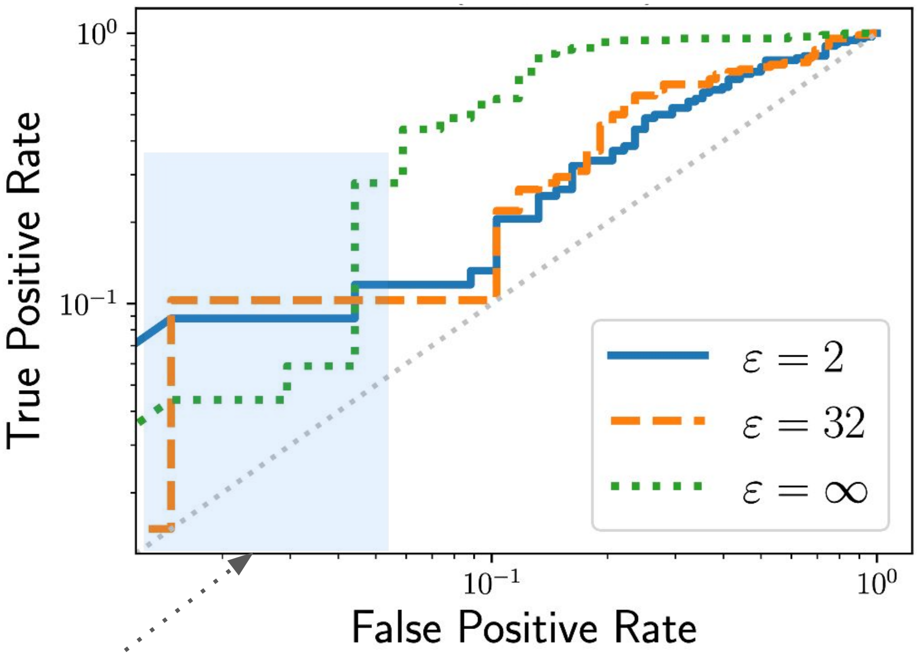
Carlini, Liu, Erlingsson, Kos, Song. **The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks.** USENIX Security 2019.

Example-level DP offers limited mitigation for user inference

AUROC:
- non-private: 88%
- ε = 32: 70%

Utility:
- DP model reaches what the private model achieves in 1/3 epoch

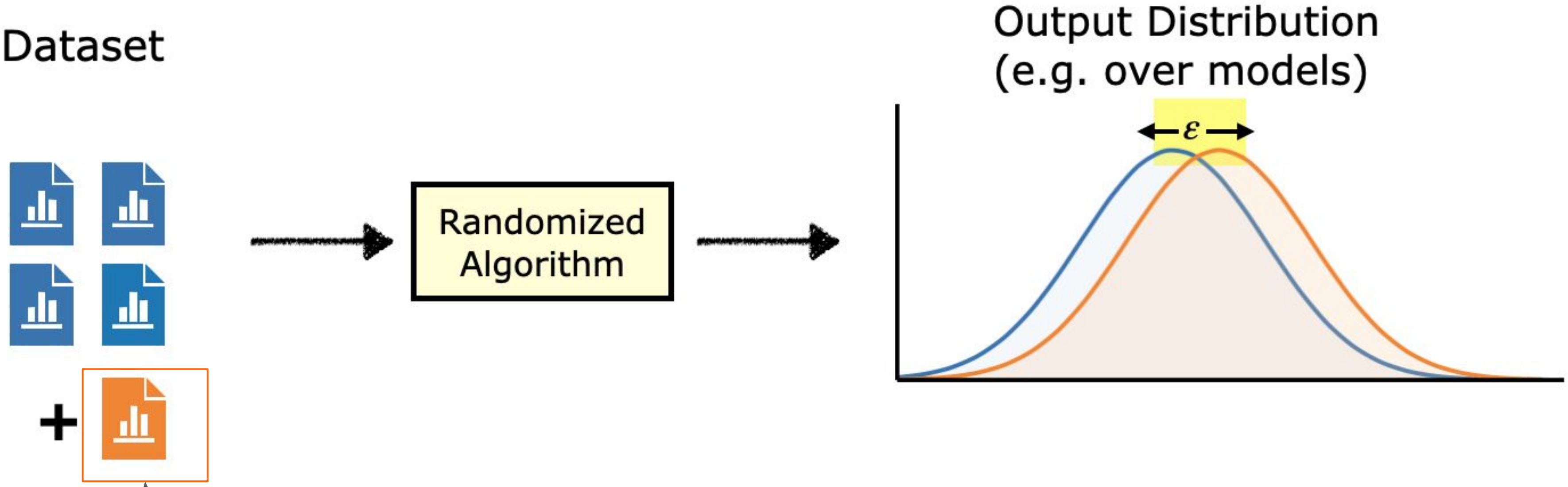
ROC Curves for Enron Emails



Example-level DP does not help here

User ~~Example~~-level Differential privacy (DP)

Unit of data = ***user***



A randomized algorithm is ε-**differentially private** if the addition of **one user's data** does not alter its output distribution by more than ε

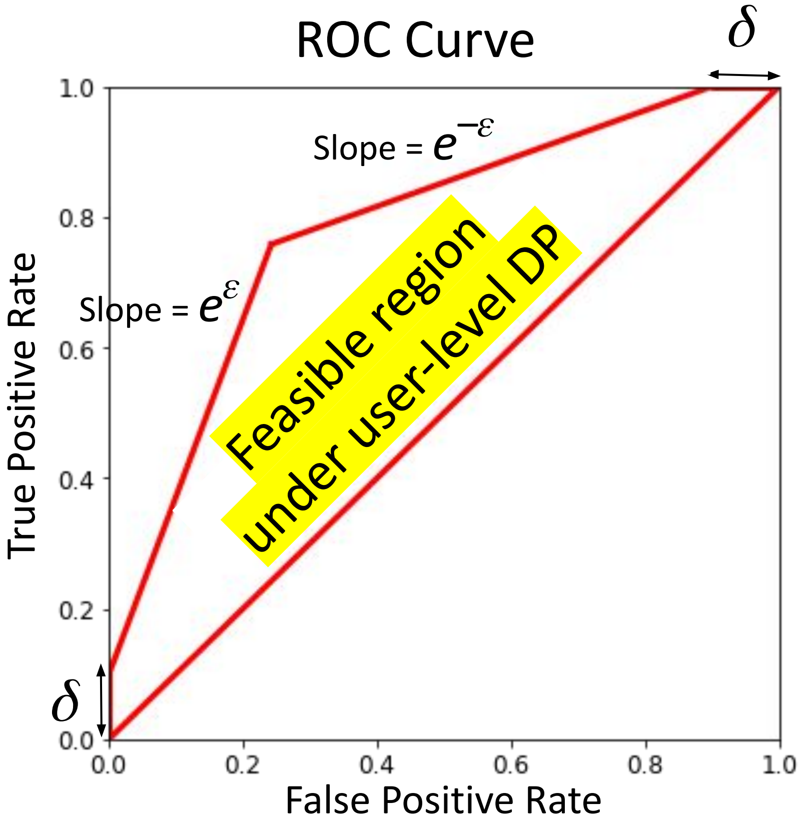
User-level DP: Provable protections against user inference

- By **differential privacy definition**:

$$\mathrm{TPR} \leq e^{\varepsilon}\,\mathrm{FPR} + \delta$$

True Positive Rate *False Positive Rate*

- Fundamental limits on the success of membership inference



Kairouz, Oh, Viswanath. **The composition theorem for differential privacy**. ICML 2015

Fine-Tuning Large Language Models with User-Level Differential Privacy

Zachary Charles
Google Research
Seattle, WA, USA
zachcharles@google.com

Arun Ganesh
Google Research
Seattle, WA, USA
arunganesh@google.com

Ryan McKenna
Google Research
Seattle, WA, USA
mckennar@google.com

H. Brendan McMahan
Google Research
Seattle, WA, USA
mcmahan@google.com

Nicole Mitchell
Google Research
San Francisco, CA, USA
nicolemitchell@google.com

Krishna Pillutla
IIT Madras
Chennai, India
krishnap@dsai.iitm.ac.in

Keith Rush
Google Research
Seattle, WA, USA
krush@google.com

Published as a conference paper at ICLR 2024

ICLR 2024

Advances in DP training

Correlated Noise Provably Beats Independent Noise for Differentially Private Learning

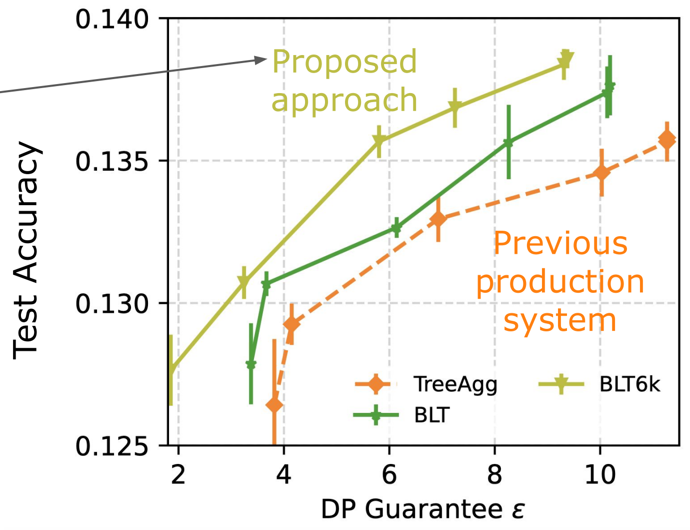
Christopher A. Choquette-Choo* Krishnamurthy (Dj) Dvijotham* Krishna Pillutla*
Arun Ganesh Thomas Steinke Abhradeep Guha Thakurta

Efficient and Near-Optimal Noise Generation for Streaming Differential Privacy*

Krishnamurthy (Dj) Dvijotham, *Google DeepMind*, dvijothamcs@gmail.com
H. Brendan McMahan, *Google Research*, mcmahan@google.com
Krishna Pillutla, *IIT Madras*, krishnap@dsai.iitm.ac.in
Thomas Steinke, *Google DeepMind*, steinke@google.com
Abhradeep Thakurta, *Google DeepMind*, athakurta@google.com

FOCS 2024

Google's production LM (Portuguese language) for next-word prediction



Plot: McMahan, Xu, Zhang (2024).

Thank you!

User Inference Attacks on Large Language Models.
EMNLP 2024 (**Oral Presentation**)

Nikhil Kandpal
U. Toronto

Alina Oprea
Northeastern U.

Peter Kairouz
Google

Chris Choquette
Google

Zheng Xu
Google