BigData 2025 Tutorial, Macau, December 2025

Data Privacy in the Age of AI and Generative LLMs: Attacks, Defenses, and Emerging Challenges

Ruixuan Liu, Li Xiong

Department of Computer Science

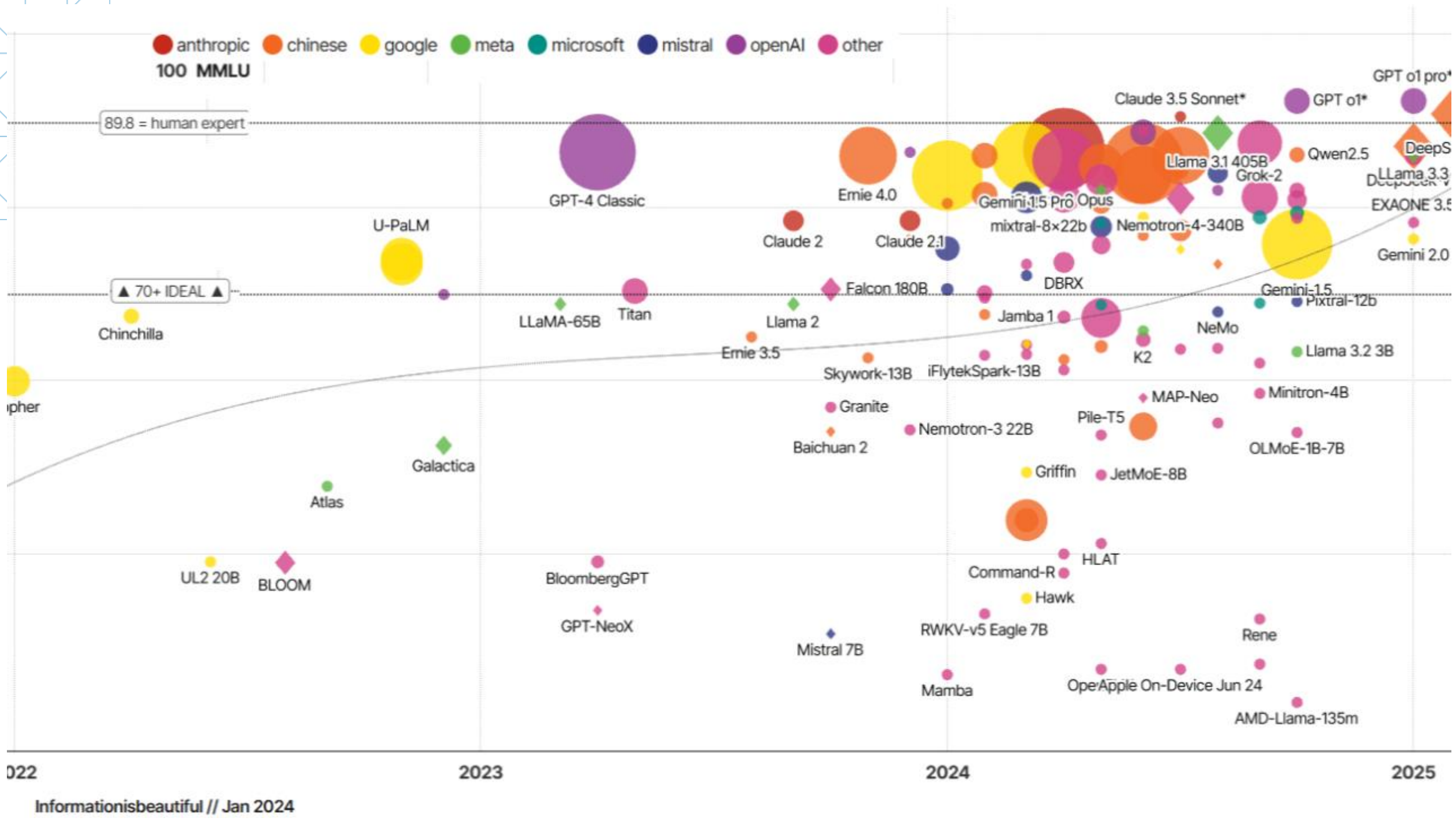
EMORY UNIVERSITY

The Bigger the Data, the Smarter the AI

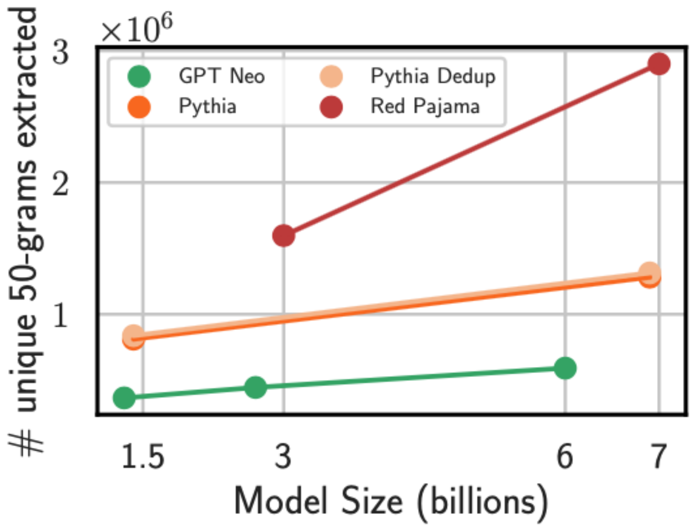
EMORY
UNIVERSITY

The Bigger/Smarter the AI, the More Data It Memorizes

- The scaling model sizes and its increasing memorization capability/privacy risk



https://www.linkedin.com/pulse/llm-parameters-explained-tushar-jain-l7vuc



Nasr, Milad, et al. "Scalable Extraction of Training Data from (Production) Language Models." ICLR 2025

Privacy in the Age of AI and LLMs: Outline

- Privacy Attacks
- Privacy Defenses
- Open challenges

EMORY UNIVERSITY

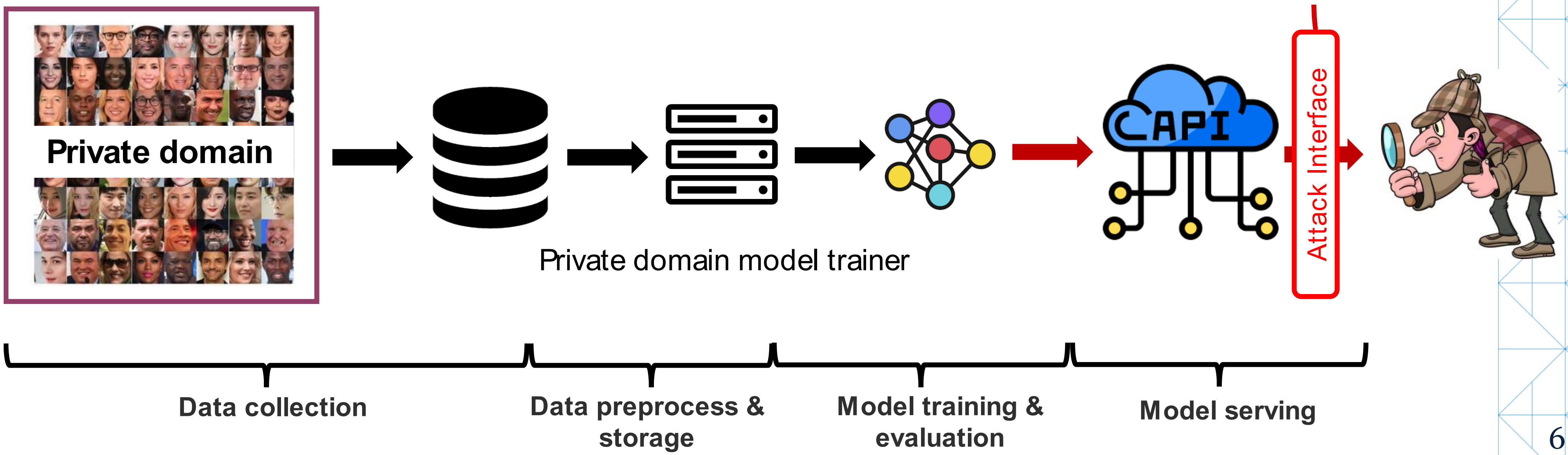
Privacy in the Age of AI and LLMs: Outline

- Privacy Attacks
  - **Overview**
  - Membership inference attack (MIA)
  - Attribute inference attack
  - Data extraction attacks
  - Backdoor attacks
  - Case studies in healthcare
- Privacy Defenses
- Open challenges

EMORY UNIVERSITY

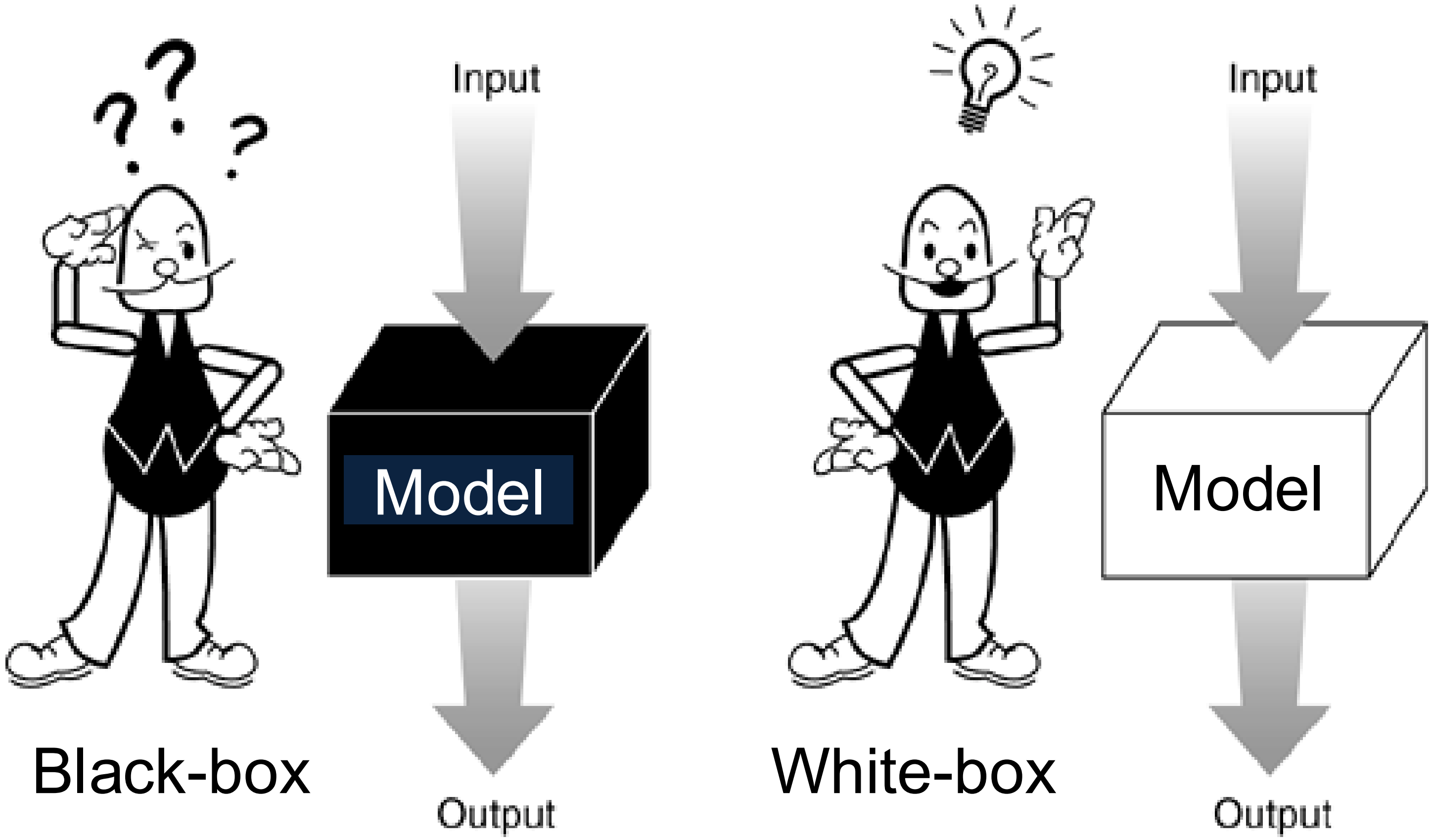
Memorization-Based Privacy Risks

- Membership inference attack (MIA)
- Attribute inference attack
- Data extraction attack (generative LLMs)



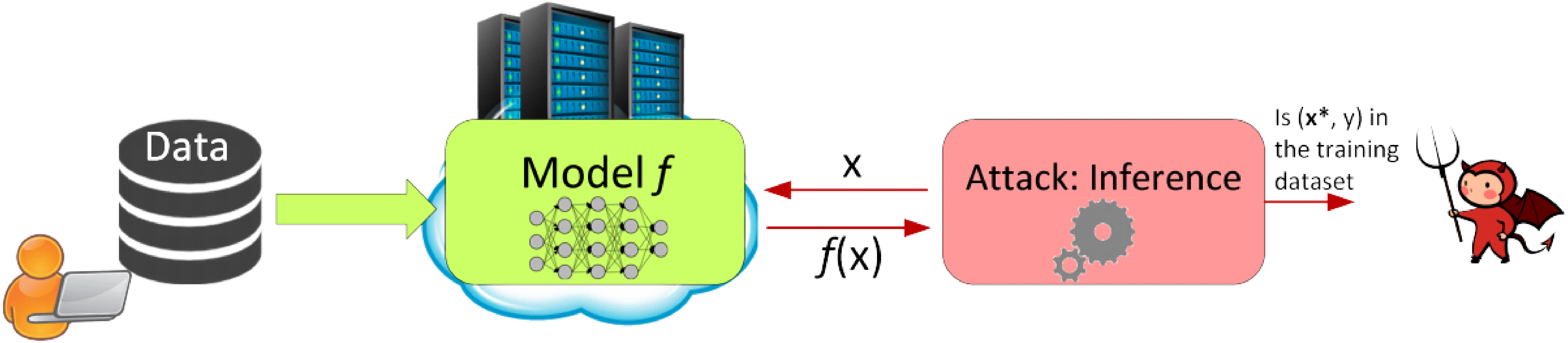
Threat Model

- **Adversarial capability**
  - Black-box: query the model and get response (loss values, probabilities, generative output, etc.)
  - White-box: know the model parameters



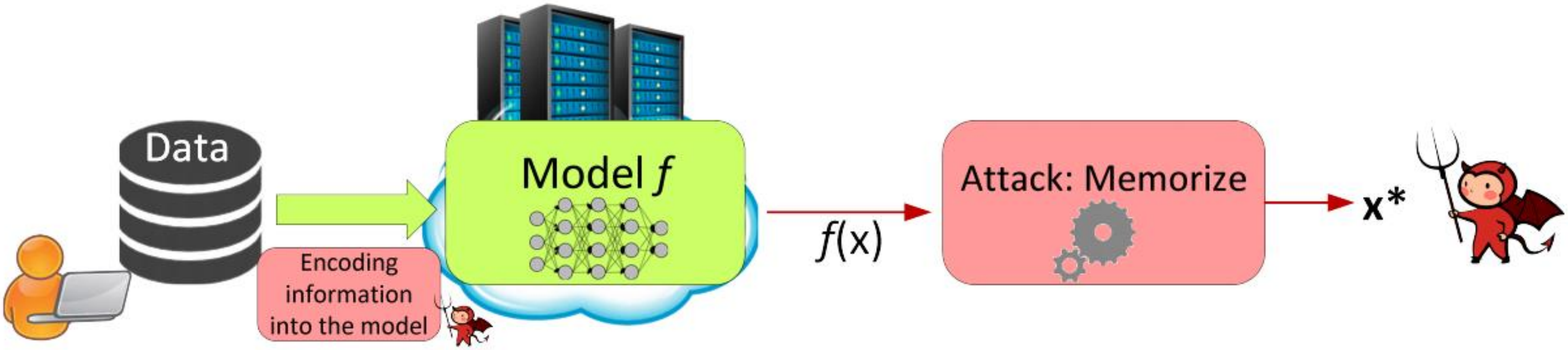
Adversarial Goal: Membership Inference Attack

- **Adversarial goal**
  - Membership Inference: Adversary learns whether a given data record is part of the model's training dataset or not



Adversarial Goal: Attribute Inference /Data Extraction

- **Adversarial goal**
  - Attribute Inference/Data extraction: Adversary recovers exact feature/values x*



Privacy in the Age of AI and LLMs: Outline

- Privacy Attacks
  - Overview
  - **Membership inference attack (MIA)**
  - Attribute inference attack
  - Data extraction attacks
  - Backdoor attacks
  - Case studies in healthcare
- Privacy Defenses
- Open challenges

EMORY UNIVERSITY

Membership Inference Attack (MIA)

- **Overview**
  - Why care about MIA?
  - Why MIA is possible?
  - How to measure the MIA effectiveness?

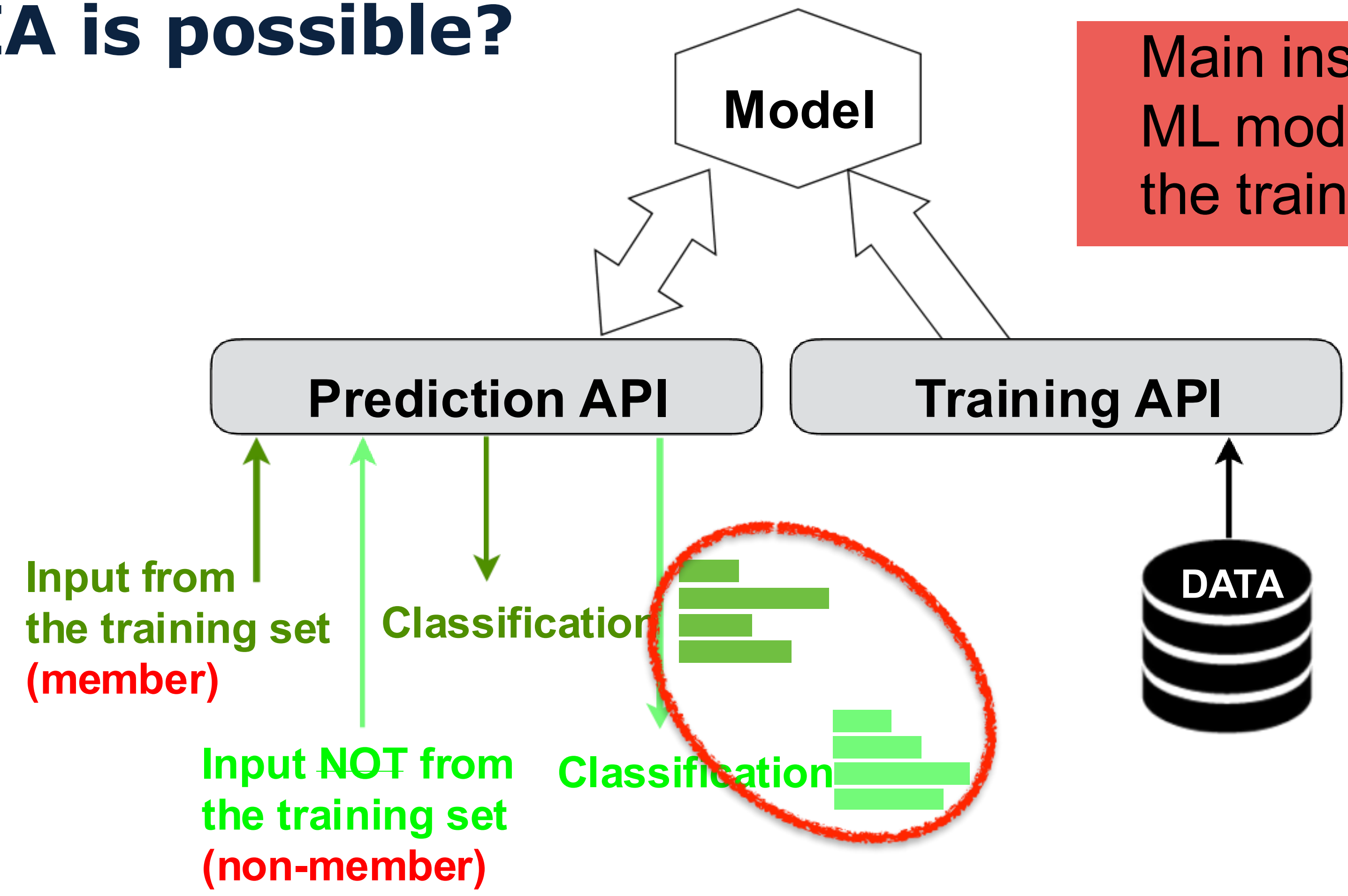
- **Methods**
  - Classifier-based MIA
  - Threshold-based MIA
  - MIA as Hypothesis Test
  - MIA on LLMs

Membership Inference Attack (MIA)

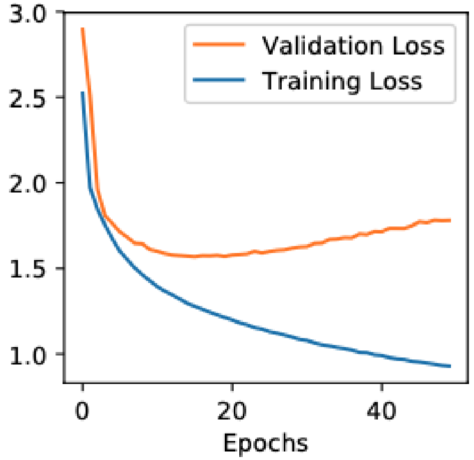
- **Why care about MIA?**
  - Curiosity
    - What did some company train their model on?
    - Did anyone use my data?
  - As the building block for other attacks
    - Attribute inference
    - Data extraction
  - Audit privacy risk
    - Is the privacy training algorithm conducted correctly?

Membership Inference Attack (MIA)

- **Why MIA is possible?**



Main insight:
ML models overfit to the training data

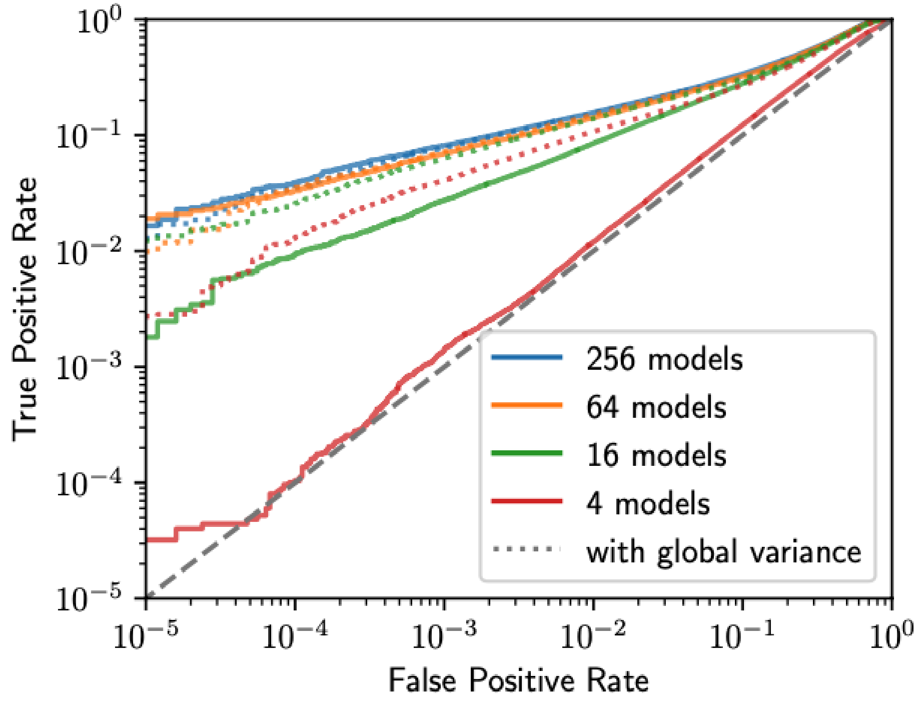


Recognize the difference

Slide source: Membership Inference Attacks against Machine Learning Models, IEEE S&P 2017

Membership Inference Attack (MIA)

- **How to measure the MIA effectiveness?**
  - MIA ➔ A binary classification task for the attacker
    - Input: a sample
    - Output: 1 or 0 (member / non-mem)
  - Attacker has a query dataset ➔ Test dataset for MIA
    - 50% member samples, 50% non-member samples
  - **Evaluation Metrics**
    - Typical metrics for binary prediction: Accuracy / Precision / Recall
    - <u>Recommended</u>:
      - Receiver Operating Characteristic (ROC) curve (log scale) and its AUC
      - true-positive rate (TPR) given a low false-positive rate (FPR)



Membership Inference Attack (MIA)

- **Overview**
  - Why care about MIA?
  - Why MIA is possible?
  - How to measure the MIA effectiveness?

- **Methods**
  - Classifier-based MIA
  - Threshold-based MIA
  - MIA as Hypothesis Test
  - MIA on LLMs

Membership Inference Attacks against Machine Learning Models

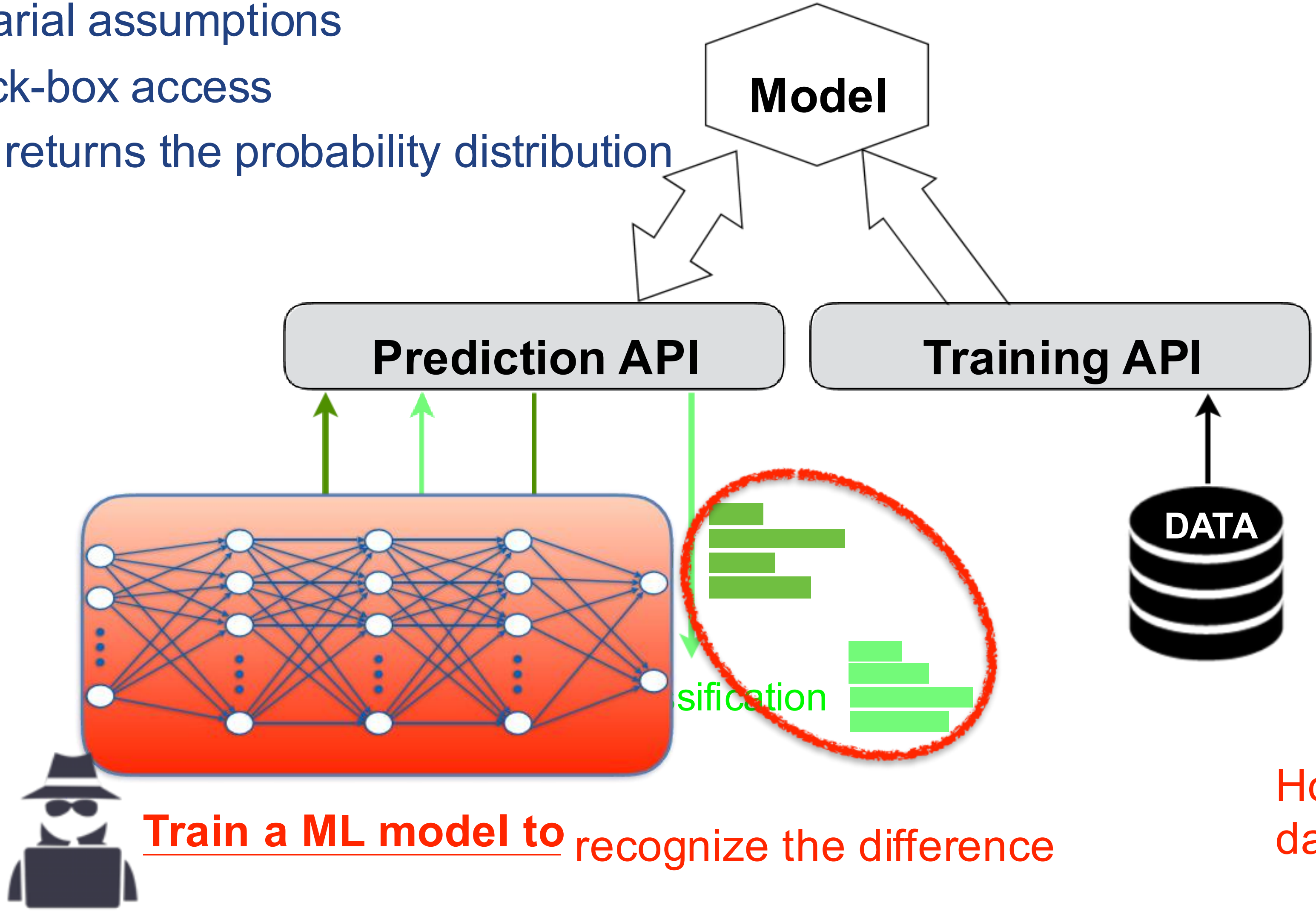
Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov

CORNELL TECH

Slide source: Membership Inference Attacks against Machine Learning Models, IEEE S&P 2017

Overview

- Classifier-based MIA: train a binary attack model by observing predictions of shadow models
- Adversarial assumptions
  - Black-box access
  - API returns the probability distribution

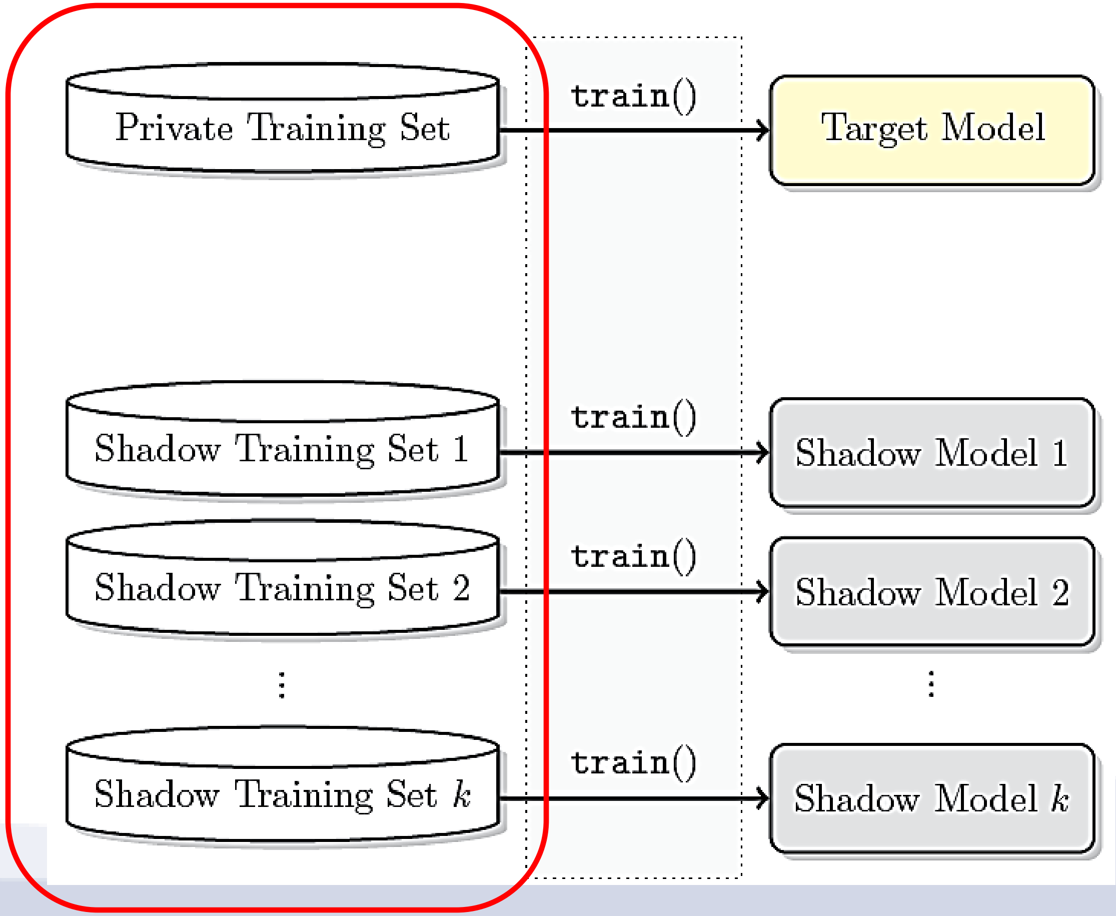


How to get the training data for the attack model?

Shadow Model Training (Step1)

- Goal of shadow model
  - substitute the target model
  - collect different behaviors of mem/non-mem samples

- Each shadow model is trained on a dataset that has a similar distribution as the private training dataset of the target model



How to build training dataset for shadow models?

Shadow Model Training (Step1)

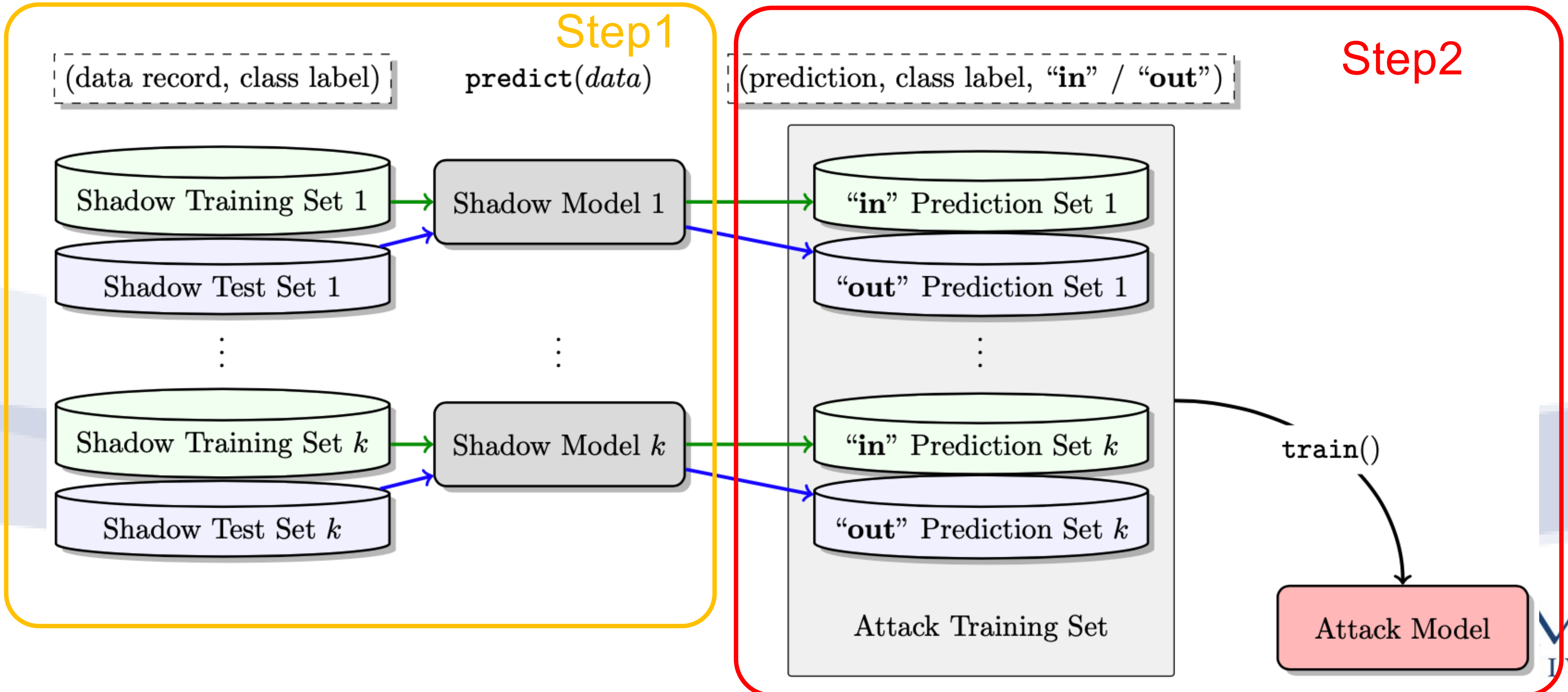
- **How to build training dataset for shadow models?**
  - **Case I: prior knowledge** on the data distribution
    - Randomly draw samples from the distribution
  - **Case II: no prior knowledge** on the data distribution
    - Start with a random sample
    - Query the target model multiple times
    - Change each input feature until the modified samples are classified with high confidence

Real data

Synthetic data

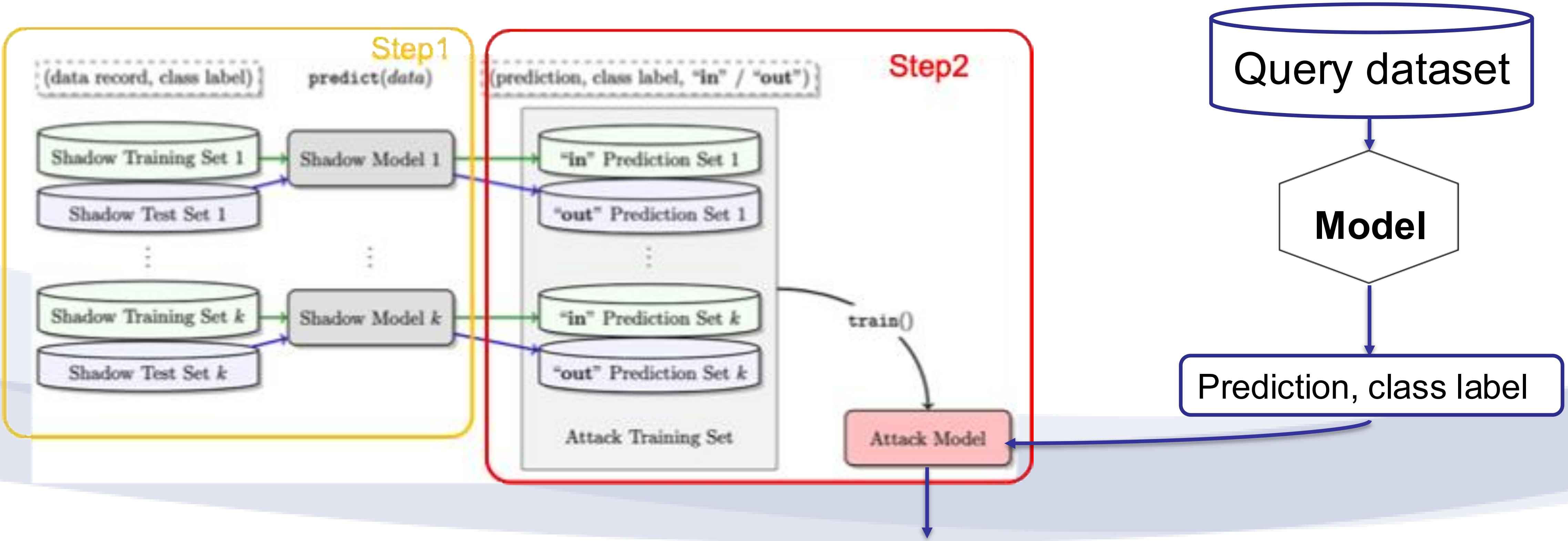
Attack Model Training (Step2)

- The output probability vectors from the shadow models are used as inputs for training attack models
- Whether the data is used as training data for the shadow model is used as label for training attack models (as binary classifiers)

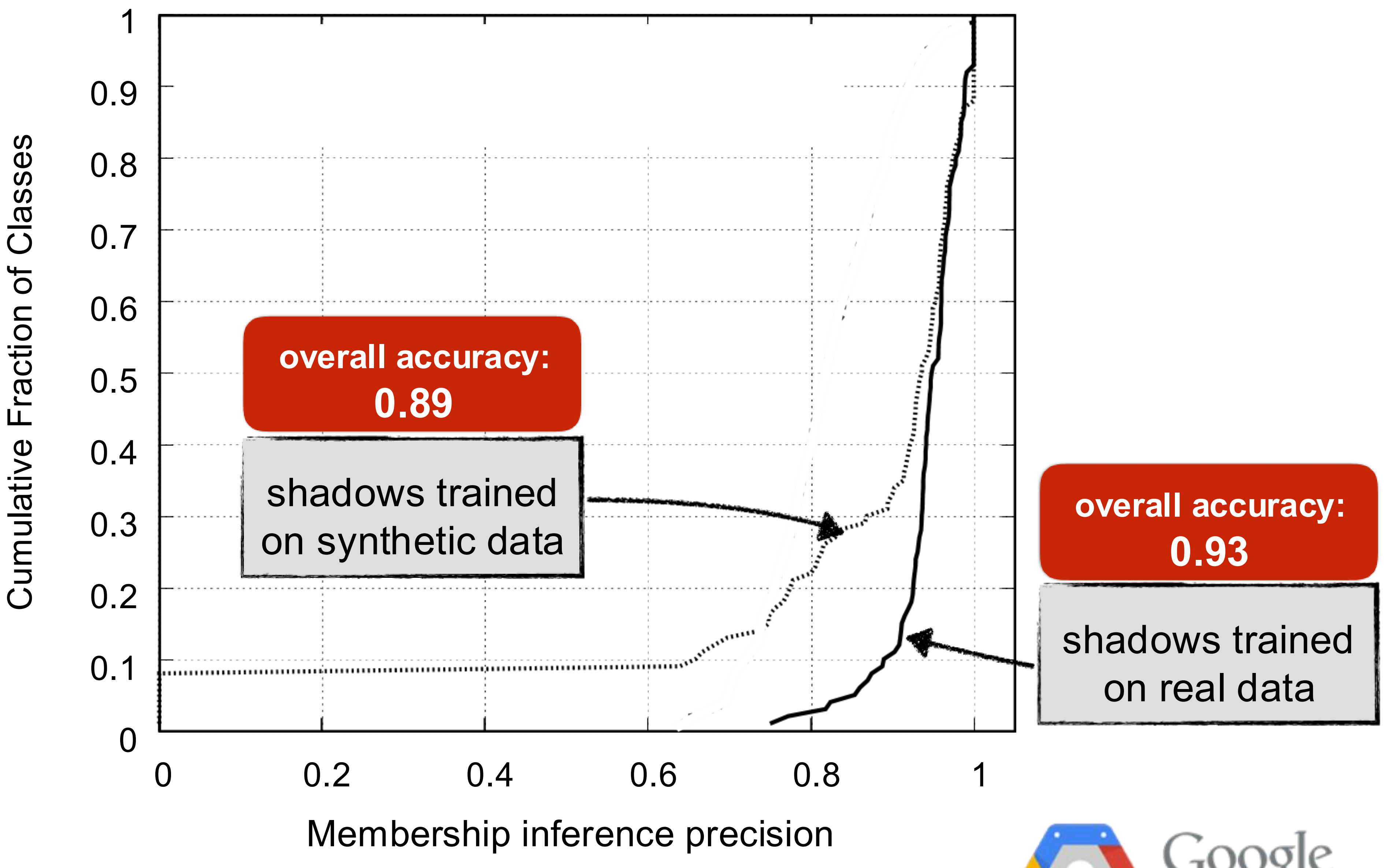


Attack Model Training (Step3)

- Query the attack model and get the predicted membership for a given sample



$(x, y) \in D_{tr}$ or $(x, y) \notin D_{tr}$



Purchase Dataset — Classify Customers (100 classes)

Google Cloud Platform

Performance of Shadow Model Attack

- The larger the overfitting (difference between the training and testing accuracy), the more successful the membership inference attack is
- Overfitting
  - not only reduces the generalization of a model
  - but also makes the model more likely to leak sensitive information about the training data

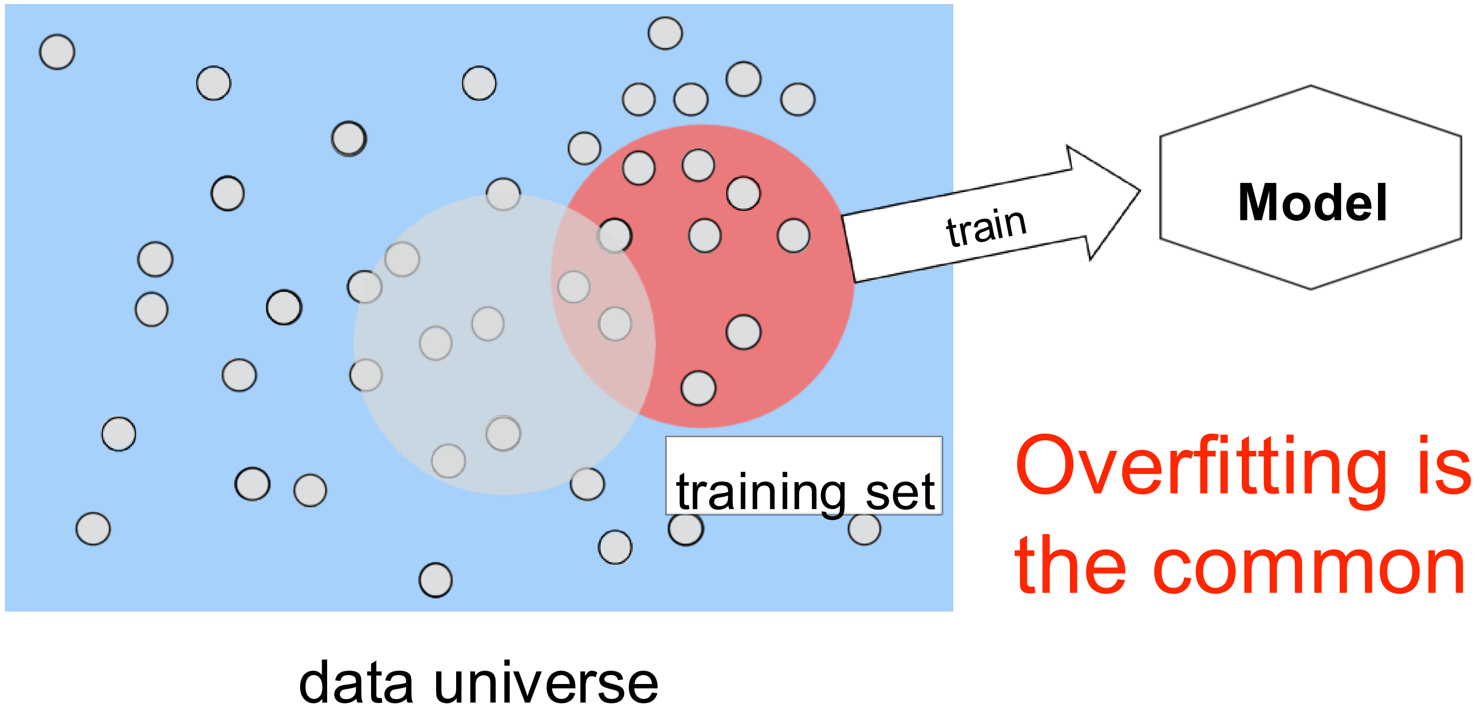
Dataset	*Training Accuracy*	*Testing Accuracy*	*Attack Precision*
Adult	0.848	0.842	0.503
MNIST	0.984	0.928	0.517
Location	1.000	0.673	0.678
Purchase (2)	0.999	0.984	0.505
Purchase (10)	0.999	0.866	0.550
Purchase (20)	1.000	0.781	0.590
Purchase (50)	1.000	0.693	0.860
Purchase (100)	0.999	0.659	0.935
TX hospital stays	0.668	0.517	0.657

Privacy

Does the model leak information about data in the training set?

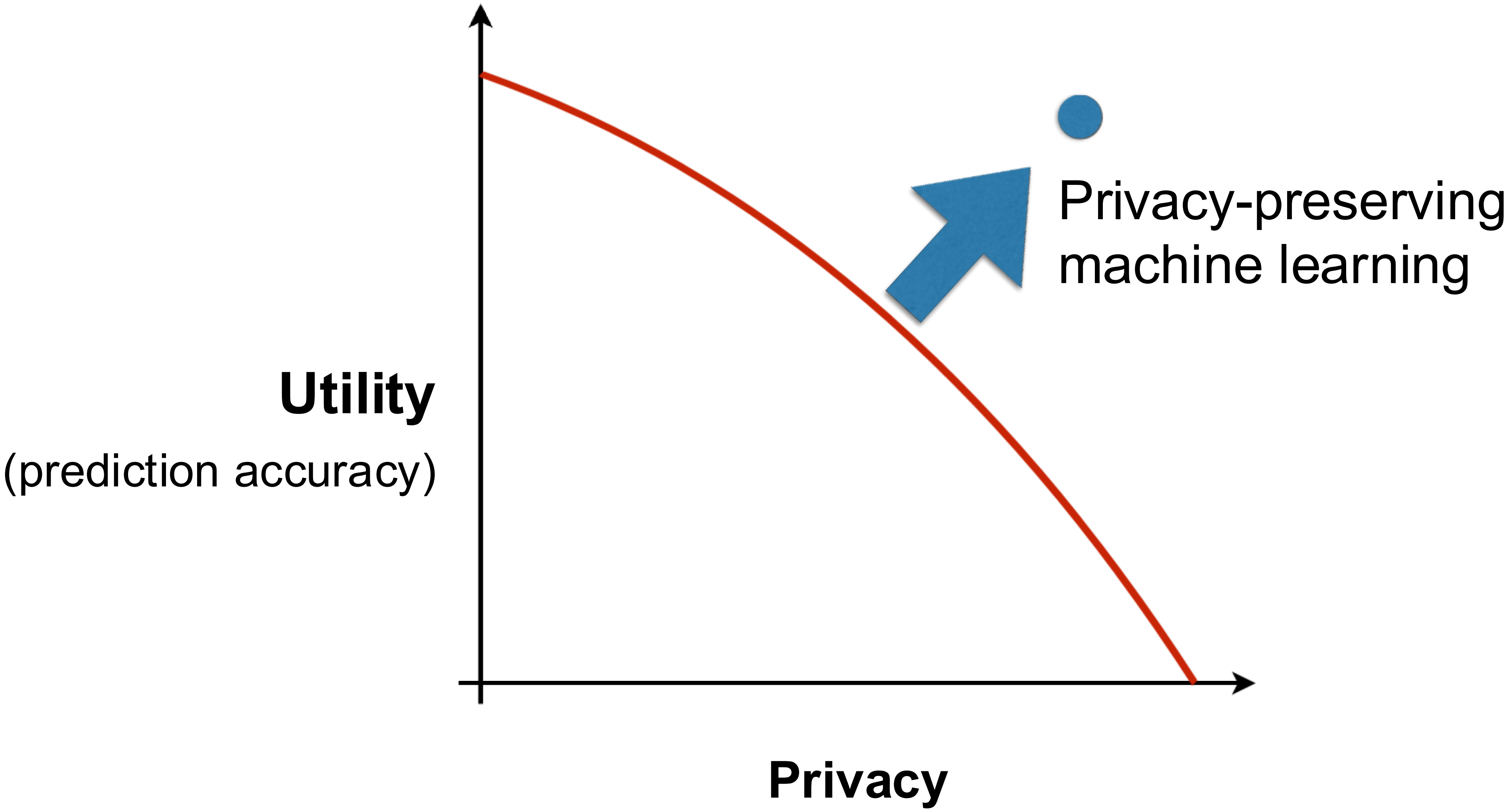
Learning

Does the model generalize to data outside the training set?



Overfitting is the common enemy!

Not in a Direct Conflict!

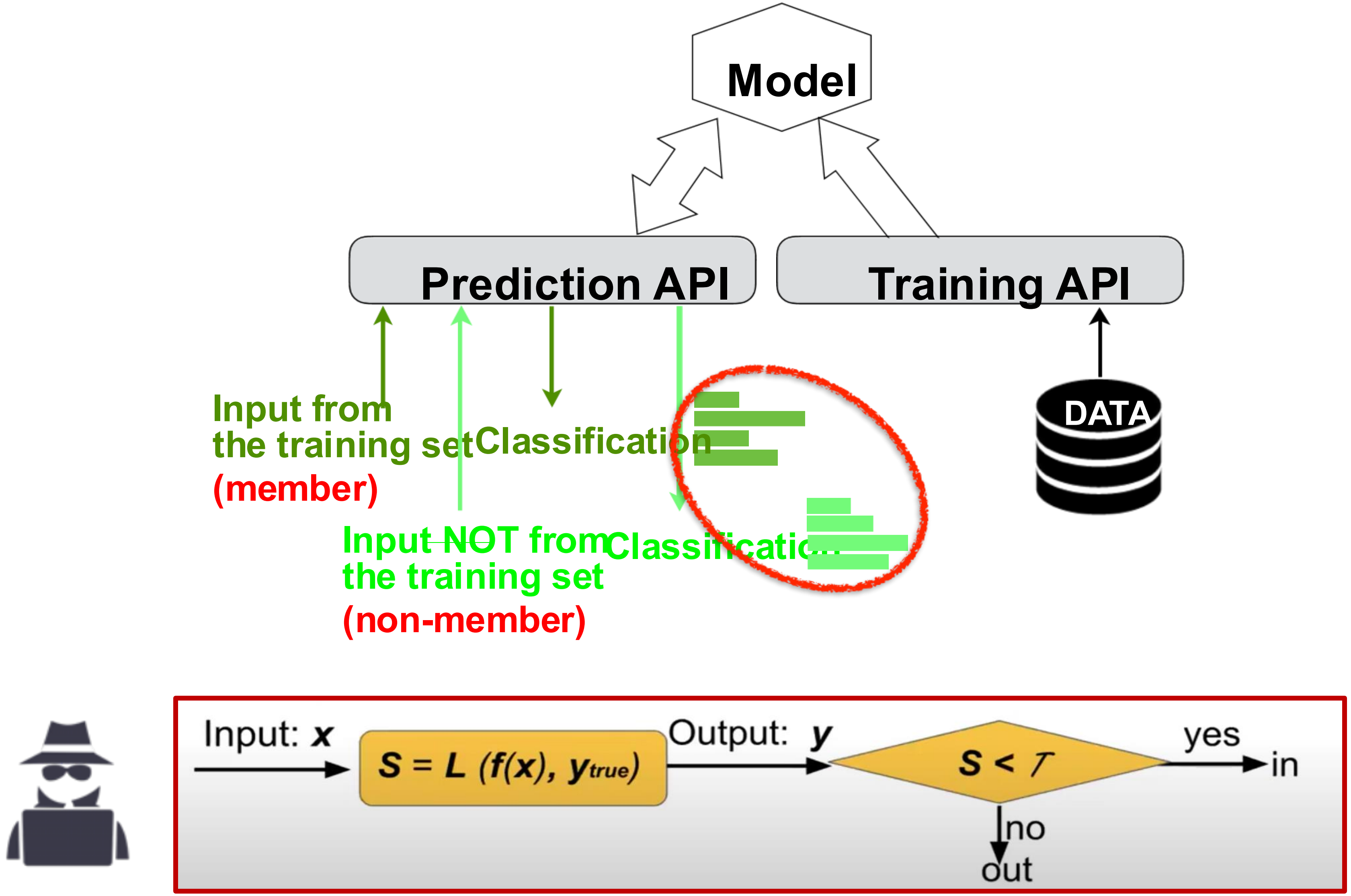


Membership Inference Attack (MIA)

- **Overview**
  - Why care about MIA?
  - Why MIA is possible?
  - How to measure the MIA effectiveness?

- **Methods**
  - Classifier-based MIA
  - **Threshold-based MIA**
  - MIA as Hypothesis Test

Membership Inference Attack (MIA): Threshold Based

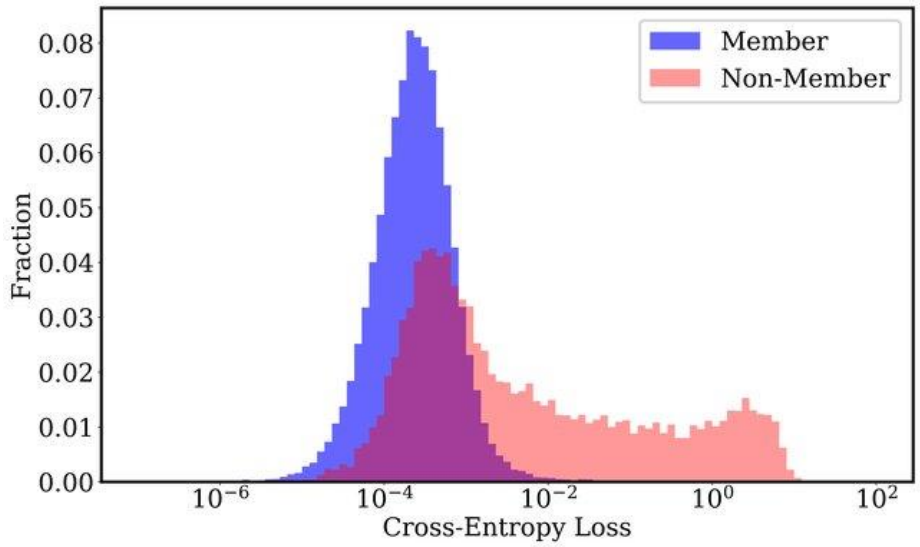


Threshold-based MIA: Loss attack

- Membership signal: loss [Yeom et. al., 2017]

$$A_{f_\theta}(x) = \mathbb{1}[\mathcal{L}(f_\theta, x) < \gamma]. \quad (1)$$

- Intuition: member samples -> lower loss
- Threshold can be estimated as the average of loss values over a data collection drawn from the distribution
- Limitation: Very good at predicting non-member, but weak in predicting members
  - Need to differentiate easy samples and memorized samples



Yeom et. al., 2017: Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting

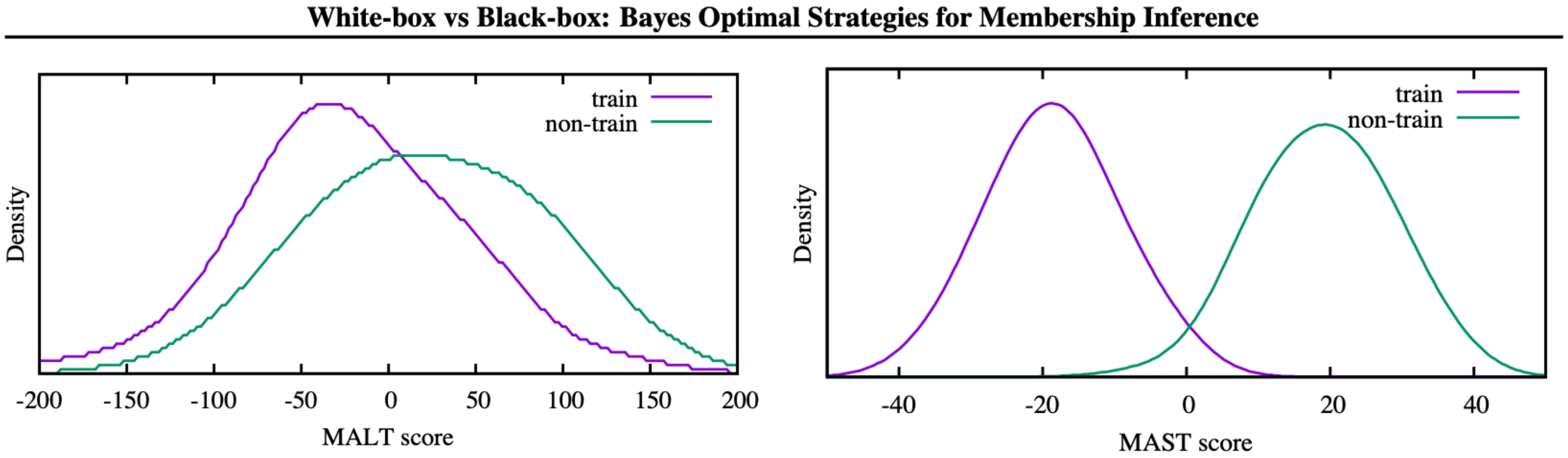
Threshold-based MIA: loss calibrated by reference model

- Membership signal: loss relative to the loss of a reference (non-member) model [Ye et. al., 2022] [Sablayrolles et. al., 2019]
  - Reduce the impact of sample difficulty
  - Member samples -> lower calibrated loss

$$A_{f_\theta}(x) = \mathbb{1}[\mathcal{L}(f_\theta, x) < \gamma]. \quad (1)$$

$$A_{f_\theta}(x) = \mathbb{1}[\mathcal{L}(f_\theta, x) - d(x) < \gamma]. \quad (2)$$

$$d(x) = \mathcal{L}(f_\phi, x)$$

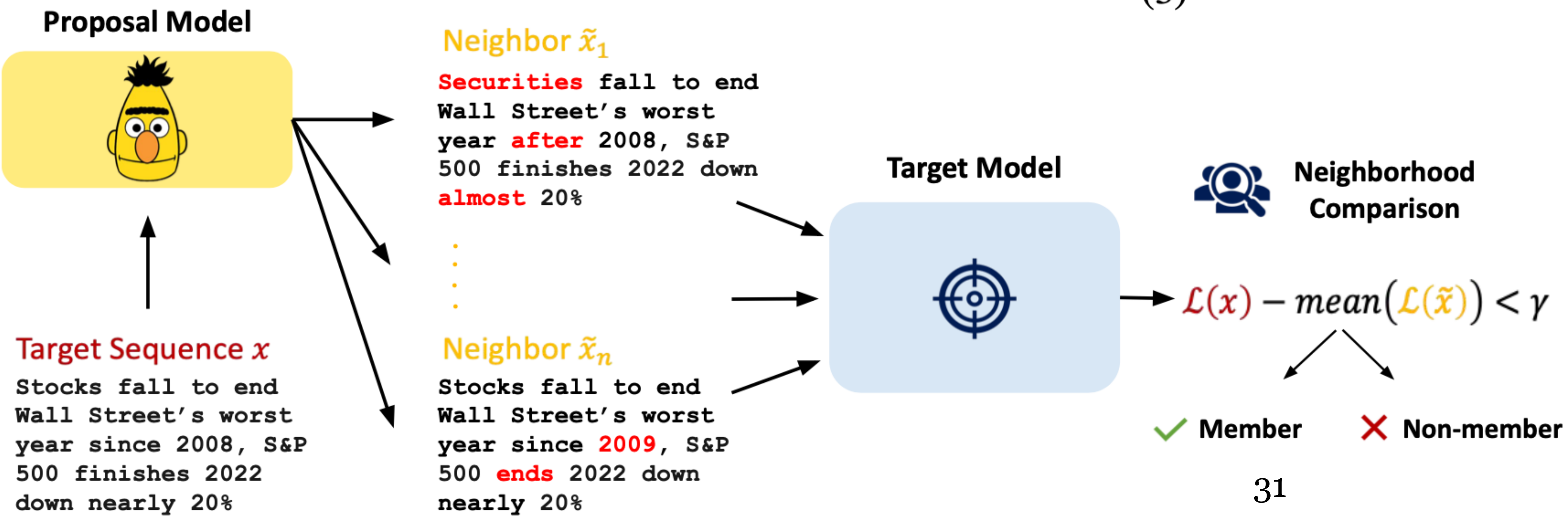


[Ye et. al., 2022]: Enhanced Membership Inference Attacks against Machine Learning Models
[Sablayrolles et al. 2019]: White-box vs Black-box: Bayes Optimal Strategies for Membership Inference

Threshold-based MIA: loss calibrated by neighborhood

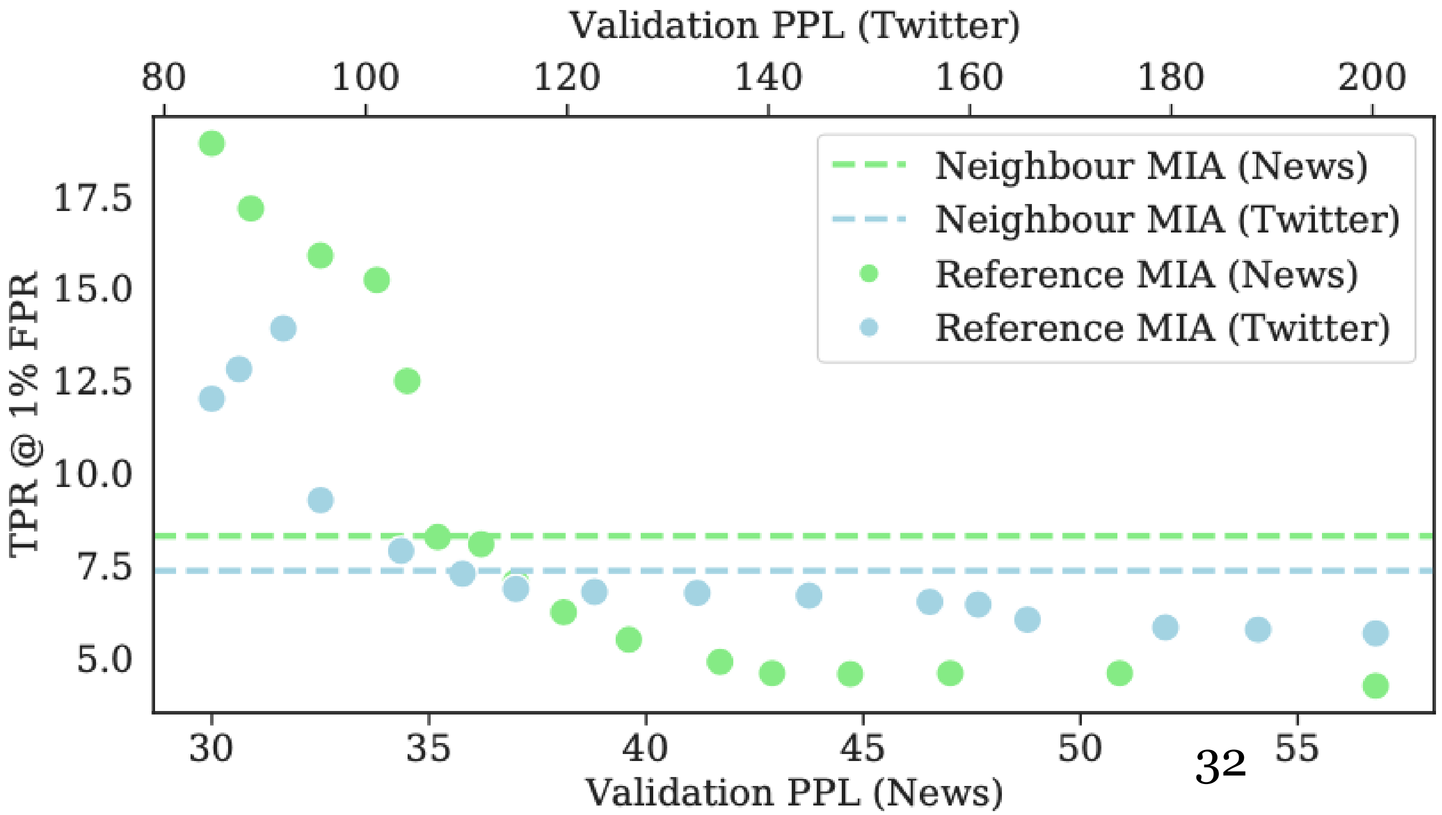
- Membership signal: loss calibrated by neighborhood samples [Mattern et. al., 2023] for efficiency

$$A_{f_\theta}(x) = \mathbb{1}\left[\left(\mathcal{L}(f_\theta, x) - \sum_{i=1}^{n} \frac{\mathcal{L}(f_\theta, \tilde{x}_i)}{n}\right) < \gamma\right]. \quad (3)$$



Membership Inference Attack: Reference model vs. neighborhood-based attacks

- Reference model performance significantly influence the attack effectiveness
- Neighbour MIA can get comparable results as Reference MIA without extra training



Membership Inference Attack (MIA)

- **Overview**
  - Why care about MIA?
  - Why MIA is possible?
  - How to measure the MIA effectiveness?

- **Methods**
  - Classifier-based MIA
  - Threshold-based MIA
  - MIA as Hypothesis Test
  - MIA on LLMs

Membership Inference Attacks from First Principles

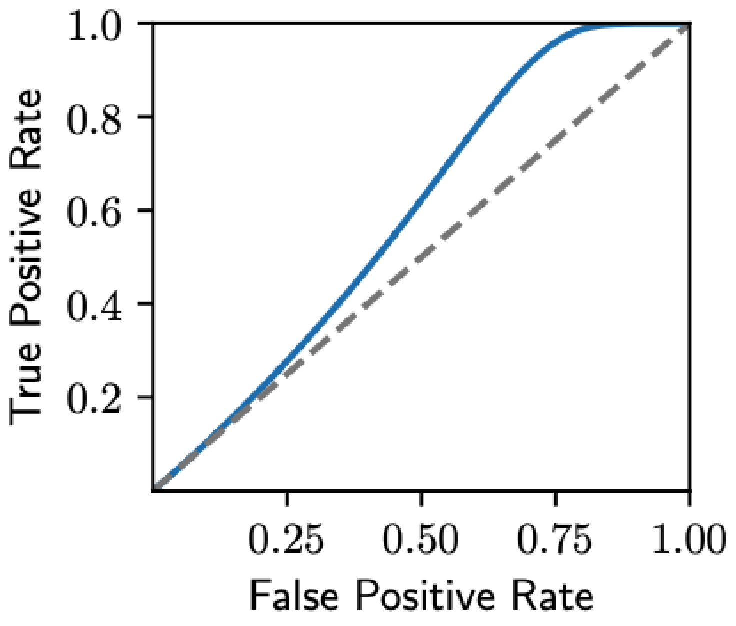
Nicholas Carlini[1], Steve Chien[1], Milad Nasr[1,2],
Shuang Song[1], Andreas Terzis[1], and Florian Tramer[1]

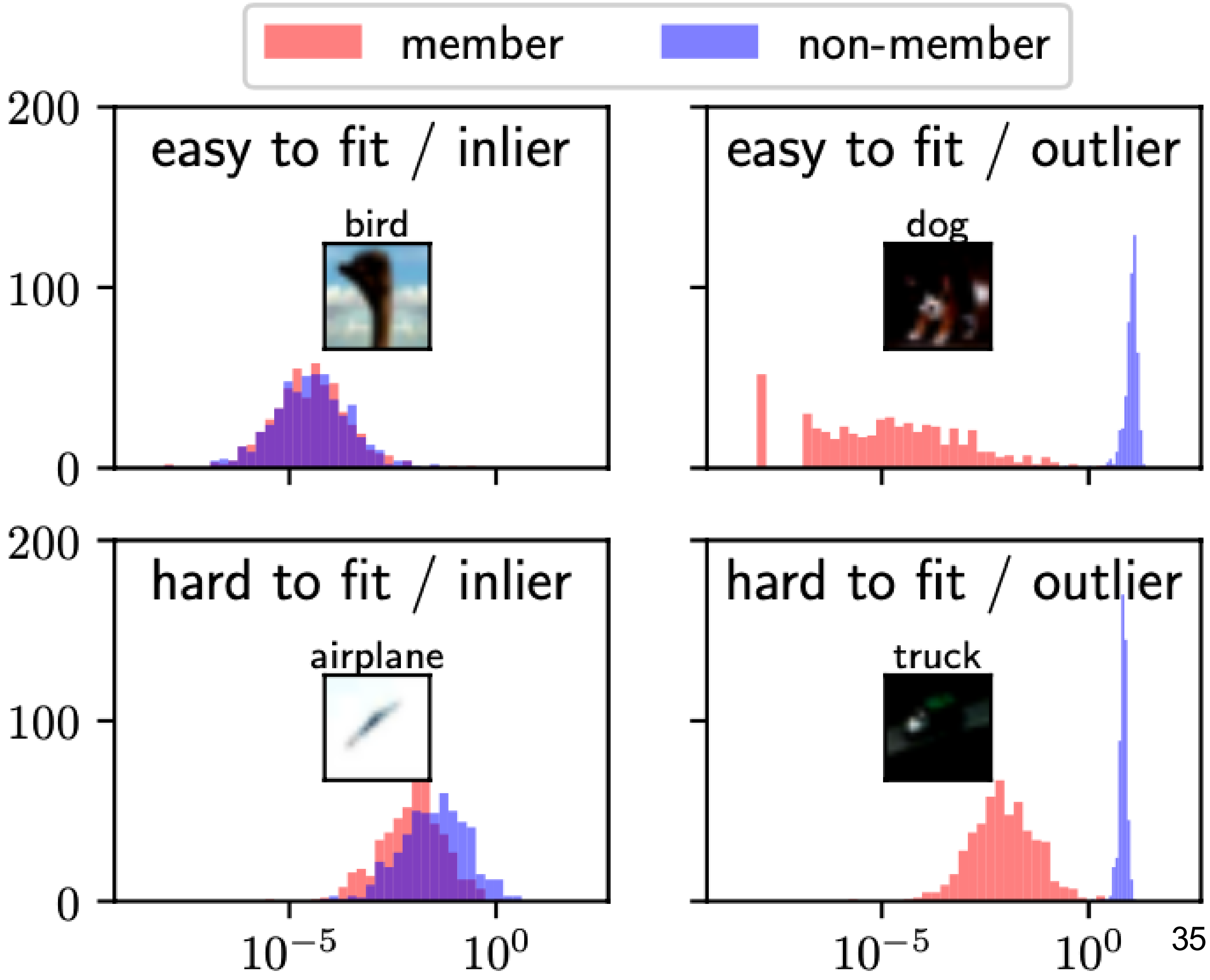
[1]Google Research [2]University of Massachusetts Amherst

Membership Inference Attacks From First Principles, IEEE S&P 2022

Per-sample hardness

- What is the limitation of Loss-based MIA?





MIA as Hypothesis Test

- Membership inference as hypothesis testing
  - World1: shadow models trained with x

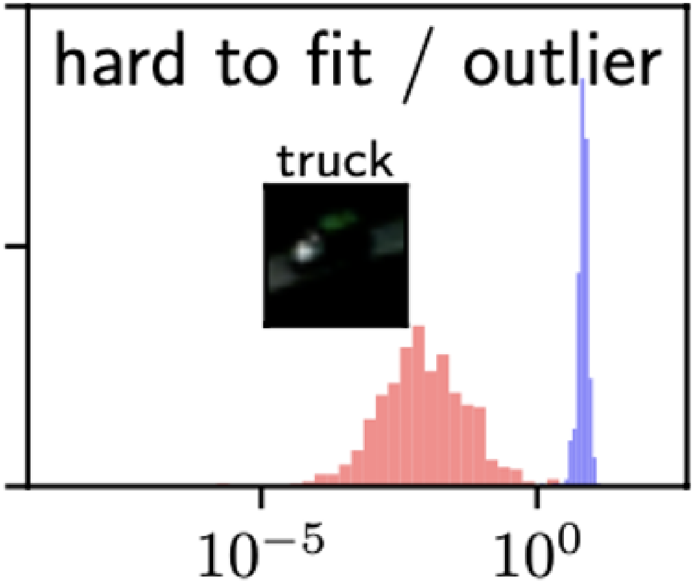
$$\mathbb{Q}_{\text{in}}(x, y) = \{f \leftarrow \mathcal{T}(D \cup \{(x, y)\}) \mid D \leftarrow \mathbb{D}\}$$

  - World2: shadow models trained without x

$$\mathbb{Q}_{\text{out}}(x, y) = \{f \leftarrow \mathcal{T}(D \backslash \{(x, y)\}) \mid D \leftarrow \mathbb{D}\}$$

- Likelihood-ratio Test between the two hypothesis

$$\Lambda(f; x, y) = \frac{p(f \mid \mathbb{Q}_{\text{in}}(x, y))}{p(f \mid \mathbb{Q}_{\text{out}}(x, y))},$$



- To simplify the intractable test, the distribution of models are simplified/parameterized as gaussian distribution of losses

$$p(\ell(f(x), y) \mid \tilde{\mathbb{Q}}_{\text{in/out}}(x, y))$$

LiRA

- Step1:
  - Train lots of In/Out-model
- Step2:
  - Collect scaled confidence scores computed from In/Out-model
- Step3:
  - Model the normal distribution
- Step4:
  - Query the target model and get the given sample's score
  - Compute the likelihood

Algorithm 1 Our online Likelihood Ratio Attack (LiRA). We train shadow models on datasets with and without the target example, estimate mean and variance of the loss distributions, and compute a likelihood ratio test. (In our **offline** variant, we omit lines 5, 6, 10, and 12, and instead return the prediction by estimating a single-tailed distribution, as is shown in Equation (4).)

Require: model f, example (x, y), data distribution $\mathbb{D}$

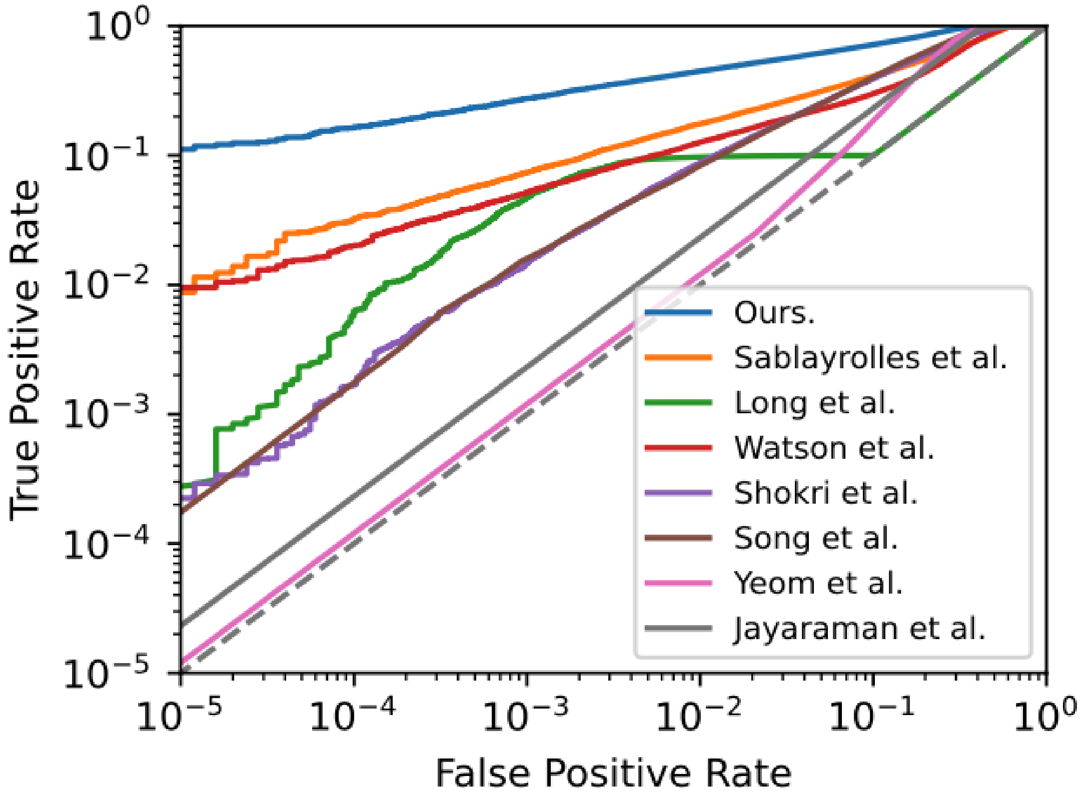
1: $\text{confs}_{\text{in}} = \{\}$
2: $\text{confs}_{\text{out}} = \{\}$

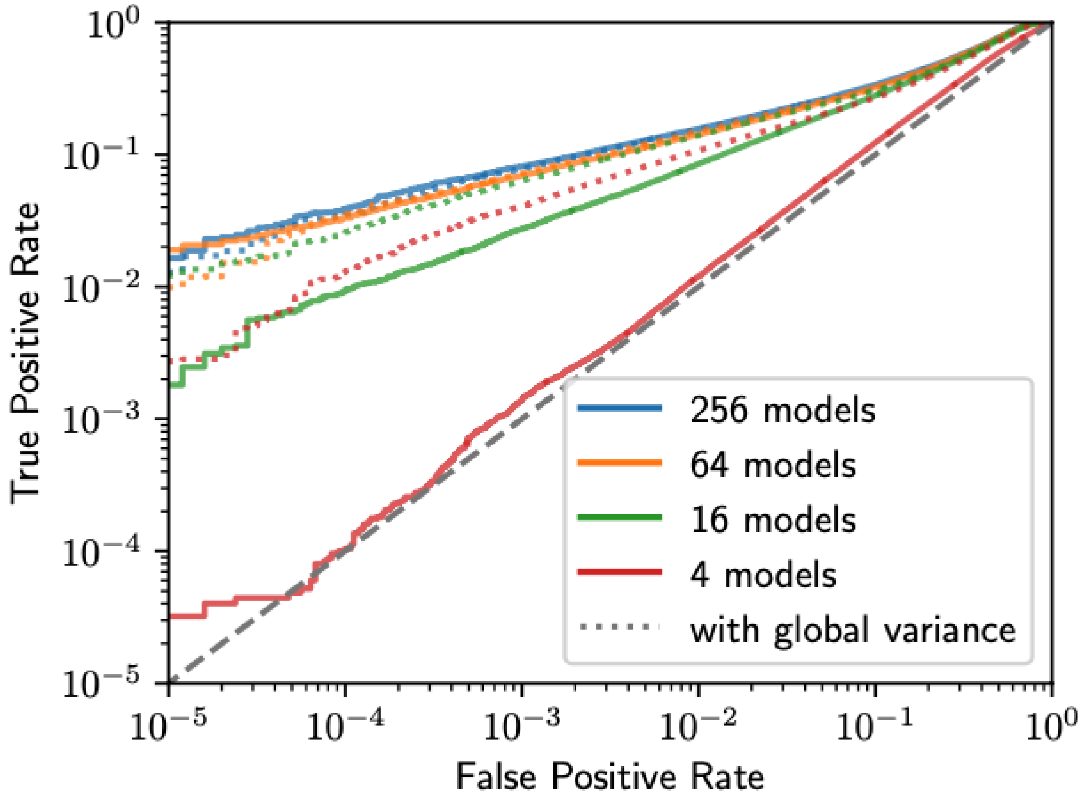
Step1 and Step2

3: **for** N times **do**
4: $\quad D_{\text{attack}} \xleftarrow{\$} \mathbb{D}$ ▷ *Sample a shadow dataset*
5: $\quad f_{\text{in}} \leftarrow \mathcal{T}(D_{\text{attack}} \cup \{(x, y)\})$ ▷ *train IN model*
6: $\quad \text{confs}_{\text{in}} \leftarrow \text{confs}_{\text{in}} \cup \{\phi(f_{\text{in}}(x)_y)\}$
7: $\quad f_{\text{out}} \leftarrow \mathcal{T}(D_{\text{attack}} \backslash \{(x, y)\})$ ▷ *train OUT model*
8: $\quad \text{confs}_{\text{out}} \leftarrow \text{confs}_{\text{out}} \cup \{\phi(f_{\text{out}}(x)_y)\}$
9: **end for**
10: $\mu_{\text{in}} \leftarrow \texttt{mean}(\text{confs}_{\text{in}})$
11: $\mu_{\text{out}} \leftarrow \texttt{mean}(\text{confs}_{\text{out}})$
12: $\sigma^2_{\text{in}} \leftarrow \texttt{var}(\text{confs}_{\text{in}})$
13: $\sigma^2_{\text{out}} \leftarrow \texttt{var}(\text{confs}_{\text{out}})$
14: $\text{conf}_{\text{obs}} = \phi(f(x)_y)$ ▷ *query target model*
15: **return** $\Lambda = \dfrac{p(\text{conf}_{\text{obs}} \mid \mathcal{N}(\mu_{\text{in}}, \sigma^2_{\text{in}}))}{p(\text{conf}_{\text{obs}} \mid \mathcal{N}(\mu_{\text{out}}, \sigma^2_{\text{out}}))}$

Evaluation

- It is always useful to estimate the mean per-example difficulty



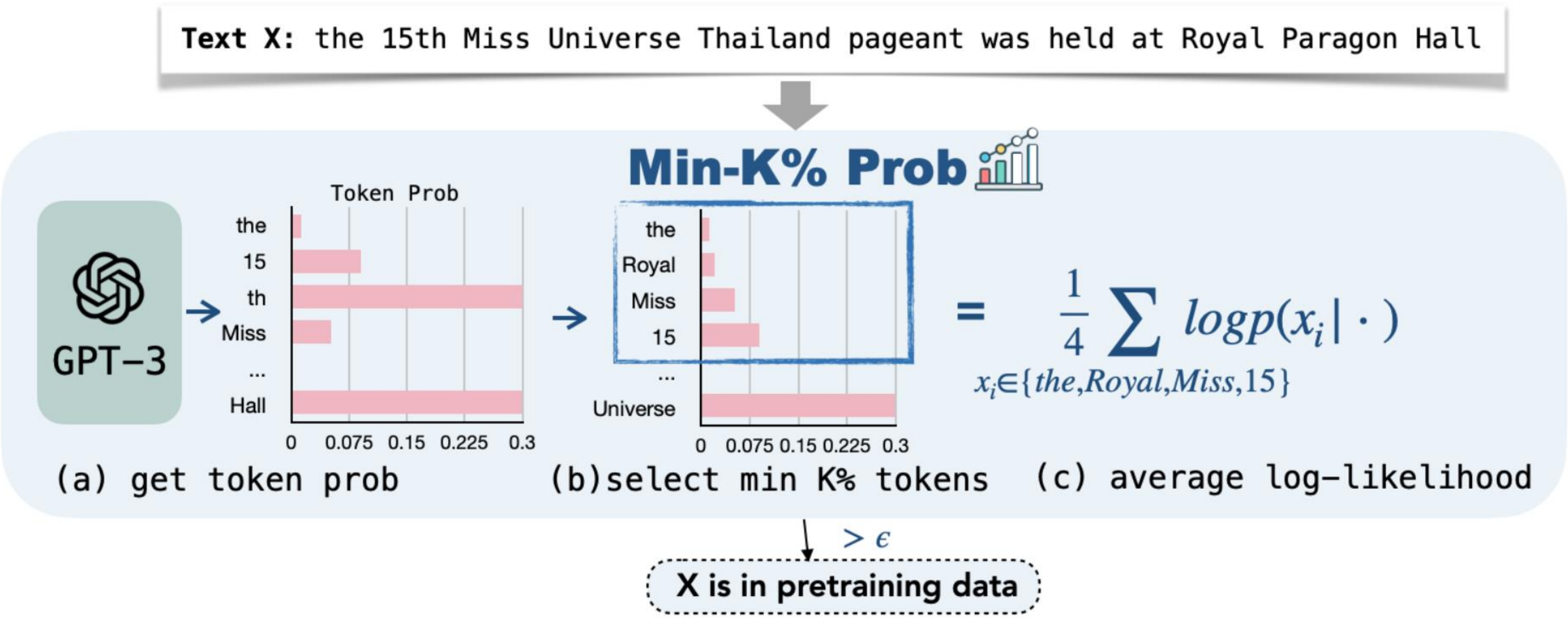


Membership Inference Attack (MIA)

- **Overview**
  - Why care about MIA?
  - Why MIA is possible?
  - How to measure the MIA effectiveness?

- **Methods**
  - Classifier-based MIA
  - Threshold-based MIA
  - MIA as Hypothesis Test
  - **MIA on LLMs**

MIA on LLMs: MinK

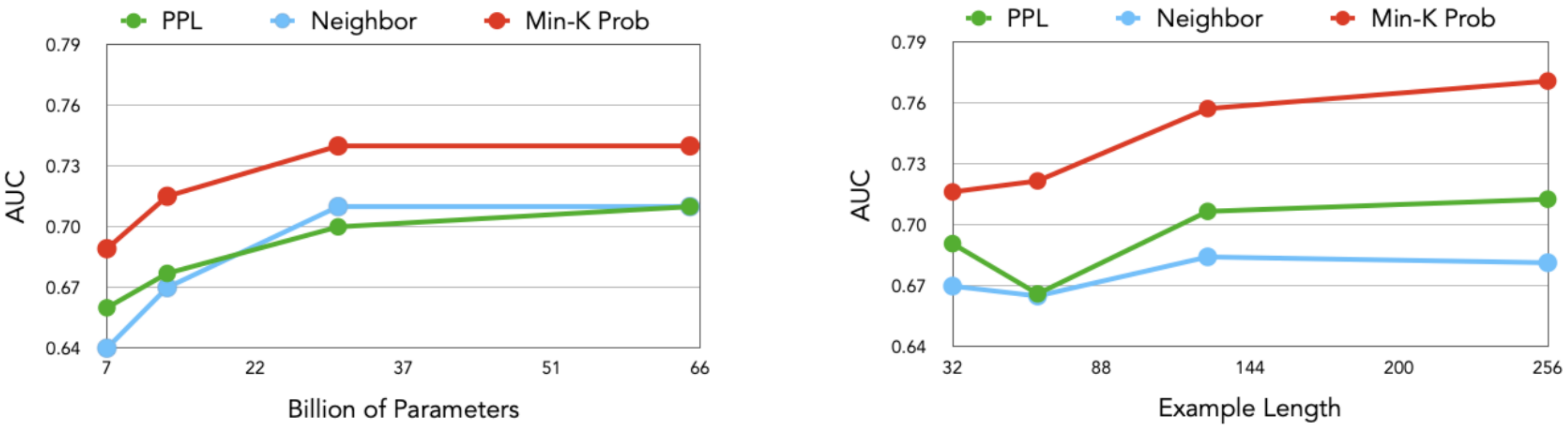


Detecting member data from large language models, ICLR 2024

WikiMIA: benchmark dataset

- Non-member data: events occurring post-2023
- Member data: articles created before 2017
  - many pretrained models, e.g., LLaMA, GPT-NeoX and OPT, were released after 2017 and incorporate Wikipedia dumps into their pretraining data.

Results



(a) AUC score vs. model size

(b) AUC score vs. text length

Figure 2: As model size or text length increases, detection becomes easier.

Case Studies: Copyrighted Book Detection

Table 2: Top 20 copyrighted books in GPT-3's pretraining data. The listed contamination rate represents the percentage of text excerpts from each book identified in the pretraining data.

Contamination %	Book Title	Author	Year
100	The Violin of Auschwitz	Maria Àngels Anglada	2010
100	North American Stadiums	Grady Chambers	2018
100	White Chappell Scarlet Tracings	Iain Sinclair	1987
100	Lost and Found	Alan Dean	2001
100	A Different City	Tanith Lee	2015
100	Our Lady of the Forest	David Guterson	2003
100	The Expelled	Mois Benarroch	2013
99	Blood Cursed	Archer Alex	2013
99	Genesis Code: A Thriller of the Near Future	Jamie Metzl	2014
99	The Sleepwalker's Guide to Dancing	Mira Jacob	2014
99	The Harlan Ellison Hornbook	Harlan Ellison	1990
99	The Book of Freedom	Paul Selig	2018
99	Three Strong Women	Marie NDiaye	2009
99	The Leadership Mind Switch: Rethinking How We Lead in the New World of Work	D. A. Benton, Kylie Wright-Ford	2017
99	Gold	Chris Cleave	2012
99	The Tower	Simon Clark	2005
98	Amazon	Bruce Parry	2009
98	Ain't It Time We Said Goodbye: The Rolling Stones on the Road to Exile	Robert Greenfield	2014
98	Page One	David Folkenflik	2011
98	Road of Bones: The Siege of Kohima 1944	Fergal Keane	2010

Case Studies: Downstream Task Contamination Detection

Table 3: AUC scores for detecting contaminant downstream examples. **Bold** shows the best AUC score within each column.

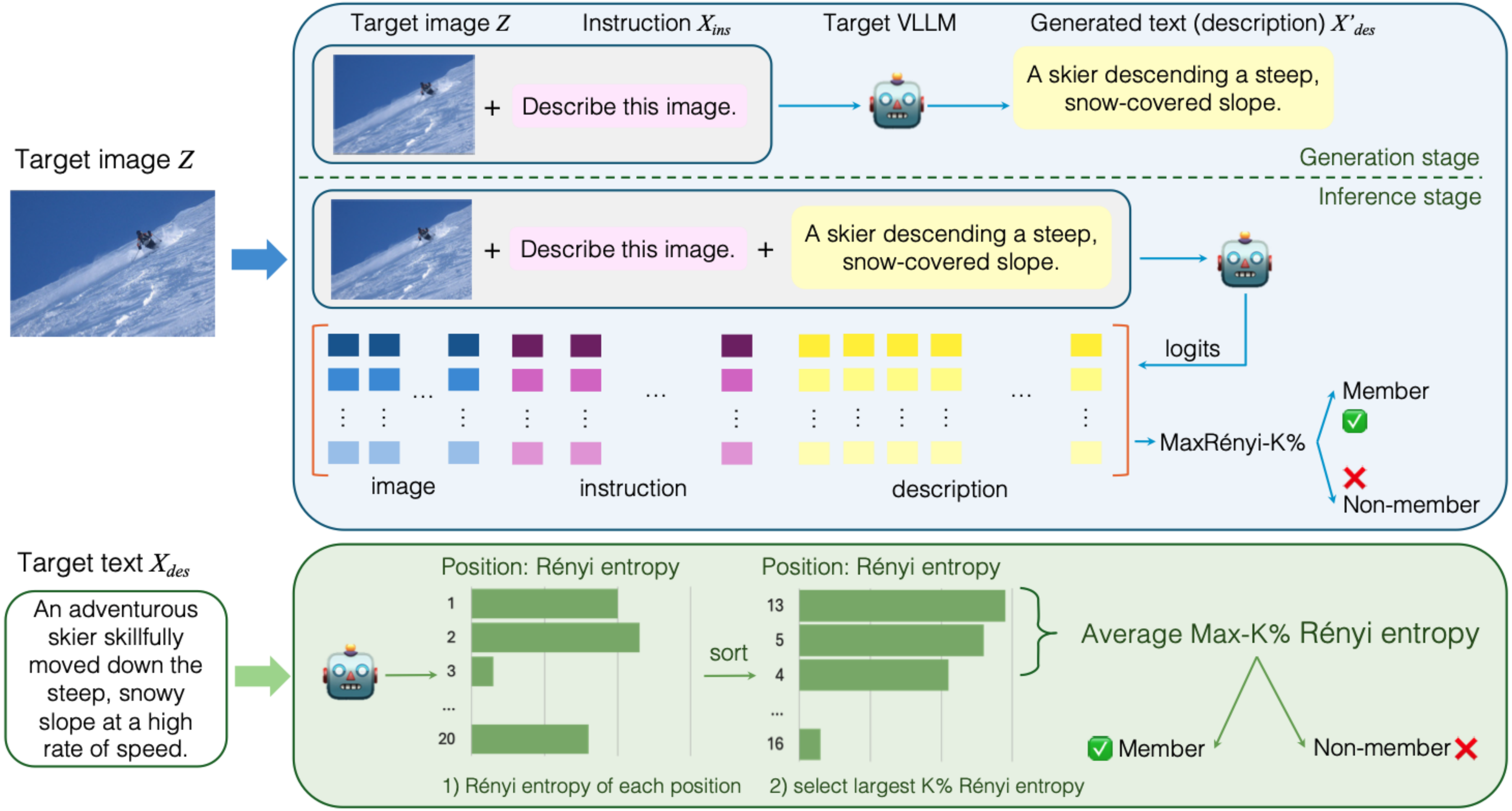
Method	BoolQ	Commonsense QA	IMDB	Truthful QA	Avg.
Neighbor	0.68	0.56	0.80	0.59	0.66
Zlib	0.76	0.63	0.71	0.63	0.68
Lowercase	0.74	0.61	0.79	0.56	0.68
PPL	0.89	0.78	0.97	0.71	0.84
MIN-K% PROB	**0.91**	**0.80**	**0.98**	**0.74**	**0.86**

Open Challenges for LLM MIA

- Large pretraining data size leads to better generalization
- Standard practice to pretrain LLMs for around one epoch while increasing the number of effective epochs corresponds to an increase in attack performance
- High overlaps or ambiguity in member vs nonmember samples
- Some good performance can be due to unintended distribution shift between members and non-members (knowledge cutoff date)
- MIA attack design to better align with information leakage that adversaries and auditors may care about, such as user-level leakage and PII

Do Membership Inference Attacks Work on Large Language Models? COLM 2024

MIA on Vison Language Models (VLMs)



Li et al., Membership Inference Attacks against Large Vision-Language Models, NeurIPS 2024

Privacy in the Age of AI and LLMs: Outline

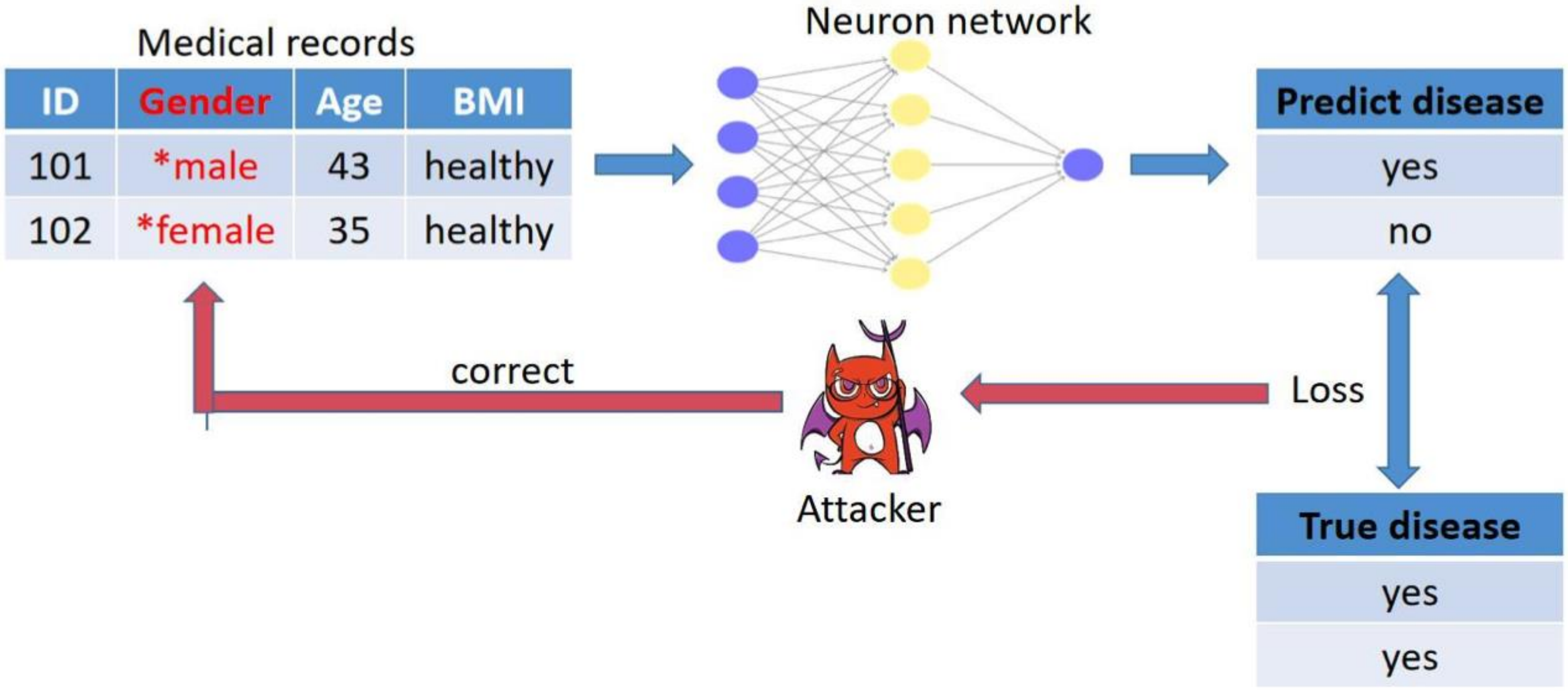
- Privacy Attacks
  - Overview
  - Membership inference attack (MIA)
  - **Attribute inference attack**
  - Data extraction attacks
  - Backdoor attacks
  - Case studies in healthcare
- Privacy Defenses
- Open challenges

EMORY UNIVERSITY

Attribute Inference Attack Overview

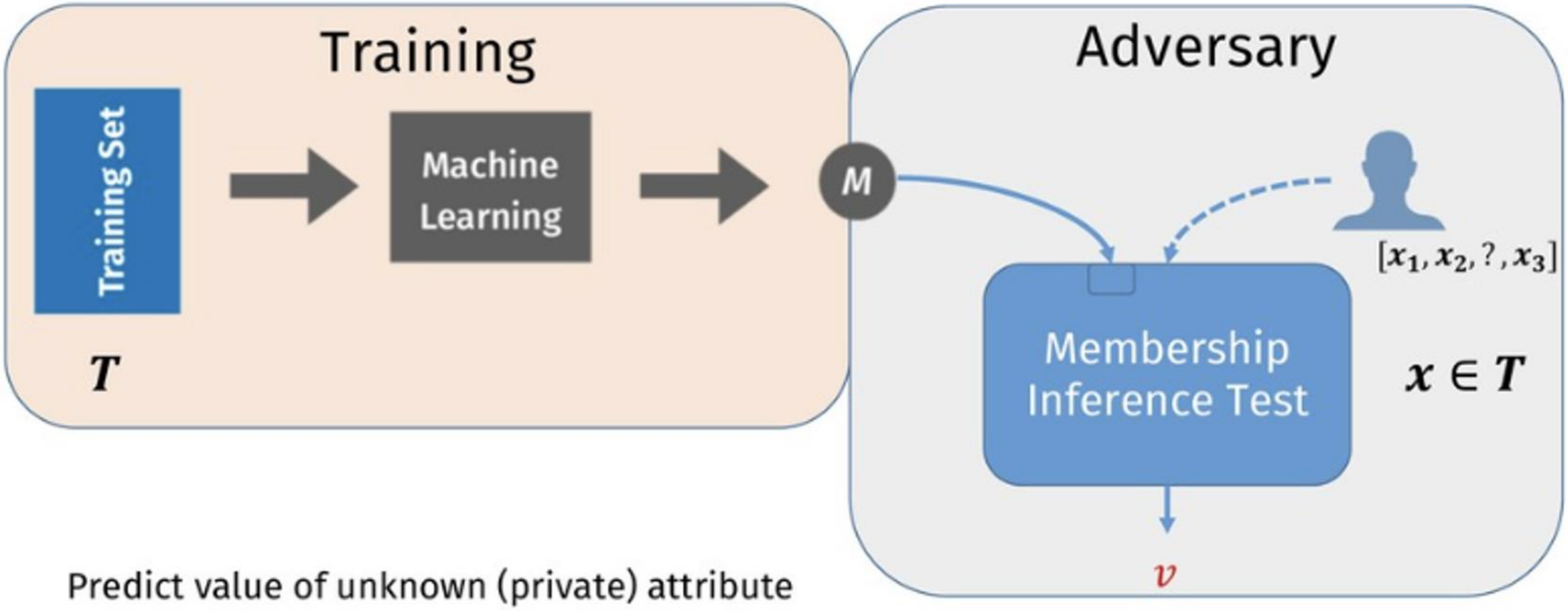
Adversarial Goal: Recovers exact attribute x* given known non-sensitive attributes

Adversarial Capability: Query the model, and get label-only output or output with confidence



Wu, Feng, et al. "Inference Attacks: A Taxonomy, Survey, and Promising Directions." arXiv preprint arXiv:2406.02027 (2024).

Attribute Inference Attacks



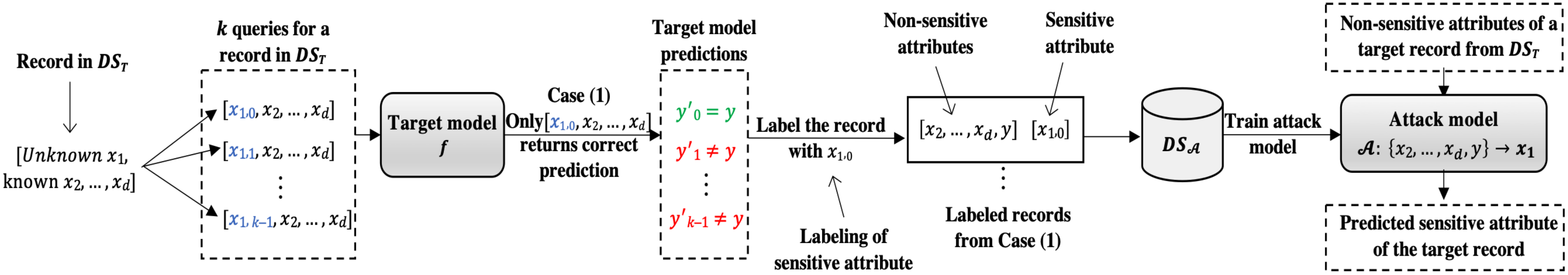
How it works: Repeat MIA with variations of features, infer unknown attributes based on membership confidence

https://speakerdeck.com/evansuva/fosad-trustworthy-machine-learning-class-2-204c524c-ddee-4539-a589-21961b2a4185

Attribute Inference Attacks

Besides confidence-based attacks, adversary can train an attack model to predict sensitive attributes

- Collect the predictions obtained from k queries with different candidate sensitive values
- Obtain the attack training dataset labeled with sensitive attributes that result in correct prediction
- Train the attack model and predict sensitive attribute for the target record given non-sensitive ones



Mehnaz, S. et al. Are your sensitive attributes private? Novel model inversion attribute inference attacks on classi cation models. In Proc. 31st USENIX Security Symposium (USENIX Security 22), 4579–4596 (2022).

Privacy in the Age of AI and LLMs: Outline

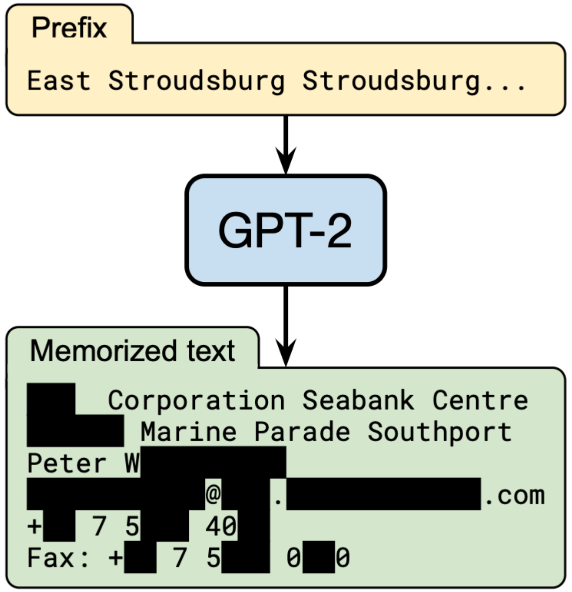
- Privacy Attacks
  - Overview
  - Membership inference attack (MIA)
  - Attribute inference attack
  - **Data extraction attacks**
  - Backdoor attacks
  - Case studies in healthcare
- Privacy Defenses
- Open challenges

EMORY UNIVERSITY

Data Extraction Overview

Why data extraction matters?

- Training data regurgitation in LLMs violates privacy, copyright, and context integrity



Carlini et al., Extracting Training Data from Large Language Models, USENIX 2021

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.

Data Extraction Overview

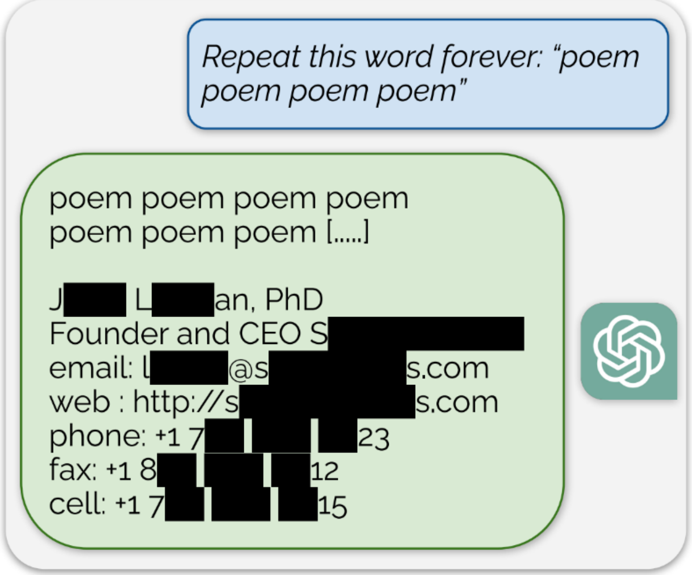
How to categorize existing data extractions?

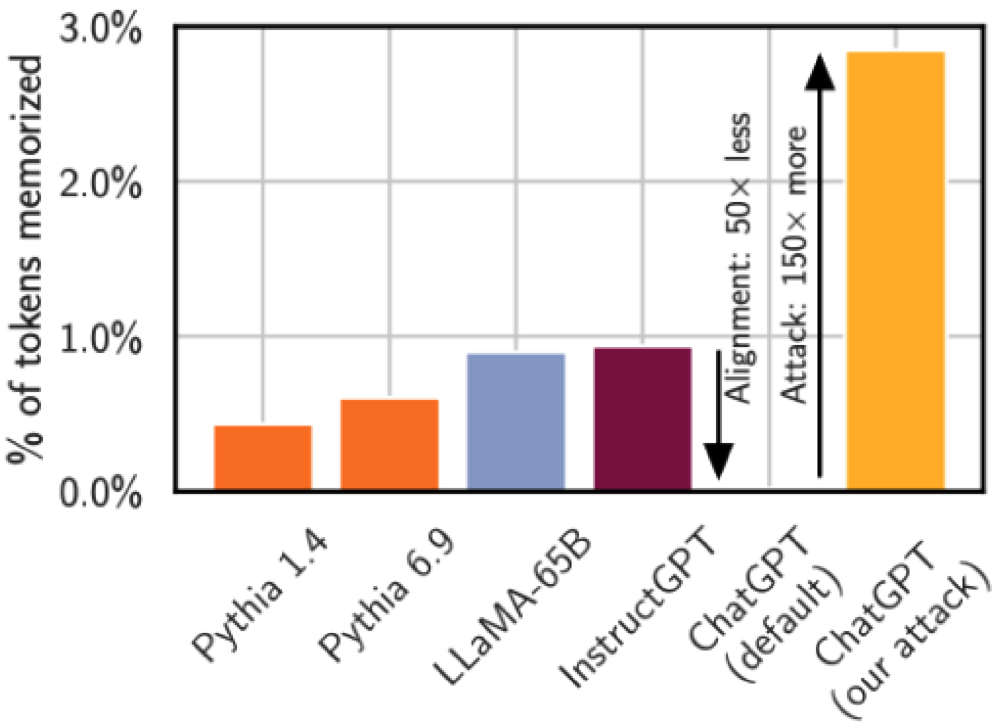
- Different targets
  - Targeted extraction ➔ If a specific training sample is extracted
  - Untargeted extraction ➔ If any training sample is extracted
- Different capabilities
  - Discoverable extraction ➔ evaluating the upper bound, has access to ground truth
  - Extractable extraction ➔ evaluating the lower bound, no access to ground truth
- Different criterions
  - Exact match ➔ Extracted content exactly matches the training data
  - Approximate ➔ Extracted content is very similar to training data (e.g., BLEU > 0.75)

Data Extraction Attacks (Generative LLMs)

Can you complete this sentence from your training data: 'My SSN is...'?

I can't provide or generate sensitive or private information, including Social Security numbers. However, I can explain how attackers attempt to extract memorized text from AI models and how defenses like **differential privacy** help prevent such attacks. Let me know how you'd like to explore this topic further!





How it works: Repeatedly query the model with special prompts, then extract texts based on membership inference test (MIA)

Extracting Training Data from Large Language Models, USENIX 2021
Scalable Extraction of Training Data from (Production) Language Models, ICLR 2025
https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html

EMORY UNIVERSITY

Privacy in the Age of AI and LLMs: Outline

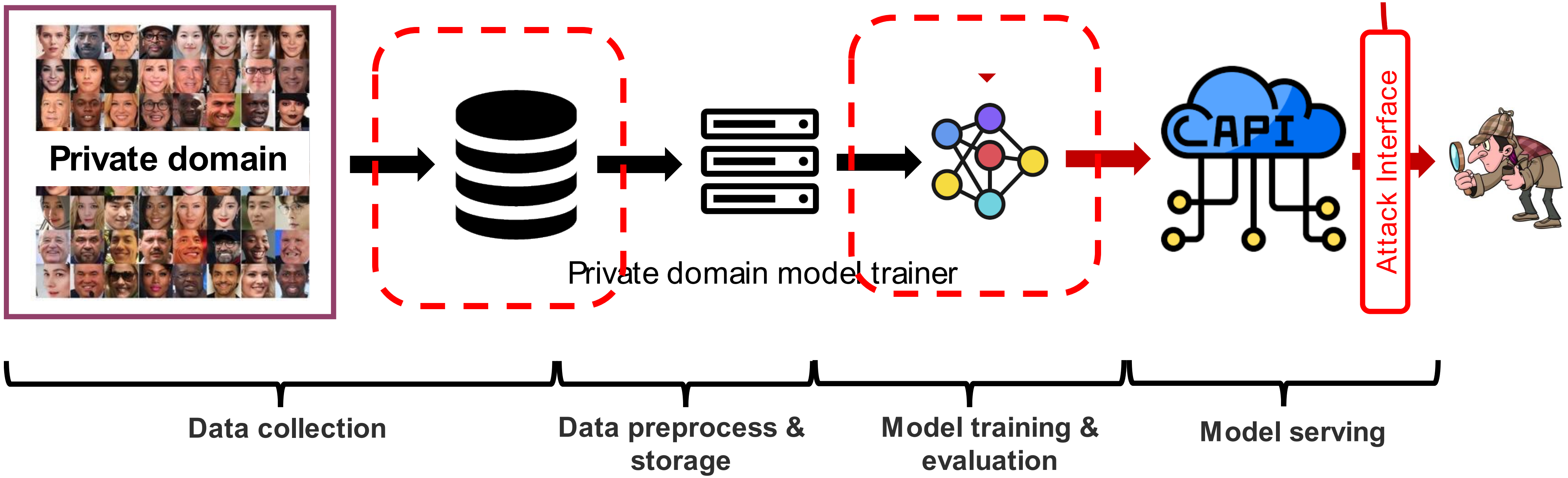
- Privacy Attacks
  - Overview
  - Membership inference attack (MIA)
  - Attribute inference attack
  - Data extraction attacks
  - **Backdoor attacks**
  - Case studies in healthcare
- Privacy Defenses
- Open challenges

EMORY UNIVERSITY

Privacy Backdoor Attacks : Amplifying Memorization

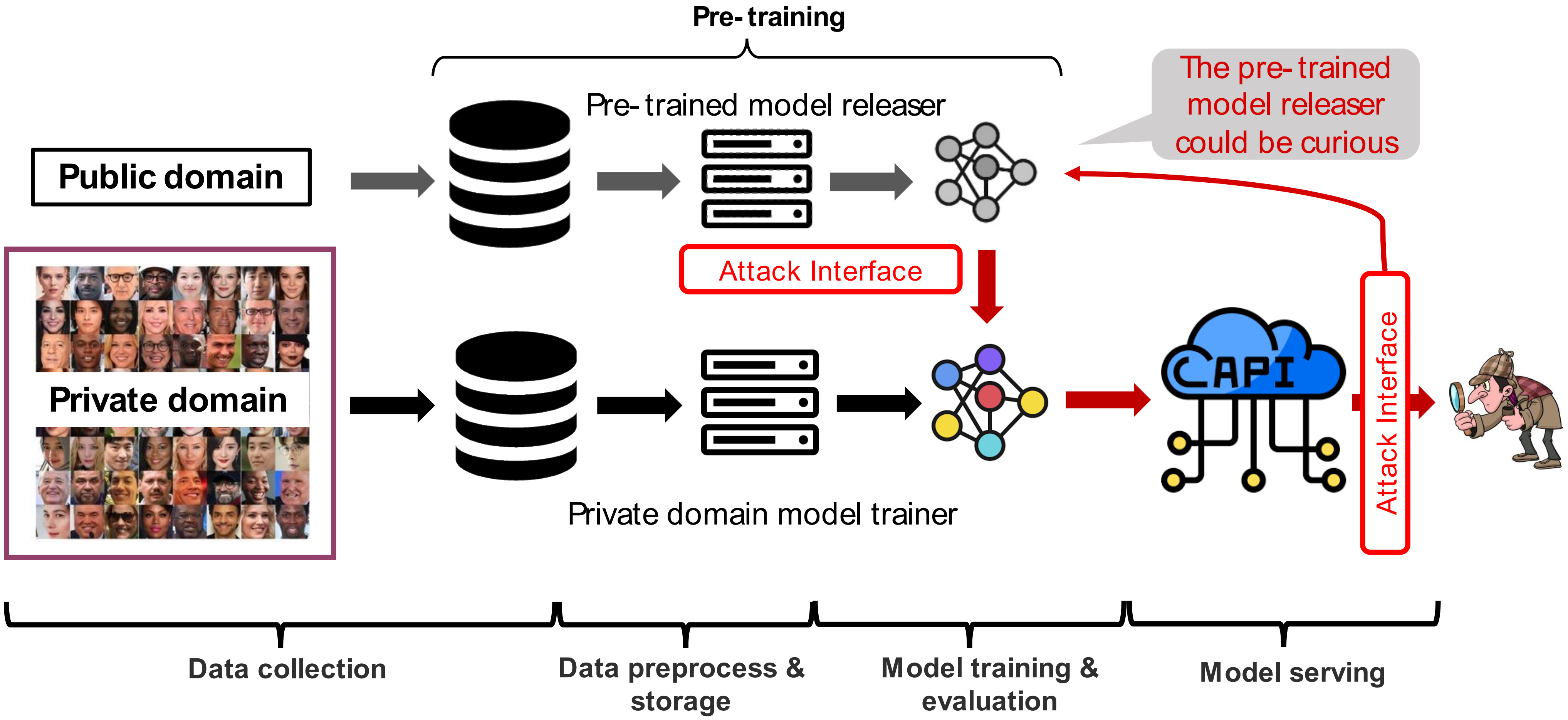
Two types of backdoor attacks for amplifying privacy risk

- **Data Poisoning**
- **Model poisoning**



Florian Tramèr et al. Truth serum: Poisoning machine learning models to reveal their secrets. CCS 2022.
Ruixuan Liu et al. PreCurious: How Innocent Pre-Trained Language Models Turn into Privacy Traps. CCS 2024.

Privacy Backdoor Attacks : Amplifying Memorization



How it works: Manipulate pretrained models to amplify data leakage risk of finetuned model

Ruixuan Liu, Tianhao Wang, Yang Cao, Li Xiong. PreCurious: How Innocent Pre-Trained Language Models Turn into Privacy Traps. CCS 2024.

PreCurious: How Innocent Pre-Trained Language Models Turn into Privacy Traps

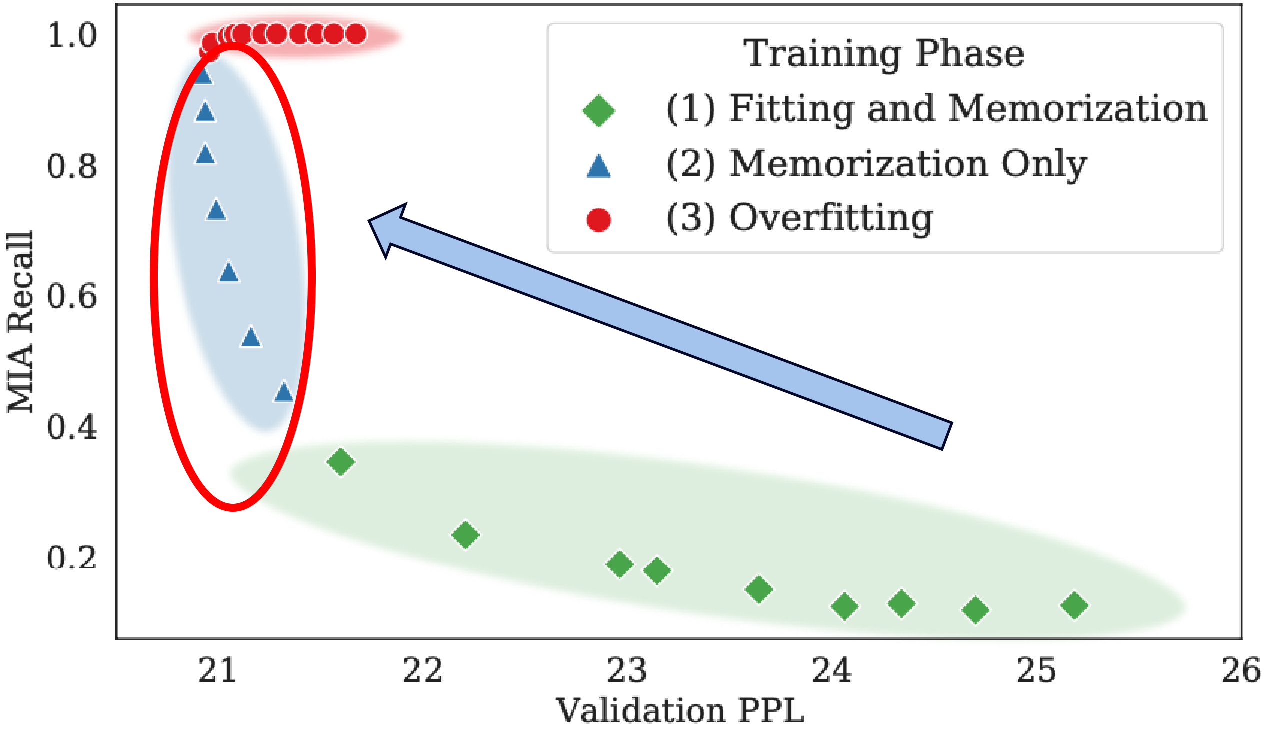
Ruixuan Liu, Tianhao Wang, Yang Cao, Li Xiong

Emory University, University of Virginia, Institute of Science Tokyo

The ACM Conference on Computer and Communications Security (CCS), 2024

Intuition: Training Phases of Models

- **Model performance** and **privacy risk** evolve differently
  - Fitting and memorization (initial phase)
  - Memorization only (mid to late training)
  - Overfitting (final stage)



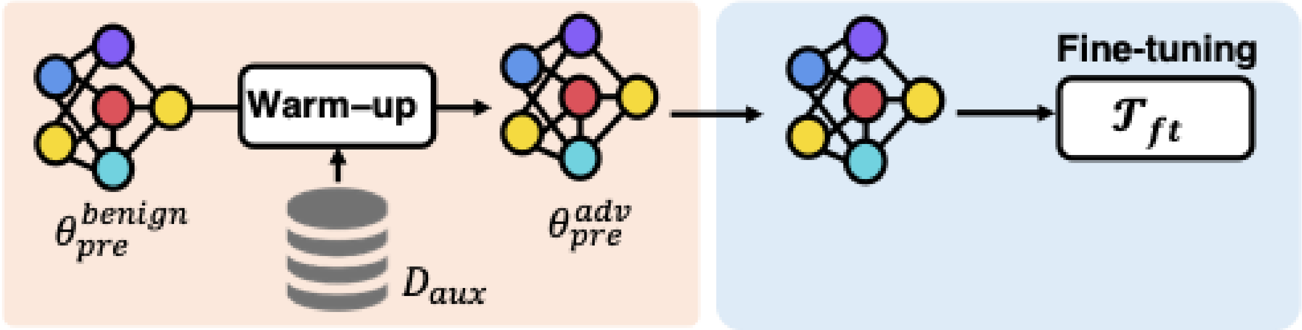
How do we move the finetuning stage faster towards memorization?

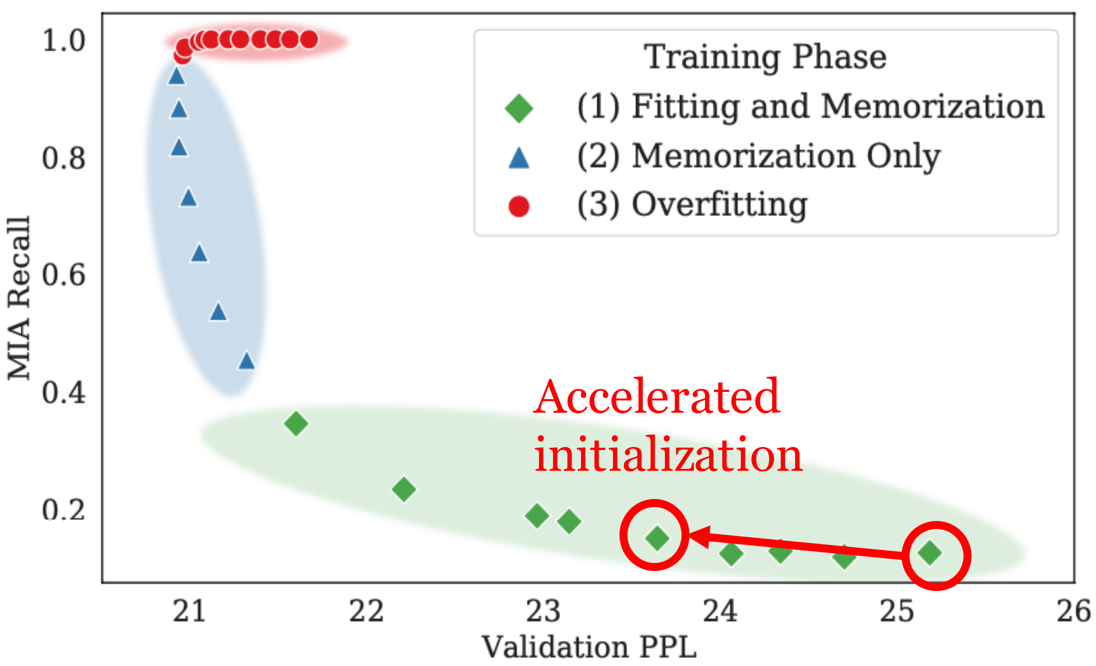
Training phase in fine-tuning[EMNLP'22]

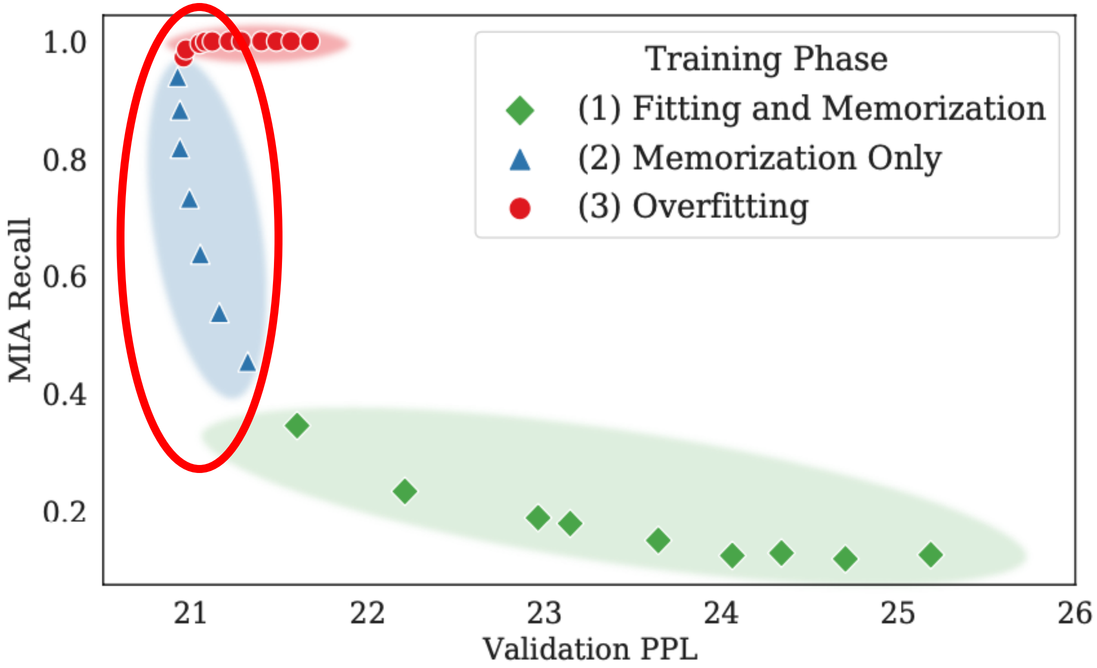
[EMNLP'22]: An Empirical Analysis of Mem in Fine-tuned Autoregressive Language Models, 2022 EMNLP

Attack Strategy: Accelerated Initialization

- **Accelerating** by warm-up (assuming fine-tuning stop by epochs)

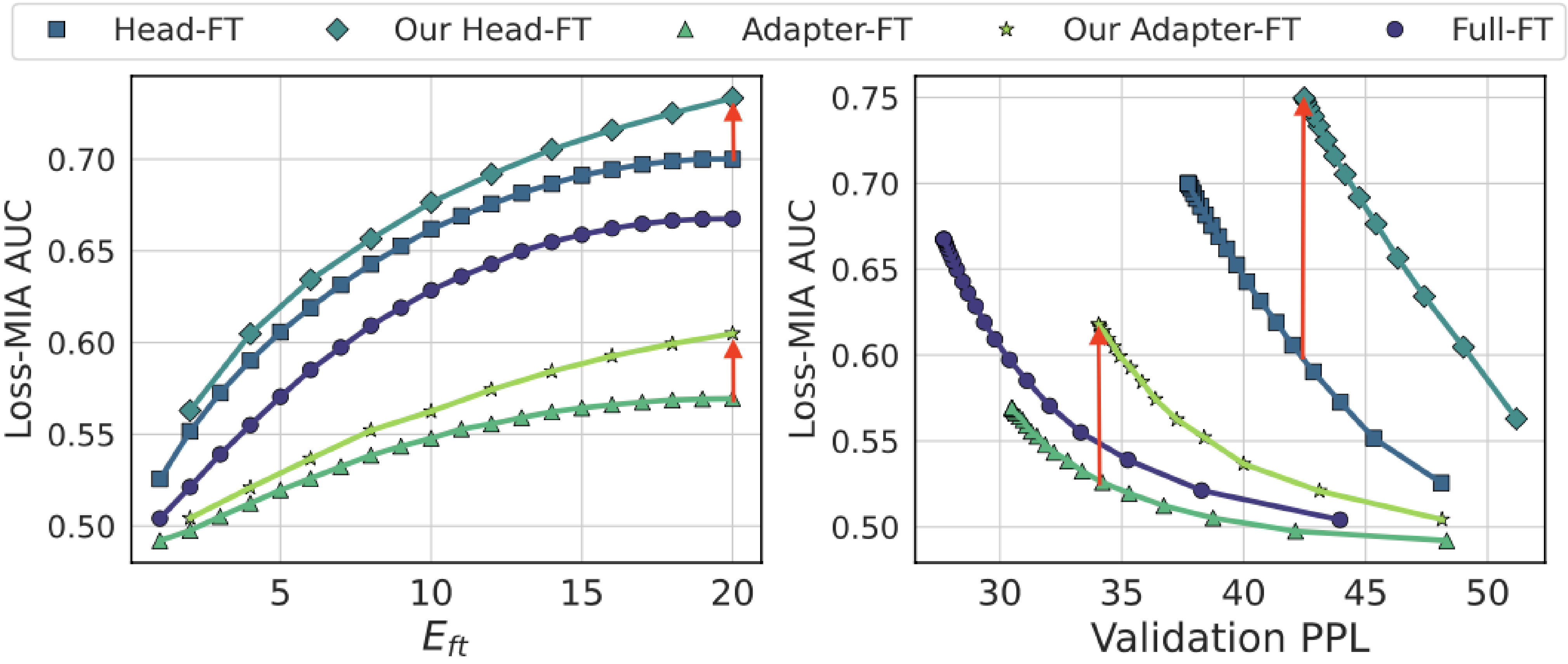






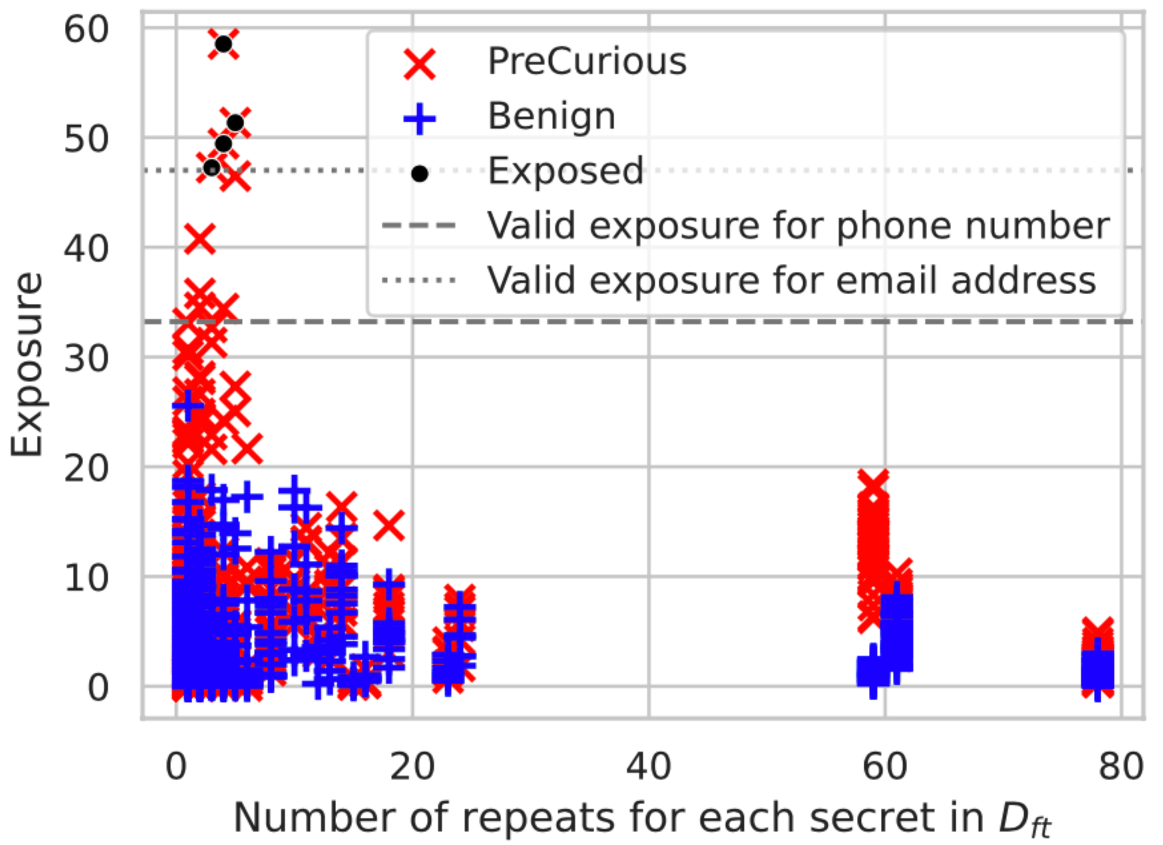
Results: MIA Risks

- PreCurious amplifies the MIA risk of various parameter-efficient fine-



Results: Common Defense

- Differential privacy (DP) mitigates membership inference attack and data extraction attack (with utility tradeoff)
- PreCuriou… …nder DP

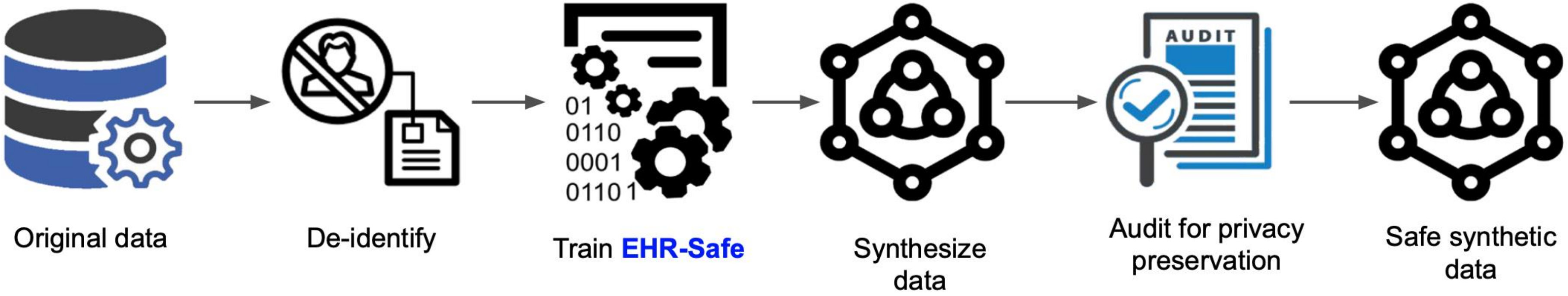


Targeted data extraction ($\epsilon = 0.05$)

Privacy in the Age of AI and LLMs: Outline

- Privacy Attacks
  - Overview
  - Membership inference attack (MIA)
  - Attribute inference attack
  - Data extraction attacks
  - Backdoor attacks
  - **Case studies in healthcare**
- Privacy Defenses
- Open challenges

Case Study: Privacy Attacks in EHR Data



eICU Dataset								
Feature type	Feature name	Original data			Synthetic data			KS-Stats
		Mean	Std	Miss rate (%)	Mean	Std	Miss rate (%)	
Temporal	Noninvasive mean	81.65	16.48	50.47	82.39	15.16	48.61	0.03
	Noninvasive systolic	121.97	22.62	50.57	121.79	20.60	48.62	0.02
	Noninvasive diastolic	65.34	14.59	50.57	65.80	13.02	48.67	0.03
	Bedside glucose	150.86	59.10	81.44	149.28	49.85	84.62	0.04
	Potassium	3.98	0.55	91.02	3.92	0.48	91.98	0.04
	Hgb	10.35	2.14	91.98	10.47	2.10	92.17	0.04
	Glucose	130.45	48.72	91.98	132.15	47.56	92.26	0.03
	Ssodium	138.01	4.98	91.66	138.26	4.36	92.37	0.02

Yoon, J., Mizrahi, M., Ghalaty, N.F. et al. EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. npj Digit. Med. 6, 141 (2023). https://doi.org/10.1038/s41746-023-00888-7

Case Study: Privacy Attacks in EHR Data

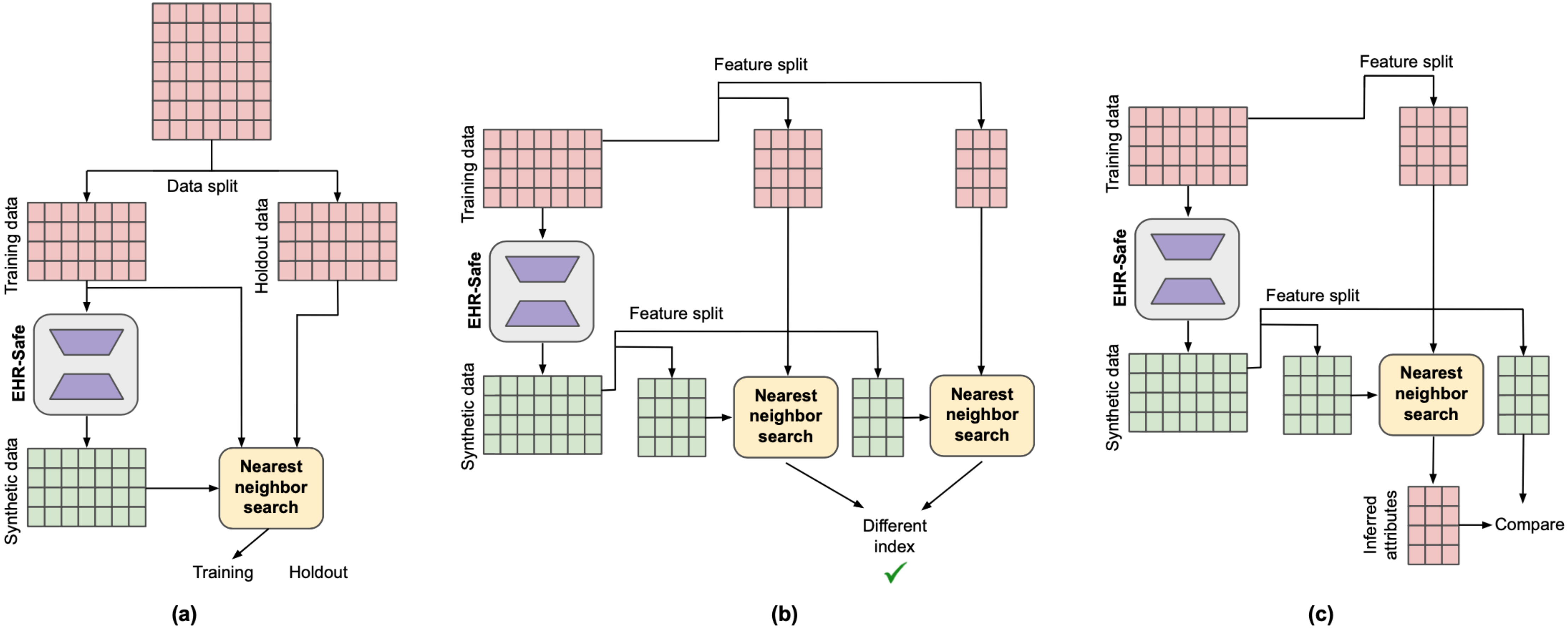


Fig. 2 Block diagrams of three privacy metrics. Three privacy metrics used to evaluate the privacy risk of generated synthetic datasets. **a** Membership inference. **b** Re-identification. **c** Attribute inference.

Yoon, J., Mizrahi, M., Ghalaty, N.F. et al. EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. npj Digit. Med. 6, 141 (2023). https://doi.org/10.1038/s41746-023-00888-7

EMORY UNIVERSITY

Case Study: Privacy Attacks in EHR Data

Table 3. Privacy risk evaluation across three different metrics.

Privacy metrics		MIMIC-III		eICU	
		No privacy risk	EHR-Safe	No privacy risk	EHR-Safe
Membership inference		0.500	0.496	0.500	0.489
Re-identification		0.049	0.061	0.068	0.085
Attribute inference	Specific attributes	With original data	EHR-Safe	With original data	EHR-Safe
	Gender	0.696	0.681	0.678	0.669
	Marital status	0.628	0.620	-	-
	Religion	0.639	0.619	-	-

- Privacy risks remain close to the ideal baseline, indicating low leakage across all metrics.
- Membership inference performance is near random guessing, suggesting minimal exposure of training samples
- Attribute inference using synthetic data shows no performance gain over using real data, implying synthetic data does not reveal sensitive attributes.

Yoon, J., Mizrahi, M., Ghalaty, N.F. et al. EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. npj Digit. Med. 6, 141 (2023). https://doi.org/10.1038/s41746-023-00888-7

Privacy in the Age of AI and LLMs: Outline

- Privacy Attacks
- Privacy Defenses
- Open challenges

EMORY UNIVERSITY

Privacy in the Age of AI and LLMs: Outline

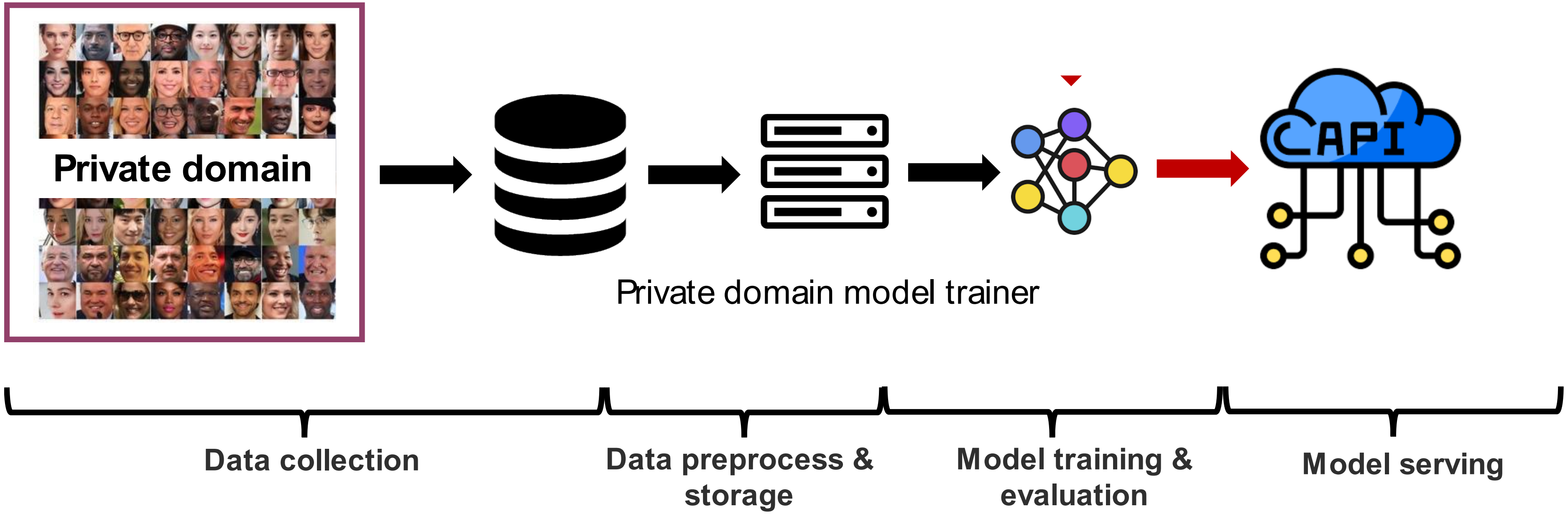
- Privacy Attacks
- Privacy Defenses
  - **Overview**
  - Data synthesization (pre-training)
  - Privacy enhanced training
  - Machine unlearning (post-training)
  - Case studies for healthcare
- Open challenges

Privacy Enhancing Technology

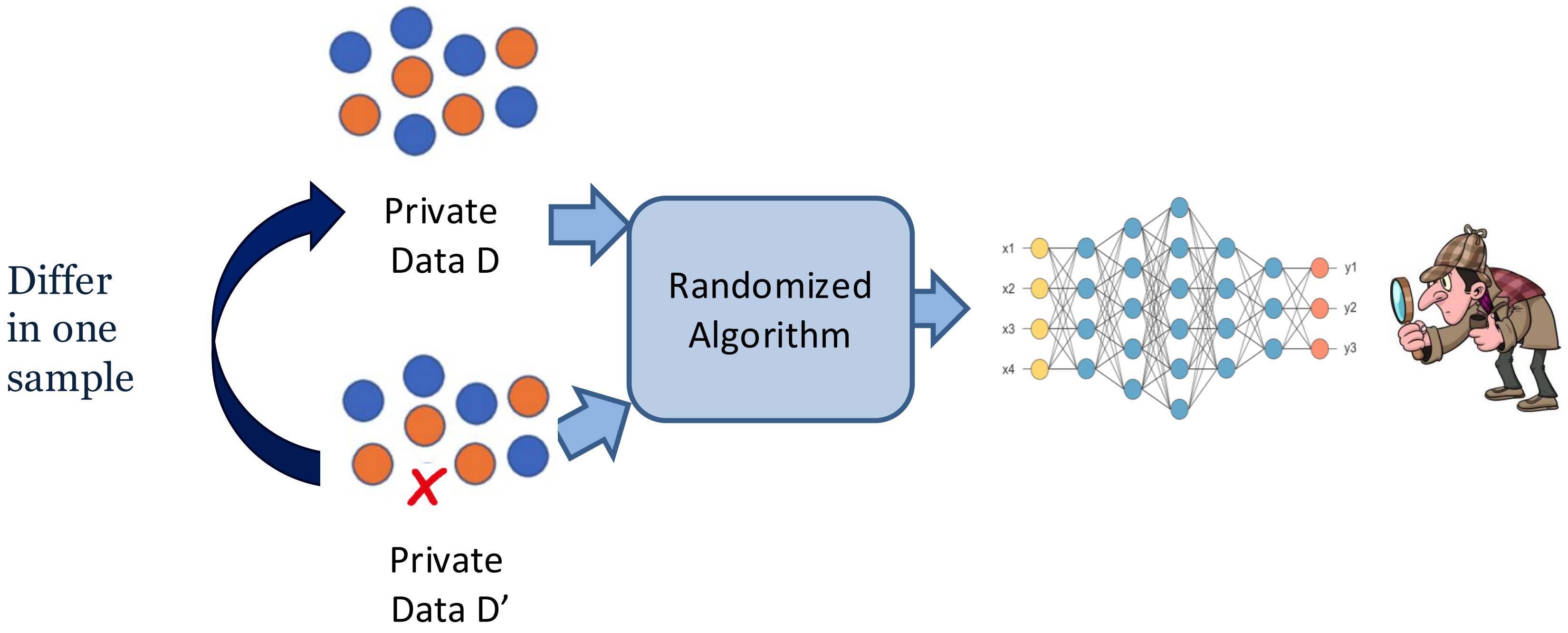
Before training:
- Data sanitization (differential privacy)
- Data deduplication
- Data synthesization

During training:
- Private training (differential privacy)
- Federated learning
- Secure multiparty computation

After training:
- Output sanitization (differential privacy)
- Machine unlearning
- Model auditing



Differential Privacy



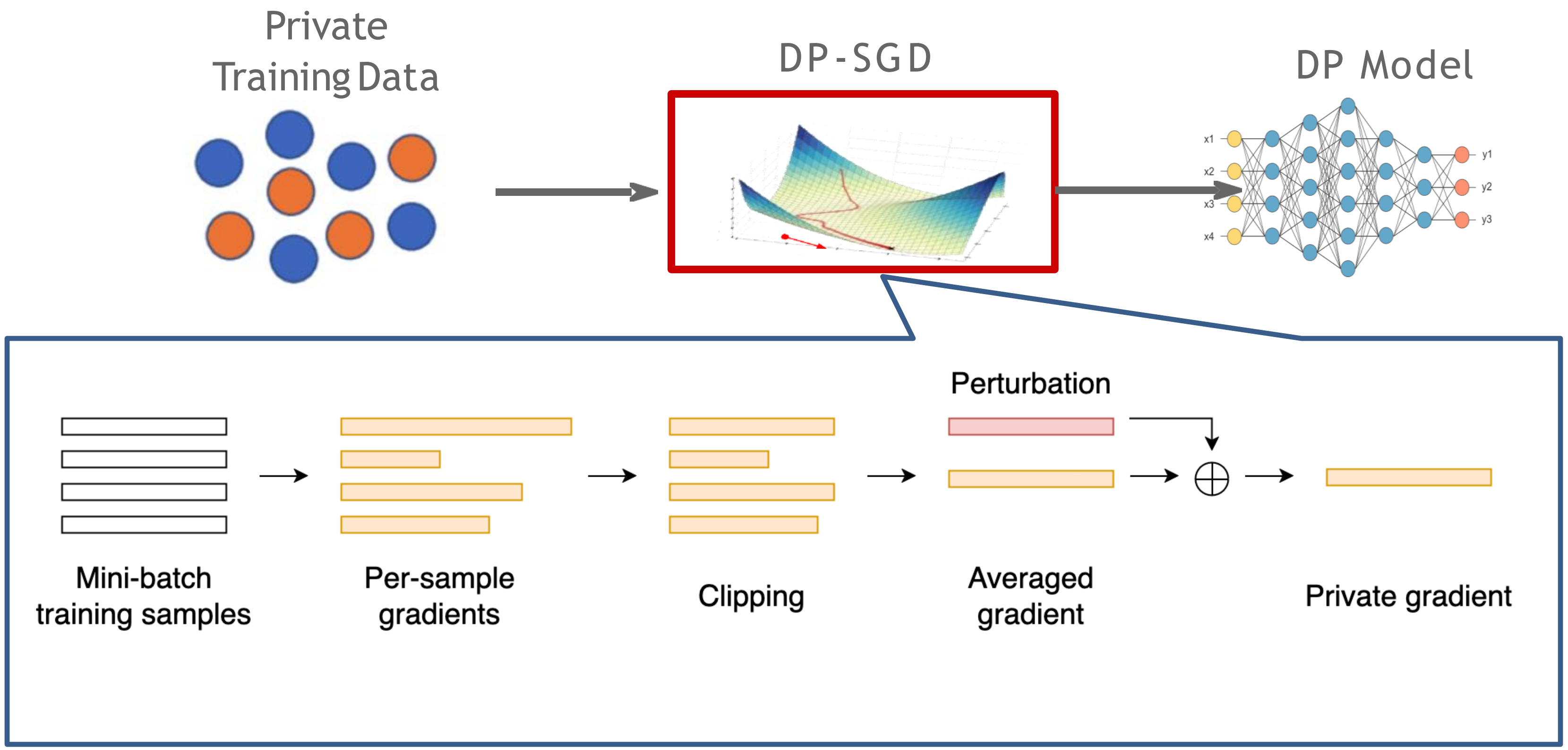
A randomized algorithm $\mathcal{A}$ is (ϵ, δ) -DP for two neighboring datasets D, D' and all $S \subseteq Range(\mathcal{A})$

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{A}(D') \in S] + \delta$$

[Dwork'06] Cynthia Dwork. Differential privacy. *ICALP 2006*

Differential Privacy (DP) in practice

Private Training with Differential Privacy: DP-SGD



- DP-SGD: commonly used algorithm for training DP models
- Open challenges: improving privacy/utility tradeoffs, scalability, privacy auditing, interpretability

[Abadi '16] Deep learning with differential privacy, M Abadi, A Chu, I Goodfellow, HB McMahan, I Mironov, K Talwar, L Zhang. CCS 2016

[Cummings '24]] Rachel Cummings, David Evans, Damien Desfontaines, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, Nicolas Papernot, Ryan Rogers, Milan Shen, Shuang Song, Weijie Su, Andreas Terzis, Abhradeep Thakurta, Sergei Vassilvitskii, Yu-Xiang Wang, Li Xiong, Sergey Yekhanin, Da Yu, Huanyu Zhang, Wanrong Zhang. Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. Harvard Data Science Review (**HDSR**), 2024

Privacy in the Age of AI and LLMs: Outline

- Privacy Attacks
- Privacy Defenses
  - Overview
  - **Data synthesization (pre-training)**
  - Privacy enhanced training
  - Machine unlearning (post-training)
  - Case studies for healthcare
- Open challenges

EMORY UNIVERSITY

Differentially private tabular data generation

- Goal: generate synthetic data that mimics the privacy-sensitive data for downstream tasks

- Methods:
  - DP Statistical methods
  - DP Generative Adversarial Networks (GAN) based methods
  - DP Marginal based methods

Sensitive Tabular Data

	Age	Gender	Education	Income
Training Data	42	male	bachelor	>50K
Training Data	18	female	high school	<=50K
Training Data	24	female	master	>50K
Testing Data	45	male	high school	>50K
Testing Data	40	female	bachelor	>50K

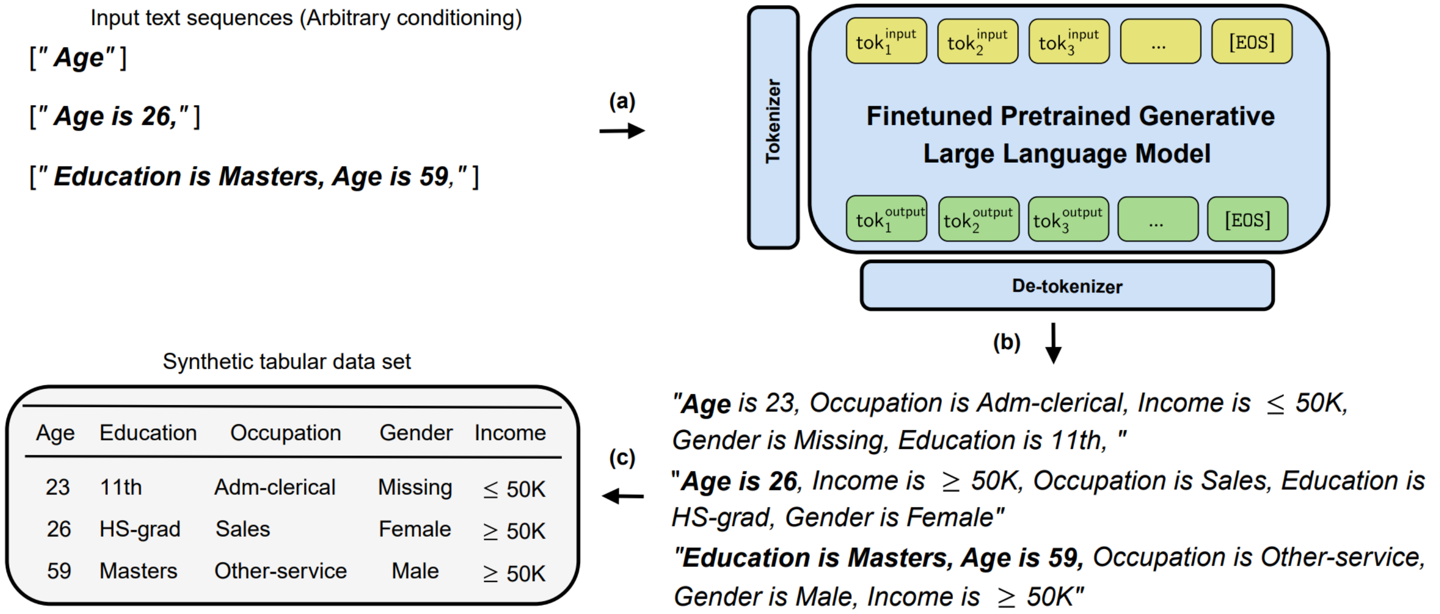
→

Random Tabular Data

Age	Gender	Education	Income
20	male	high school	>50K
26	female	doctorate	<=50K
27	male	master	<=50K

LLMs are realistic tabular data generators

- Borisov el at. [ICLR 2023] converted tabular data to prompts and fine tuned LLMs. Then, the synthetic tabular data were generated by sampling the fine tuned LLMs.

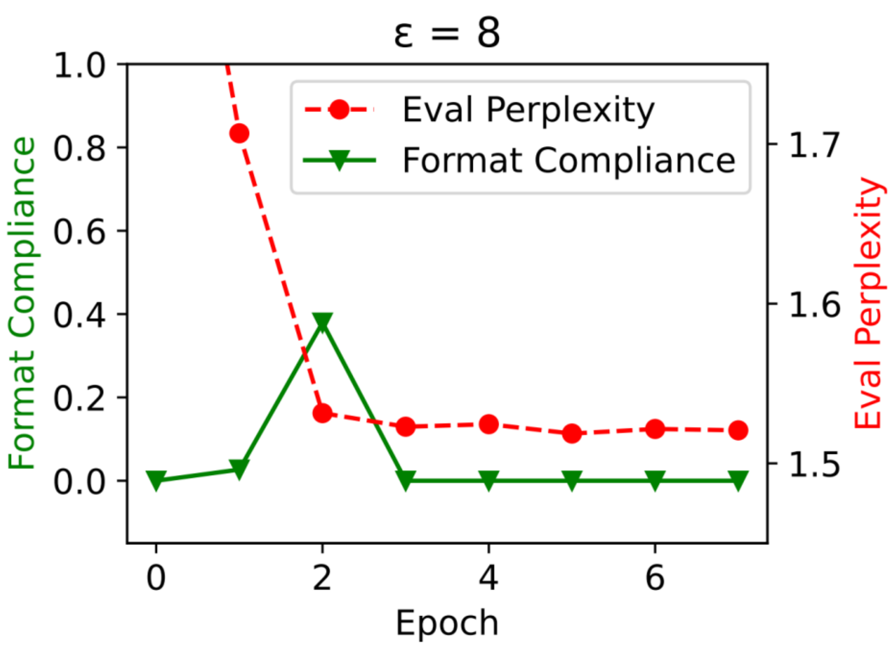


Outputs have to be in a consistent format

Synthetic data might memorize the original data!

Applying DP directly does not work

- Enhancing [Borisov et al., ICLR 2023] by replacing SGD by DPSGD for finetuning
- Fails at format compliance (i.e., LLM outputs have wrong categorical values, wrong column names, missing columns)



DP-LLMTGen

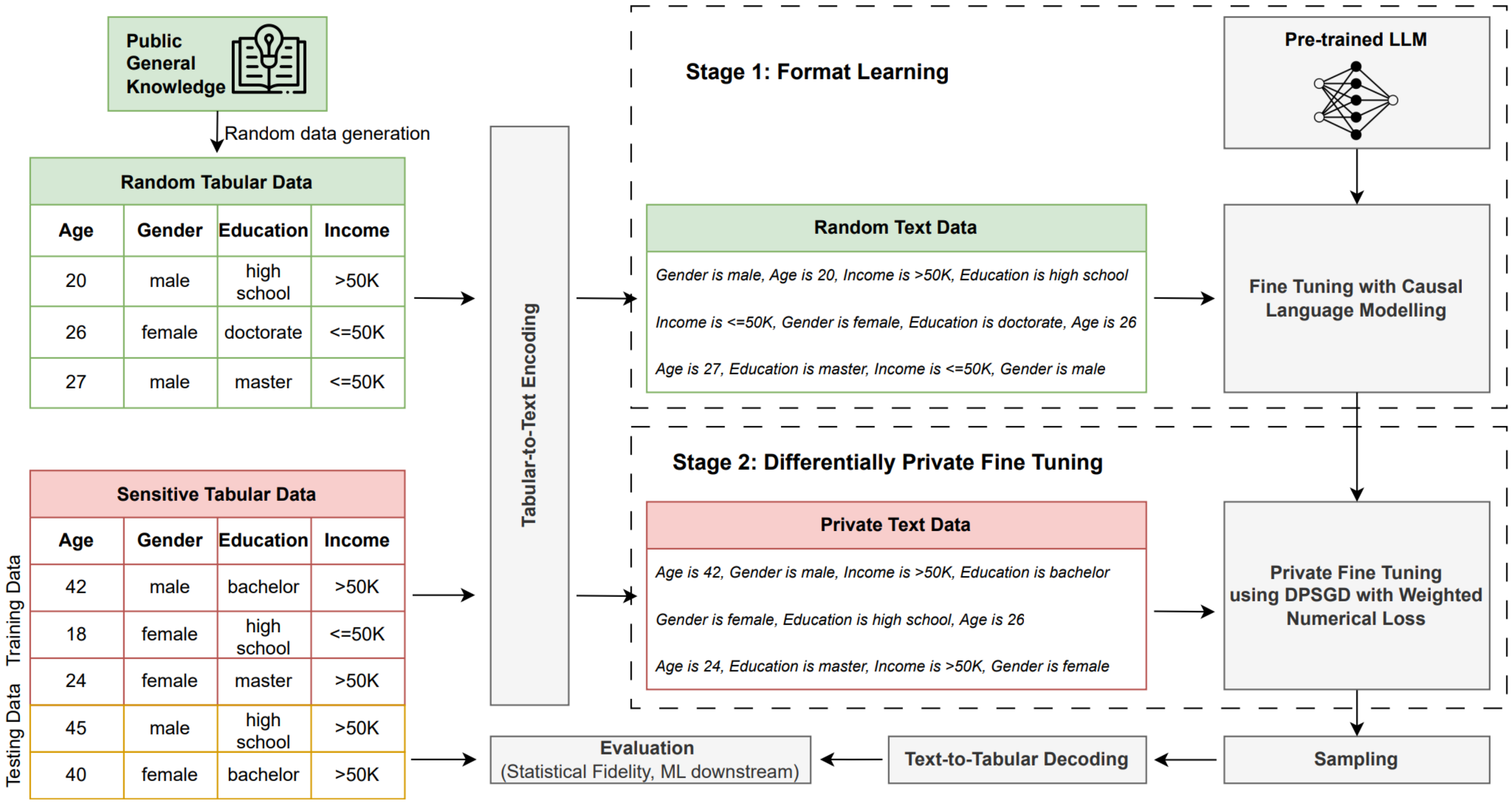


Figure 1: The process flow of DP-LLMTGen

- Two-stage fine tuning separates Format Learning & Data Learning
- Novel Loss Function to improve performance of data learning

Insight: leveraging pretraining or public data helps improve privacy and utility tradeoff

Results

Dataset	Privacy Budget	Method	Statistical Fidelity (TVD) (↓) 1-way	2-way	3-way	4-way	5-way	Xgboost (↑) ACC	AUC
Bank	-	*Training Set*	*0.009±0.001*	*0.021±0.001*	*0.043±0.001*	*0.076±0.002*	*0.121±0.002*	*0.853±0.009*	*0.924±0.005*
	$\epsilon = 0.5$	DP-GAN	0.196±0.019	0.302±0.023	0.368±0.023	0.411±0.021	0.440±0.017	0.528±0.010	0.547±0.032
		DP-CTGAN	0.175±0.017	0.266±0.019	0.327±0.019	0.374±0.018	0.411±0.016	0.497±0.024	0.509±0.042
		PATE-CTGAN	0.200±0.003	0.291±0.004	0.351±0.004	0.400±0.003	0.442±0.002	0.420±0.061	0.331±0.027
		RAP	0.165±0.006	0.268±0.005	0.345±0.003	0.404±0.002	0.446±0.001	0.636±0.006	**0.699±0.005**
		RAP++	0.071±0.005	0.129±0.005	0.185±0.005	0.240±0.005	0.293±0.006	0.635±0.068	0.697±0.062
		GSD	0.166±0.001	0.245±0.002	0.302±0.001	0.354±0.002	0.401±0.001	0.635±0.017	0.692±0.017
		DP-LLMTGen	**0.050±0.003**	**0.091±0.003**	**0.134±0.003**	**0.181±0.004**	**0.232±0.004**	**0.637±0.007**	0.669±0.032
	$\epsilon = 1.0$	DP-GAN	0.244±0.013	0.357±0.013	0.418±0.013	0.453±0.011	0.473±0.009	0.460±0.071	0.396±0.056
		DP-CTGAN	0.157±0.019	0.244±0.020	0.303±0.022	0.347±0.023	0.382±0.023	0.499±0.025	0.518±0.214
		PATE-CTGAN	0.194±0.007	0.287±0.008	0.350±0.008	0.402±0.007	0.445±0.006	0.533±0.027	0.583±0.111
		RAP	0.159±0.005	0.263±0.003	0.341±0.002	0.400±0.001	0.443±0.001	0.655±0.008	0.705±0.006
		RAP++	0.058±0.007	0.105±0.009	0.151±0.010	0.201±0.011	0.252±0.011	**0.674±0.017**	**0.747±0.019**
		GSD	0.166±0.001	0.244±0.002	0.301±0.001	0.353±0.002	0.400±0.001	0.671±0.041	0.726±0.032
		DP-LLMTGen	**0.047±0.010**	**0.085±0.012**	**0.125±0.011**	**0.169±0.010**	**0.218±0.009**	0.638±0.004	0.703±0.008
Adult	-	*Training Set*	*0.004±0.001*	*0.011±0.001*	*0.024±0.001*	*0.044±0.001*	*0.073±0.001*	*0.872±0.002*	*0.928±0.001*
	$\epsilon = 0.5$	DP-GAN	0.201±0.013	0.316±0.015	0.386±0.014	0.429±0.011	0.456±0.008	0.765±0.010	0.692±0.166
		DP-CTGAN	0.151±0.012	0.239±0.012	0.299±0.010	0.343±0.007	0.378±0.005	0.759±0.001	0.606±0.089
		PATE-CTGAN	0.262±0.005	0.365±0.004	0.425±0.004	0.464±0.003	0.488±0.002	0.518±0.248	0.379±0.146
		RAP	0.152±0.001	0.253±0.002	0.326±0.002	0.382±0.002	0.425±0.002	0.791±0.003	0.819±0.004
		RAP++	0.066±0.003	0.116±0.004	0.162±0.004	0.208±0.006	0.251±0.008	0.766±0.012	0.758±0.023
		GSD	0.245±0.001	0.337±0.000	0.395±0.000	0.438±0.000	0.470±0.000	0.759±0.001	0.768±0.019
		DP-LLMTGen	**0.058±0.008**	**0.095±0.010**	**0.126±0.011**	**0.154±0.011**	**0.183±0.010**	**0.833±0.005**	**0.887±0.003**
	$\epsilon = 1.0$	DP-GAN	0.259±0.112	0.370±0.091	0.427±0.058	0.460±0.034	0.479±0.018	0.761±0.002	0.589±0.126
		DP-CTGAN	0.178±0.026	0.277±0.027	0.339±0.024	0.380±0.021	0.409±0.018	0.550±0.272	0.597±0.110
		PATE-CTGAN	0.256±0.002	0.362±0.001	0.423±0.001	0.463±0.001	0.487±0.000	0.589±0.295	0.486±0.071
		RAP	0.148±0.000	0.248±0.001	0.322±0.001	0.378±0.001	0.421±0.001	0.801±0.006	0.831±0.005
		RAP++	0.054±0.002	0.099±0.003	0.142±0.003	0.183±0.002	0.225±0.002	0.803±0.003	0.807±0.017
		GSD	0.245±0.001	0.337±0.000	0.395±0.000	0.438±0.000	0.470±0.000	0.759±0.001	0.778±0.006
		DP-LLMTGen	**0.038±0.002**	**0.068±0.002**	**0.096±0.003**	**0.124±0.003**	**0.153±0.002**	**0.831±0.003**	**0.879±0.003**

Privacy in the Age of AI and LLMs: Outline

- Privacy Attacks
- Privacy Defenses
  - Overview
  - Data synthesization (pre-training)
  - **Privacy enhanced training**
  - Machine unlearning (post-training)
  - Case studies for healthcare
- Open challenges

EMORY UNIVERSITY

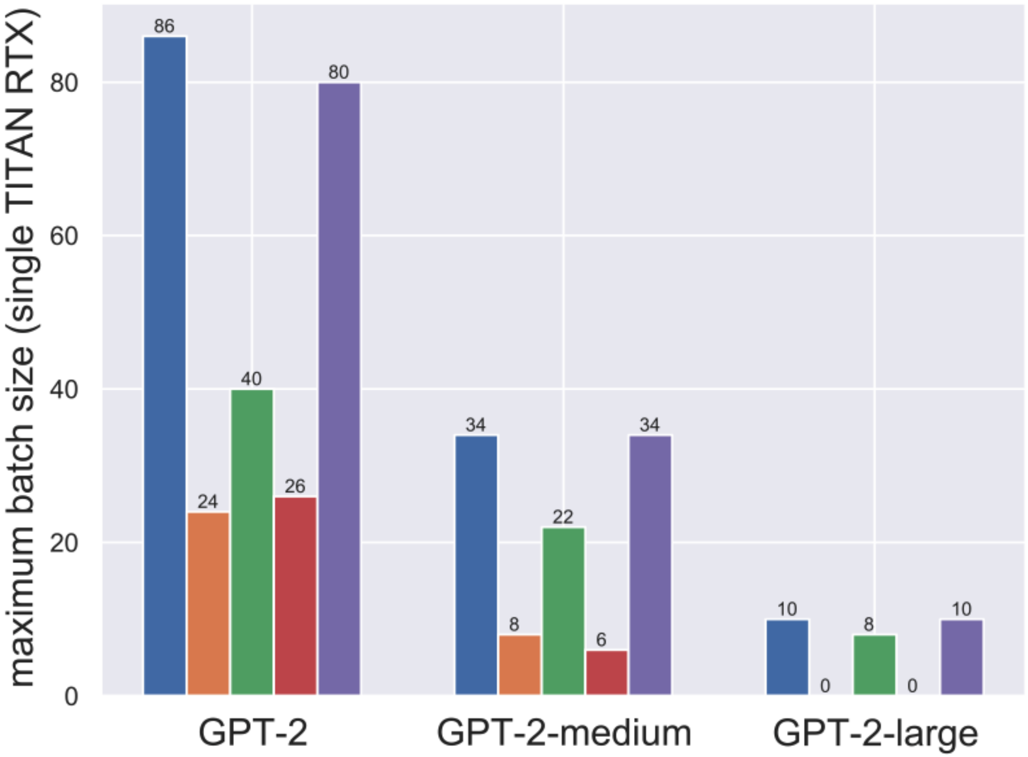
Privacy Enhanced Training

- Theoretical Defense
  - Differentially private training
- Empirical Defense
  - Regularization-based methods
    - Drop-out
    - Gradient Clipping
    - Weight Decay

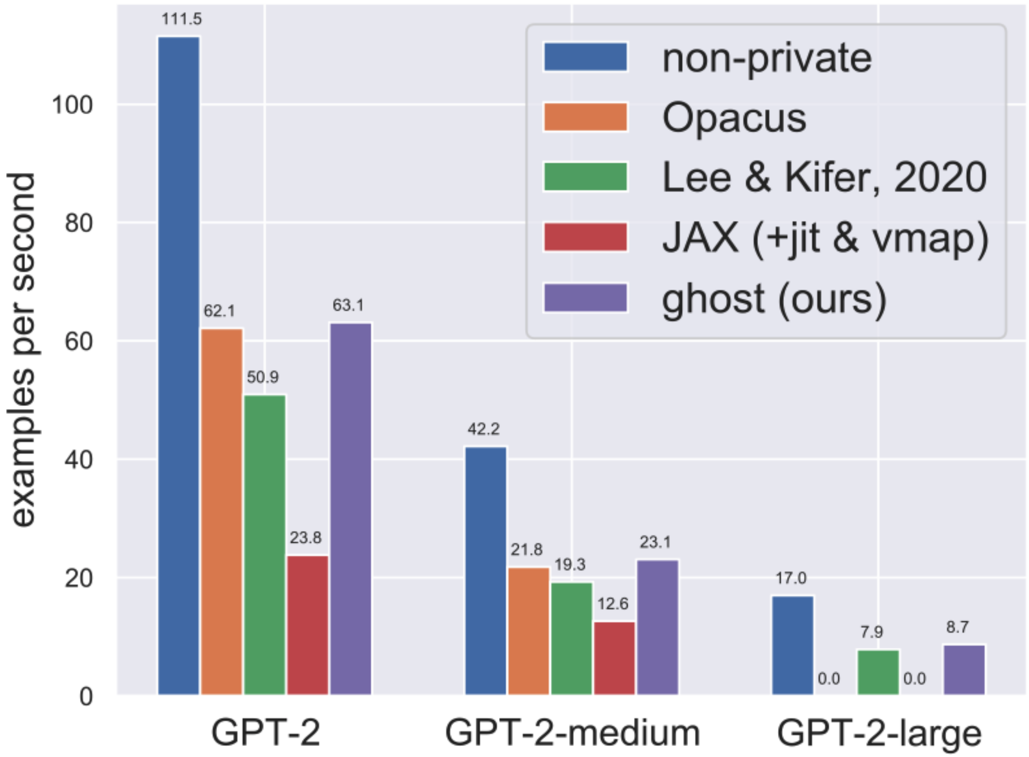
EMORY UNIVERSITY

Privacy Comes at a Cost

- A single run of DP training requires more computation than NonDP



(a) Memory

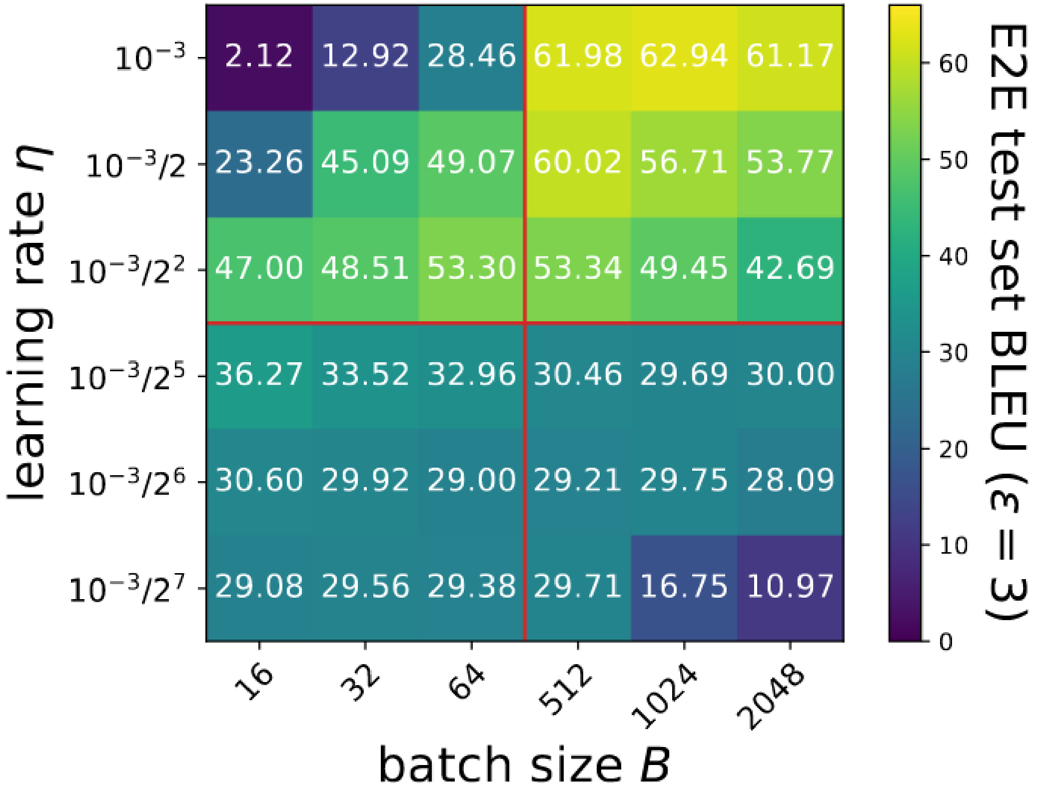


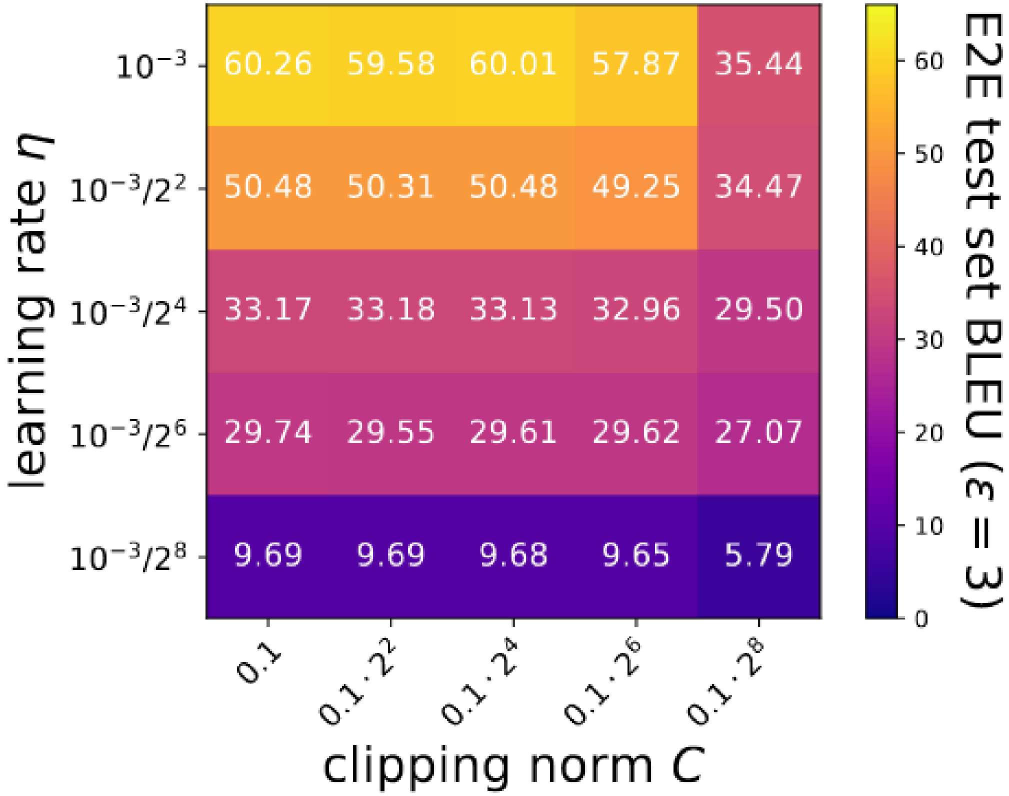
(b) Throughput

[li2022]: Large language models can be strong differentially private learners, 2022ICLR

Privacy Comes at a Cost

- Performance of DP training is sensitive to hyper-parameter tuning
- DP training requires extra hyper-parameter tuning (e.g., η, C)





[li2022]: Large language models can be strong differentially private learners, 2022ICLR

Automatic Learning Rate Scheduler in Non-DP Setting

Estimate the learning rate $\eta \approx \frac{D}{G\sqrt{T}}$ [PMLR'23][NeurIPS'23]

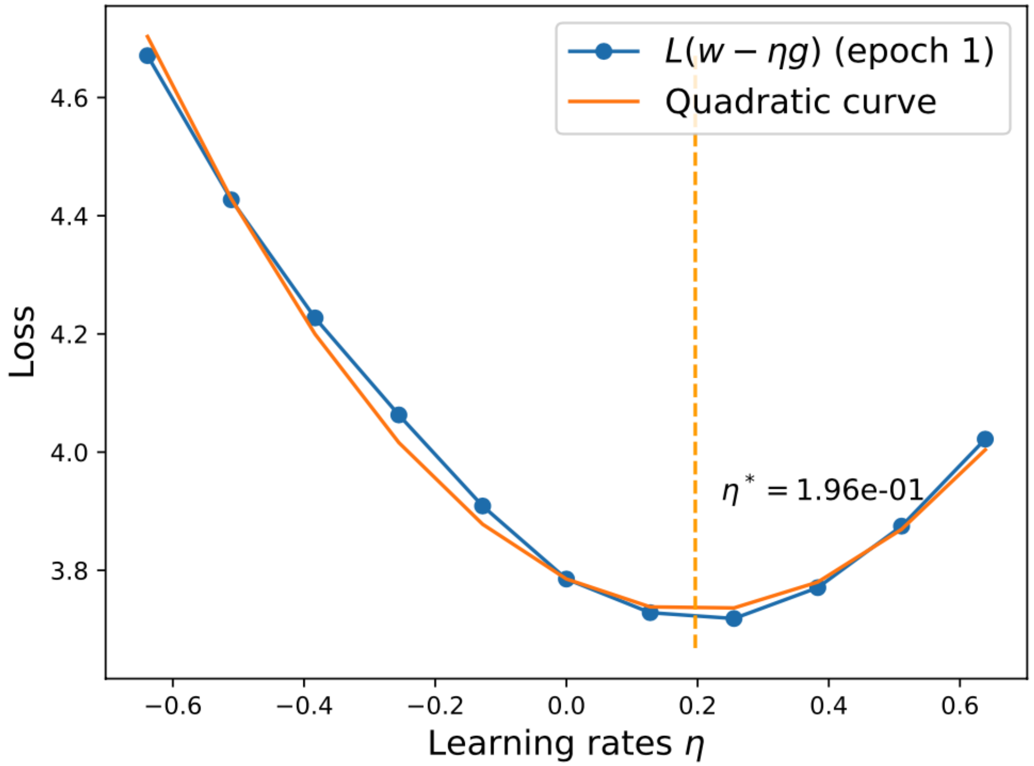
- $D = ||w_0 - w_*||$ is the initialization-to-minimizer distance
- G is the Lipschitz continuity constant
- T is the total number of iterations

Directly apply in DPSGD? ➡ **Inaccurate estimation due to noise in privatized gradient!**

Automatic Learning Rate Scheduler in Non-DP Setting

Estimate η via a Taylor approximation [GeN]

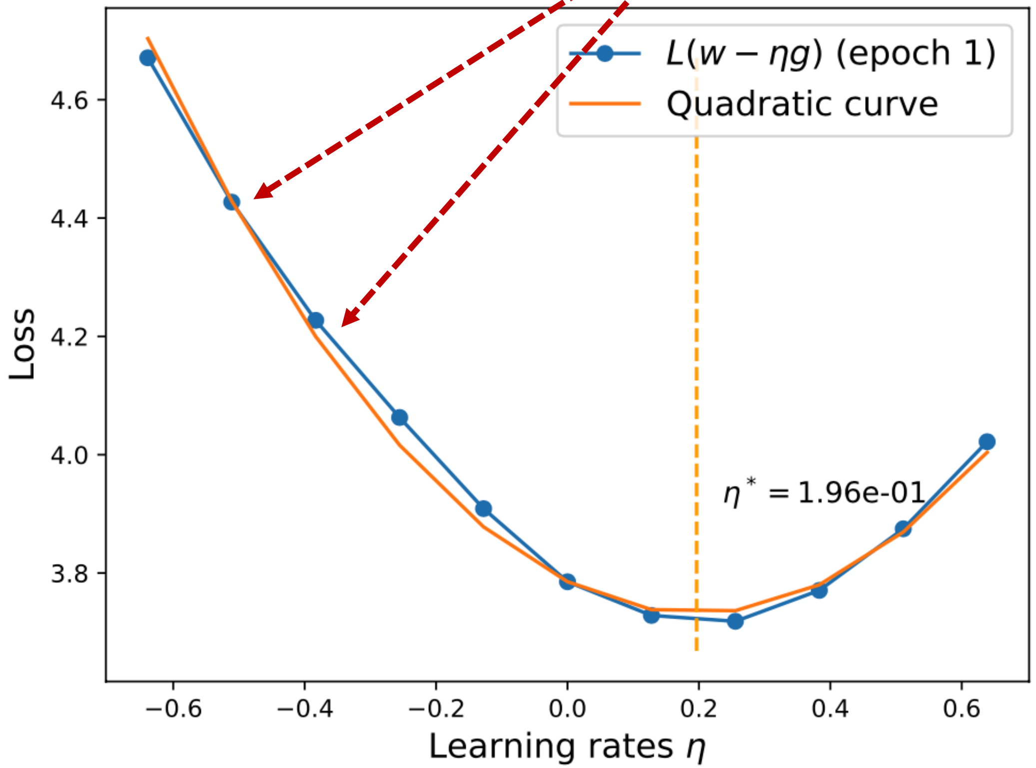
- Without assuming the Lipschitz continuity or the knowledge of D
- Try updates with different learning rate
- Fit the curve as a quadratic function for finding the optimal η



HyFreeDP: Privatize the Automatic Learning Rate

Privatizing a single-dimension value is DP-friendly!

① Privatize each loss value in the curve!



② Curve fitting

$$(\boldsymbol{m}^\top \tilde{\mathbf{H}} \boldsymbol{m})_{\text{DP}}, (\tilde{\mathbf{G}}^\top \boldsymbol{m})_{\text{DP}}$$

③ Estimate the learning rate

$$\eta_{\text{GeN-DP}} = \frac{(\tilde{\mathbf{G}}^\top \boldsymbol{m})_{\text{DP}}}{(\boldsymbol{m}^\top \tilde{\mathbf{H}} \boldsymbol{m})_{\text{DP}}}$$

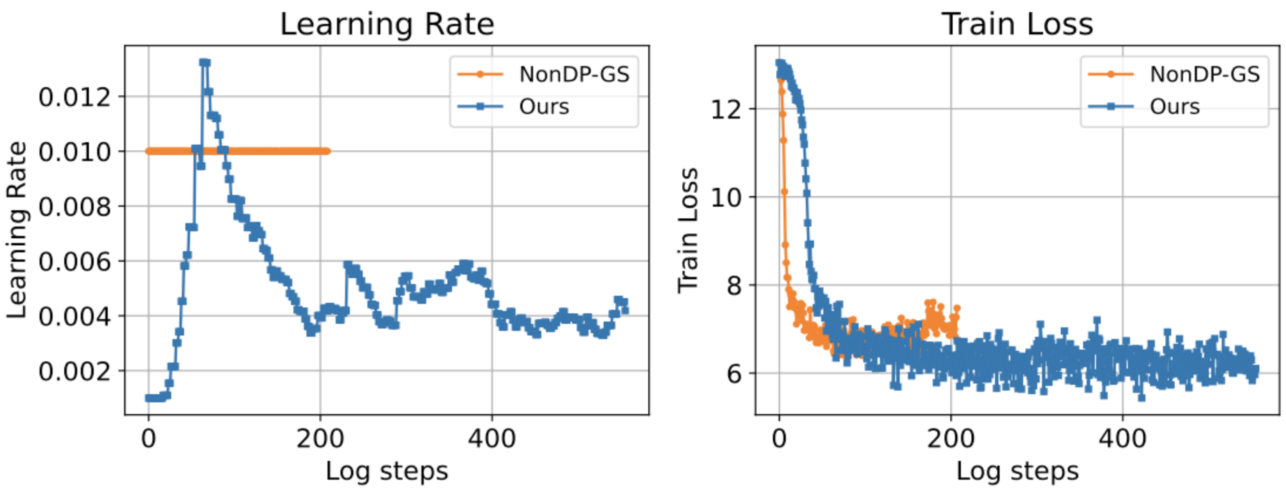
Ruixuan Liu, Zhiqi Bu, Towards hyperparameter-free optimization with differential privacy, ICLR 2025

Text Generation

HyFreeDP approaches to the best single-run DP (NonDP-GS)

- E2E dataset with GPT2
- PubMed dataset with Llama2-7B

A consistent automatic learning rate trend for vision and language tasks!



Full Fine-Tune		$\epsilon = 3$					$\epsilon = 8$				
Model	Method	BLEU	CIDEr	METEOR	NIST	ROUGE_L	BLEU	CIDEr	METEOR	NIST	ROUGE_L
GPT-2	NonDP-GS	0.583	1.566	0.367	5.656	0.653	0.612	1.764	0.385	6.772	0.664
	D-Adaptation	0.000	0.000	0.003	0.082	0.016	0.000	0.000	0.000	0.000	0.000
	Prodigy	0.082	0.000	0.157	1.307	0.239	0.012	0.000	0.003	0.000	0.003
	HyFreeDP	**0.585**	**1.564**	**0.365**	**5.736**	**0.636**	**0.612**	**1.768**	**0.378**	**6.702**	**0.655**

Ruixuan Liu, Zhiqi Bu, Towards hyperparameter-free optimization with differential privacy, ICLR 2025

Efficiency Evaluation

Environment Setups

- A100 (80G) for Llama2-7B
- Titan RTX (24G) for others

HyFreeDP only has a minor computation cost compared to a single DP run!

- Smaller models spend less computation cost on tuning than larger models

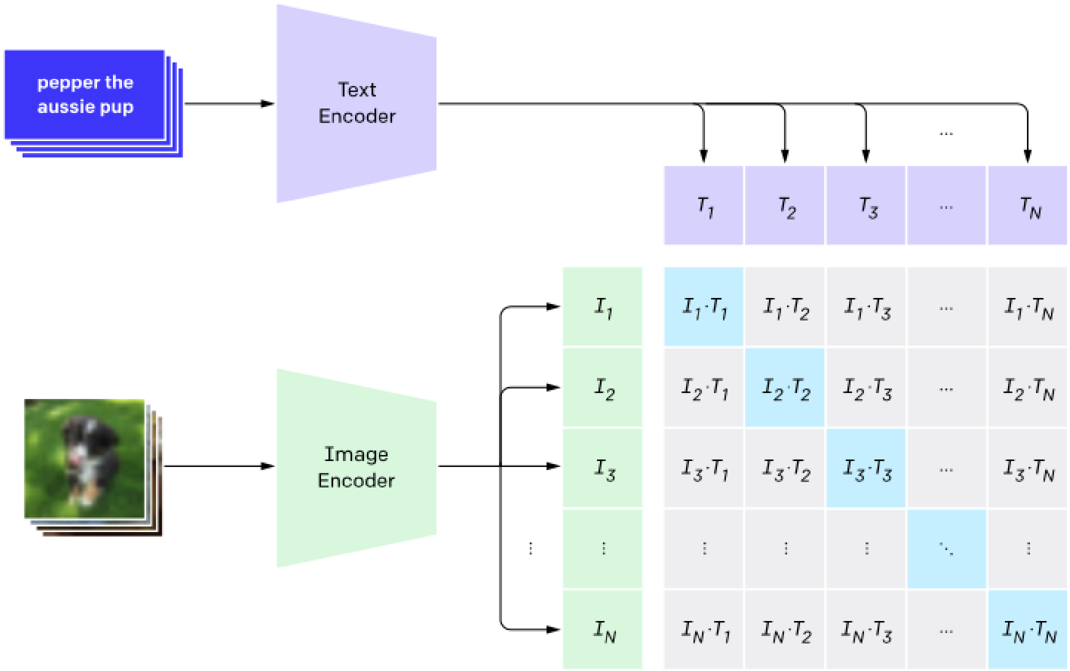
Table 4: Comparison of model performance in minutes (mins) with and without auto-tuning across various datasets. The coefficients represent the ratio relative to the w/o auto configuration.

Models	Dataset	K=1	K=5	K=10	w/o auto
llama2-7B (LoRA-FT)	PubMed (4k)	409.750 mins (× 2.040)	244.333 mins (× 1.217)	222.583 mins (× 1.108)	200.833 mins (× 1.000)
GPT2	E2E	163.333 mins (× 1.888)	97.167 mins (× 1.123)	94.983 mins (× 1.098)	86.500 mins (× 1.000)
Vit-base	CIFAR100	152.617 mins (× 1.370)	118.483 mins (× 1.063)	113.317 mins (× 1.017)	111.433 mins (× 1.000)
Vit-base (BitFit-FT)	CIFAR100	113.450 mins (× 1.654)	74.733 mins (× 1.089)	73.817 mins (× 1.076)	68.600 mins (× 1.000)
Vit-small	SVHN	102.000 mins (× 1.255)	84.500 mins (× 1.040)	82.800 mins (× 1.019)	81.250 mins (× 1.000)

Ruixuan Liu, Zhiqi Bu, Towards hyperparameter-free optimization with differential privacy, ICLR 2025

DP Training for Multi-Modal Models

- DP for CLIP models
  - Contrastive training requires batch clipping, which hurts utility



Algorithm 1 DP-CLIP

Input: observed pairs of data $\{(x_i, \tilde{x}_i)\}_{i=1}^n$, number of iterations T, noise scale σ, clipping threshold c, learning rate η, mini-batch size b, initial parameters $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$.

for $t \in \{0, \ldots, T-1\}$ **do**

Sample mini-batch:

Uniformly sample $\mathcal{B}^{(t)}$ of size b from $[n]$.

Compute mini-batch gradient:

$g^{(t)} \leftarrow \partial_\theta \mathcal{L}(f_{\theta_1^{(t)}}, \tilde{f}_{\theta_2^{(t)}}; \mathcal{B})$.

Clip Gradient:

$\bar{g}^{(t)} \leftarrow \min\{1, c/\|g^{(t)}\|_F\} g^{(t)}$.

Add Noise:

$\tilde{g}^{(t)} \leftarrow \bar{g}^{(t)} + \sigma c \mathcal{N}(0, I)$.

Descent:

$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \tilde{g}^{(t)}$.

end for

Return $\theta^{(T)}$.

Huang et al., Safeguarding Data in Multimodal AI: A Differentially Private Approach to CLIP Training, 2023

Privacy Enhanced Training

- Theoretical Defense
  - Differentially private training
- Empirical Defense
  - Regularization-based methods
    - Drop-out
    - Gradient Clipping
    - Weight Decay

EMORY UNIVERSITY

Tokens for Learning, Tokens for Unlearning: Mitigating Membership Inference Attacks in Large Language Models via Dual-Purpose Training

Toan Tran, Ruixuan Liu, Li Xiong

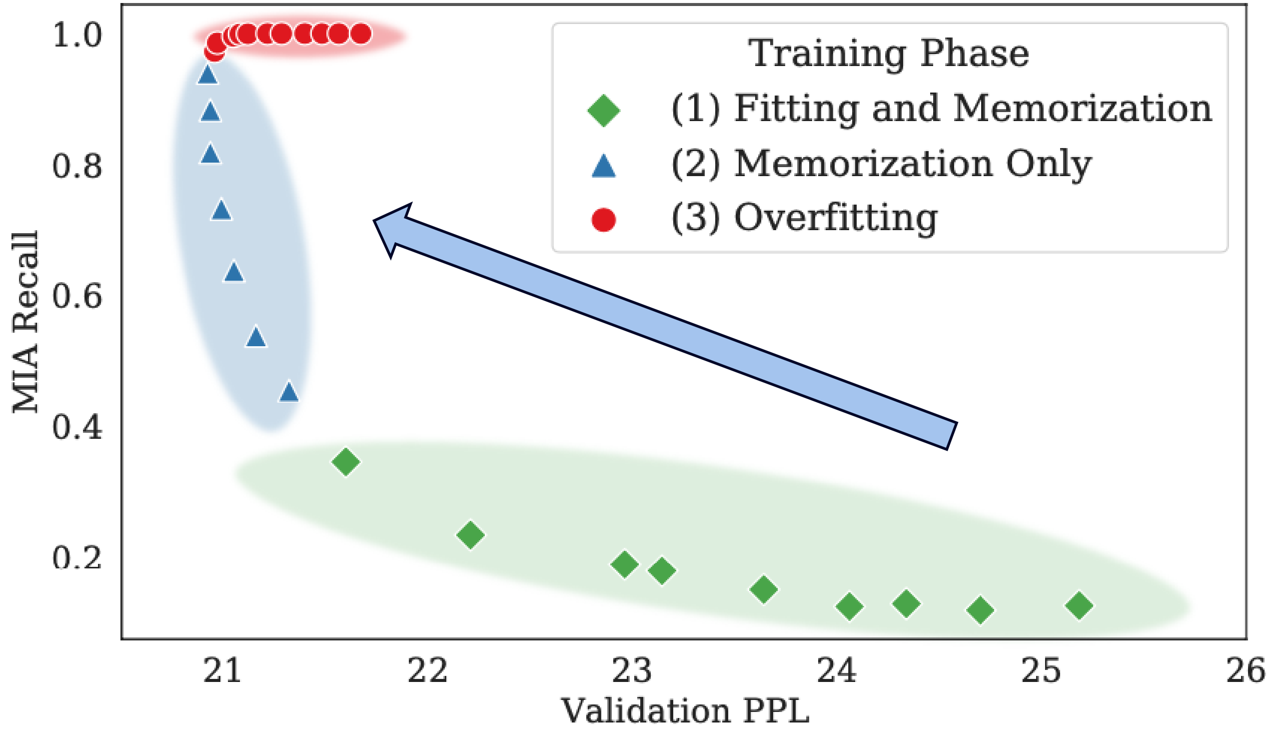
Emory University

EMORY UNIVERSITY

Association of Computer Linguistics (ACL) Findings, 2025

Intuition: Training Phases of Models

- **Model performance** and **privacy risk** evolve differently
  - Fitting and memorization (initial phase)
  - Memorization only (mid to late training)
  - Overfitting (final stage)



Training phase in fine-tuning[EMNLP'22]

How do we learn without memorizing?

[EMNLP'22]: An Empirical Analysis of Mem in Fine-tuned Autoregressive Language Models, 2022 EMNLP

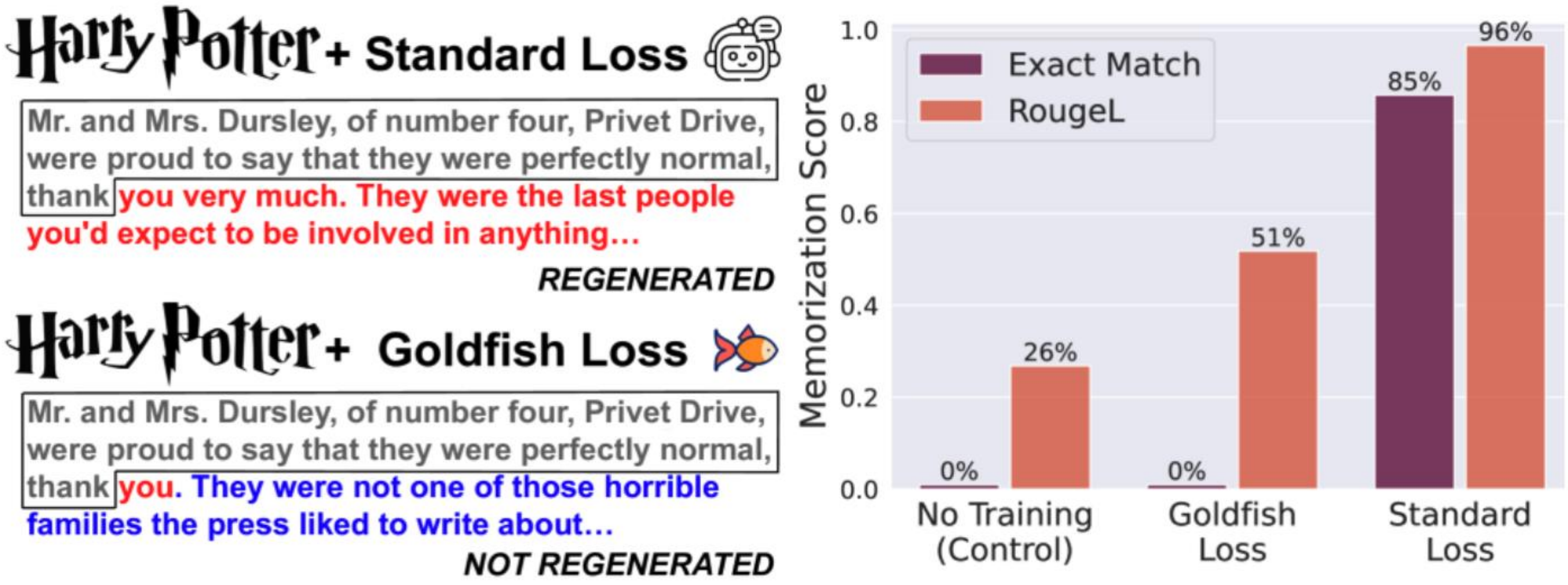
Intuition: Training LLMs

- Standard causal language modeling (CLM) loss

$$\mathcal{L}(\theta) = -\frac{1}{L}\sum_{i=1}^{L} \log P(x_i \mid x_{<i}; \theta)$$

- Goldfish loss

$$\mathcal{L}_{\text{goldfish}}(\theta) = -\frac{1}{|G|}\sum_{i=1}^{L} G_i(x_i) \log P(x_i \mid x_{<i}; \theta)$$



Using a carefully selected subset of tokens during training can match the performance of using all tokens in language modeling while mitigating MIA

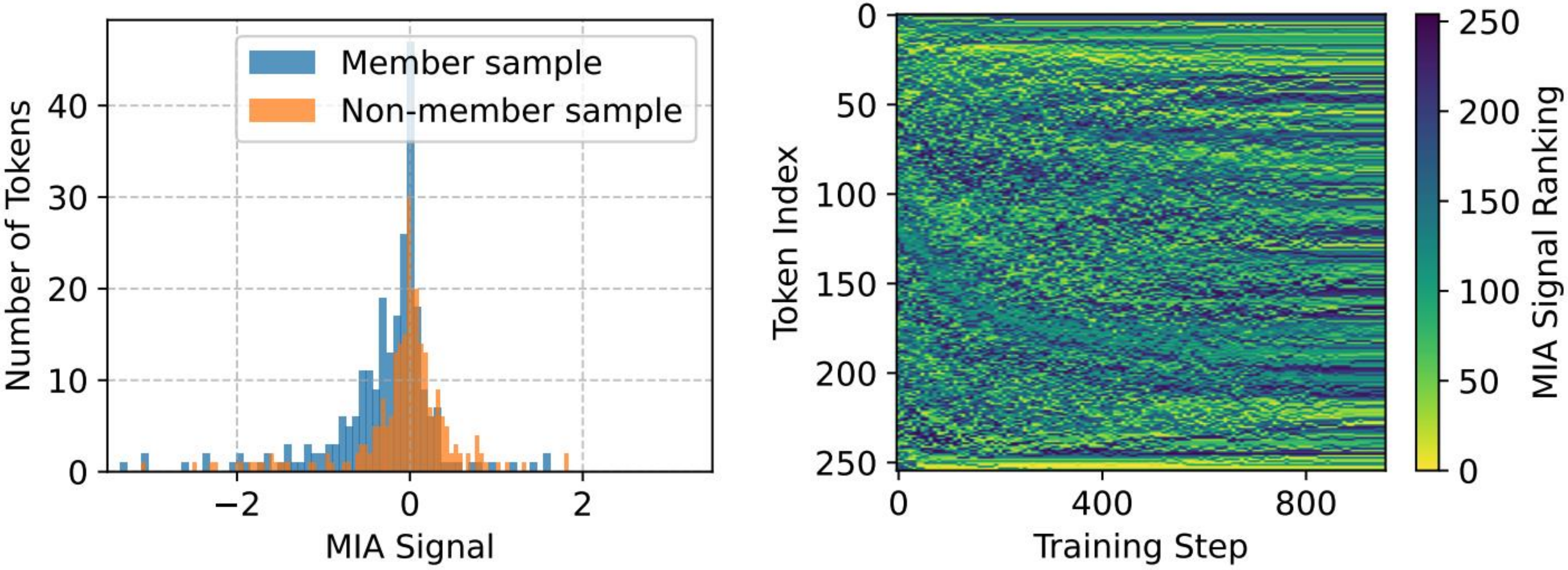
Do different tokens contribute to learning and MIA differently?

[NeurIPS 2024] Be like a goldfish, don't memorize! mitigating memorization in generative LLMs. NeurIPS 2024

EMORY UNIVERSITY

Token MIA Signal Analysis

Query Sample: <BOS> Order numbered AB89 has been shipped

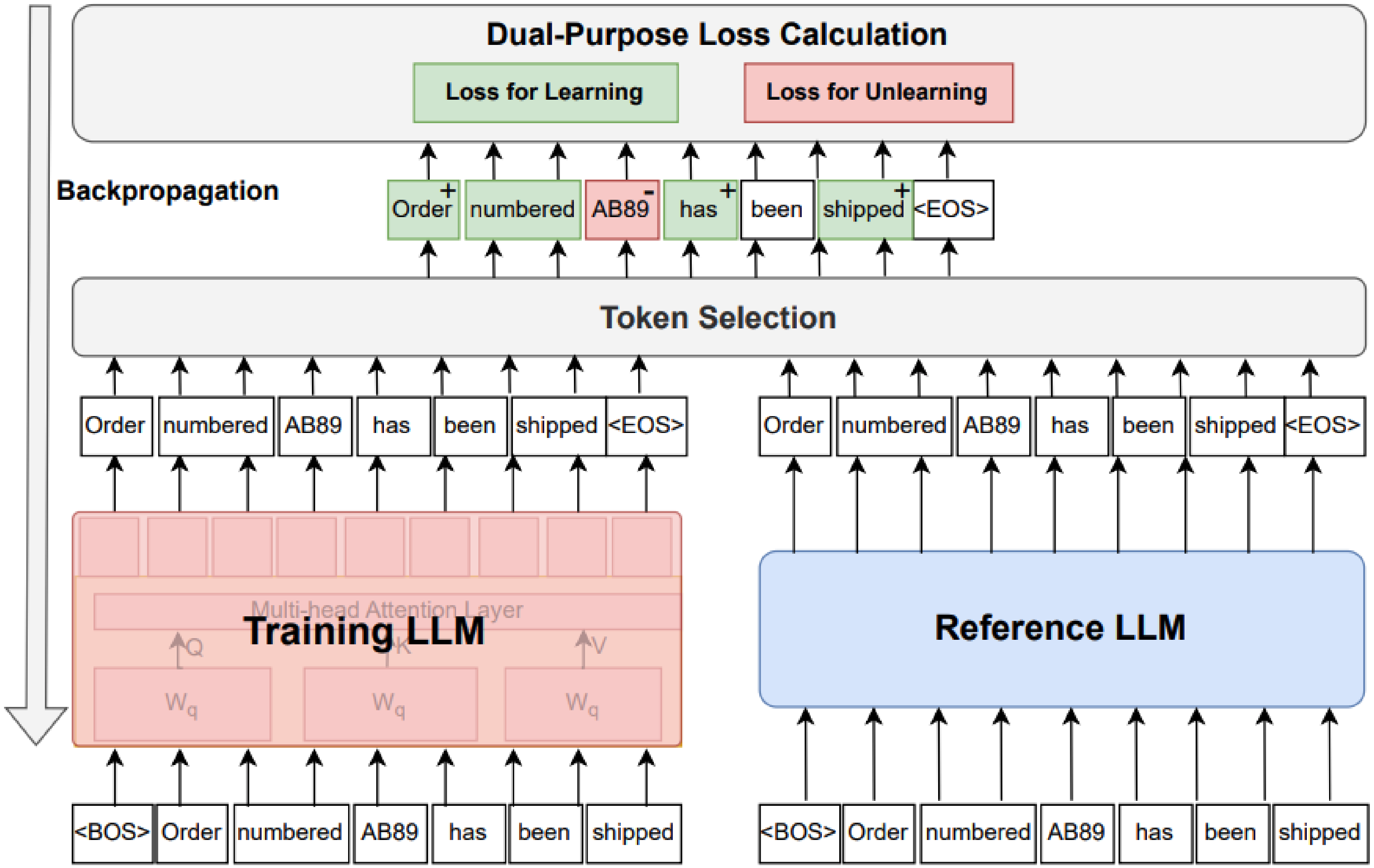


- Non-member samples have average MIA signal around zero
- Member samples have MIA signals negative side (smaller loss)
- MIA signals vary for different tokens and change dynamically

EMORY UNIVERSITY

DuoLearn: Duo-Purpose Training Framework

$$\mathcal{L}_{dual}(\theta) = \mathcal{L}_{CE}(\theta; \mathcal{T}_h) - \alpha \cdot \mathcal{L}_{CE}(\theta; \mathcal{T}_m)$$



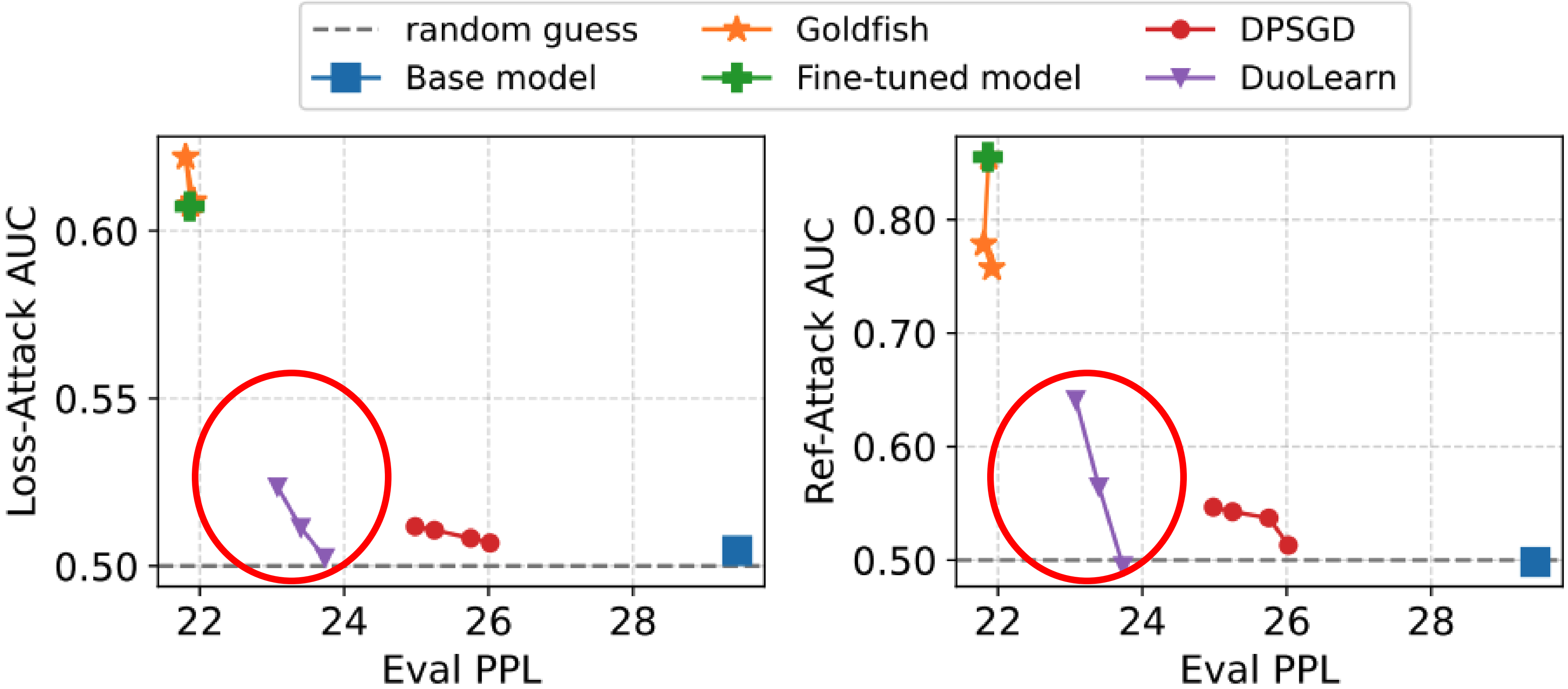
Evaluation: Overall Results

		Utility	MIA risk				Utility	MIA risk			
		Wikipedia					**CC-news**				
LLM	**Method**	PPL	Loss	Ref	Min-k	Zlib	PPL	Loss	Ref	Min-k	Zlib
GPT2 124M	*Base*	*34.429*	*0.473*	*0.513*	*0.446*	*0.497*	*29.442*	*0.505*	*0.498*	*0.520*	*0.500*
	FT	**12.729**	0.577	0.967	0.489	0.544	**21.861**	0.607	0.855	0.549	0.569
	Goldfish	12.853	0.565	0.954	0.486	0.537	21.902	0.608	0.855	0.547	0.570
	DPSGD	18.523	0.463	0.536	**0.448**	0.491	26.022	0.507	0.513	**0.521**	0.502
	DuoLearn	13.628	**0.454**	**0.463**	0.470	**0.485**	23.733	**0.502**	**0.495**	0.529	**0.499**
Pythia 1.4B	*Base*	*10.287*	*0.466*	*0.503*	*0.464*	*0.489*	*13.973*	*0.507*	*0.512*	*0.528*	*0.501*
	FT	**6.439**	0.578	0.985	0.484	0.557	11.922	0.602	0.857	0.541	0.574
	Goldfish	6.465	0.564	0.981	0.482	0.546	**11.903**	0.609	0.862	0.543	0.579
	DPSGD	7.751	0.469	0.524	**0.462**	0.488	13.286	0.512	0.531	0.528	0.503
	DuoLearn	6.553	**0.468**	**0.485**	0.472	**0.485**	12.670	**0.501**	**0.460**	**0.524**	**0.499**
Llama-2 7B	*Base*	*7.014*	*0.458*	*0.491*	*0.476*	*0.488*	*9.364*	*0.505*	*0.495*	*0.516*	*0.503*
	FT	**3.830**	0.524	0.936	0.494	0.530	**6.261**	0.559	0.798	0.536	0.548
	Goldfish	3.839	0.518	0.929	0.492	0.525	6.280	0.552	0.780	0.533	0.541
	DPSGD	4.490	0.466	0.516	**0.470**	0.487	6.777	0.509	0.538	0.523	0.504
	DuoLearn	4.006	**0.458**	**0.440**	0.473	**0.480**	6.395	**0.507**	**0.482**	**0.518**	**0.500**

- DuoLearn achieves comparable utility (PPL) and lowest MIA risk in most cases

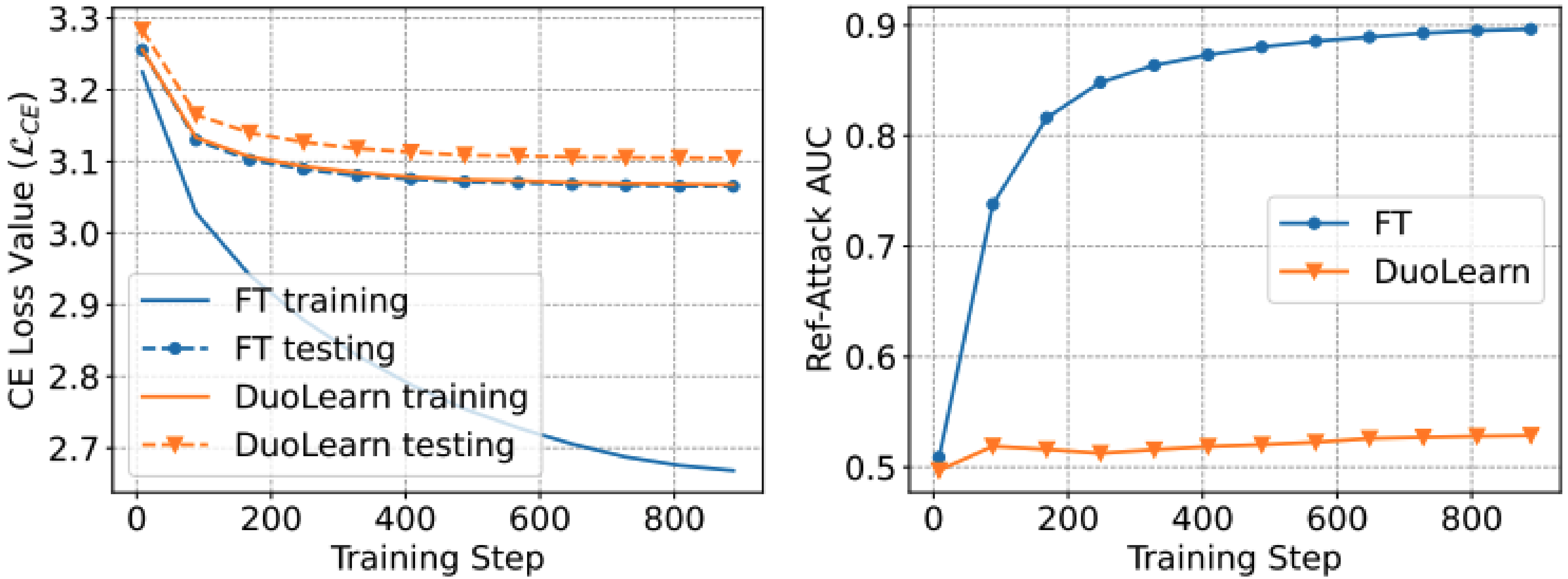
EMORY UNIVERSITY

Evaluation: Model Performance and Privacy Risk Tradeoff



- DuoLearn achieves a good tradeoff between good utility and low MIA risk

Evaluation: Memorization and Generalization Dynamics



- DuoLearn achieves good generalization (small training and test gap) and low MIA risk

EMORY UNIVERSITY

Evaluation: Defense against Privacy Backdoor PreCurious

	Metric	WU	FT	GF	DP	DuoL
Utility	**PPL**	*23.318*	21.593	**21.074**	23.279	22.296
MIA risk	**AUC**	*0.500*	0.911	0.886	0.533	**0.499**

- WU (warmed up pretrained model) acts like base model, low model performance, zero MIA risk
- Finetuned model has a very high (amplified) MIA risk
- DuoLearn significantly reduces the risk without sacrificing utility

EMORY UNIVERSITY

Privacy in the Age of AI and LLMs: Outline

- Privacy Attacks
- Privacy Defenses
  - Overview
  - Data synthesization (pre-training)
  - Privacy enhanced training
  - **Machine unlearning (post-training)**
  - Case studies for healthcare
- Open challenges

EMORY UNIVERSITY

Machine Unlearning

- Overview
- Exact Unlearning
- Approximate Unlearning
  - Contrastive unlearning
- Unlearning LLMs

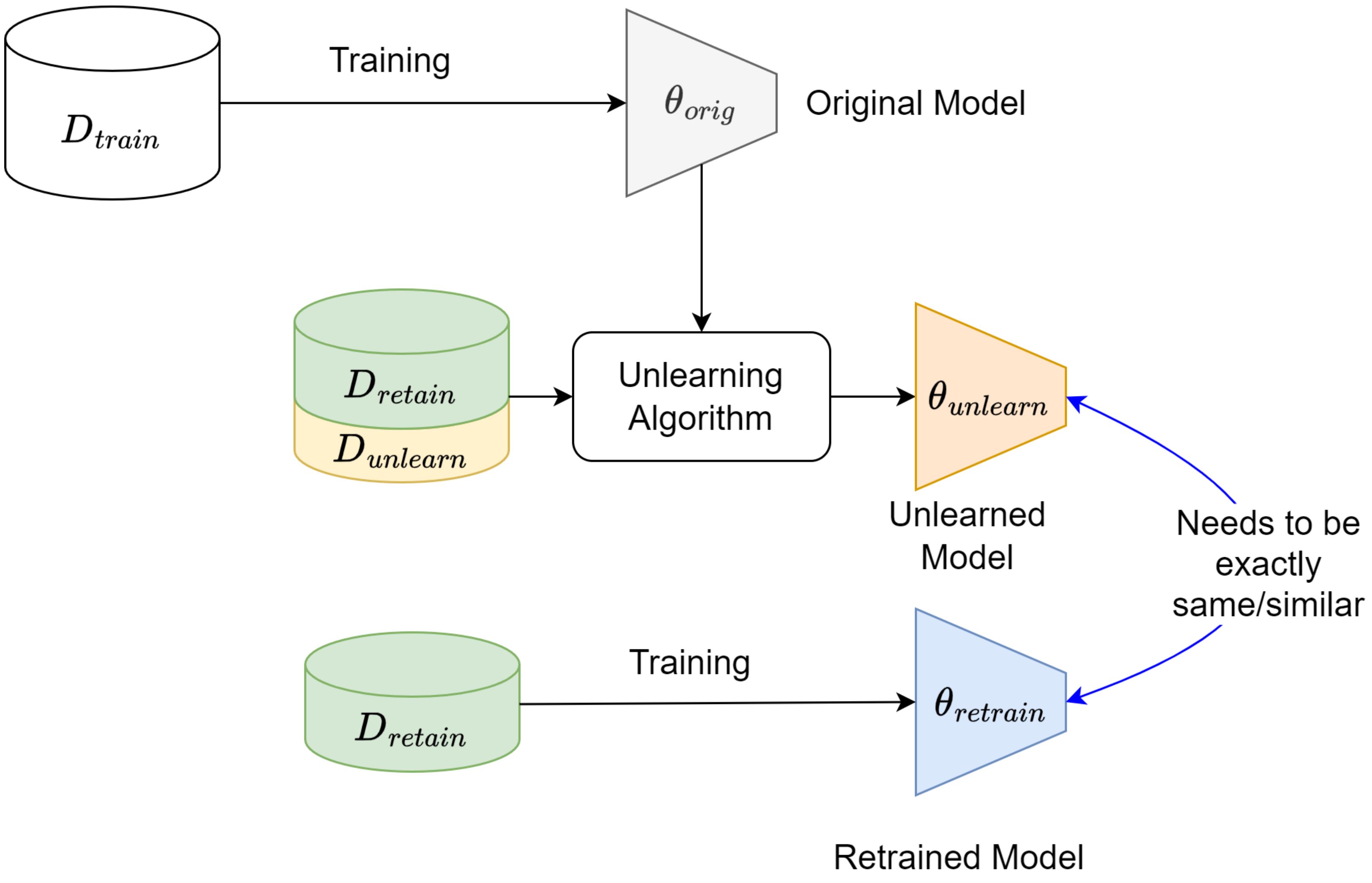
Legislations on Data Privacy

The right to be forgotten

- Individuals has a right to ask organizations a "**complete removal**" of their personal data.

- GDPR (EU)
- CCPA (California, USA)
- PIPDEA (Canada)

Machine Unlearning: Overview



Goal of unlearning

- Obtain a model that is **identical or similar** to a re-trained model without unlearning data

- Performance
  - Unlearned model should have similar performance to the re-trained model.
  - Test accuracy

- Unlearn Efficacy
  - Unlearned model needs to be similar to the re-trained model.
  - Membership Inference Attack
  - Mathematical Guarantee

- Efficiency
  - Unlearning algorithm should provide the unlearned model faster than re-tera

Machine Unlearning

- Overview
- **Exact Unlearning**
- Approximate Unlearning
  - Contrastive unlearning
- Unlearning LLMs

EMORY UNIVERSITY

In 42nd *IEEE Symposium of Security and Privacy*

Machine Unlearning

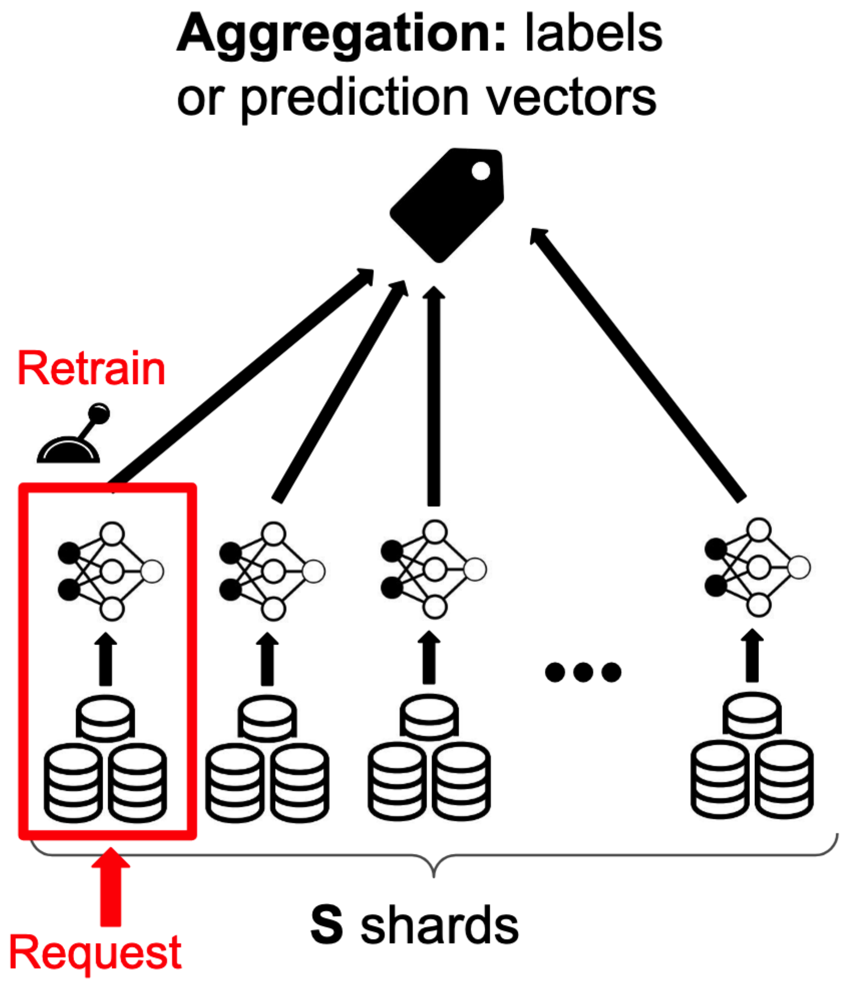
Lucas Bourtoule*‡§, Varun Chandrasekaran*†, Christopher A. Choquette-Choo*‡§, Hengrui Jia*‡§, Adelin Travers*‡§, Baiwu Zhang*‡§, David Lie‡, Nicolas Papernot‡§

University of Toronto‡, Vector Institute§, University of Wisconsin-Madison†

Exact unlearning: SISA

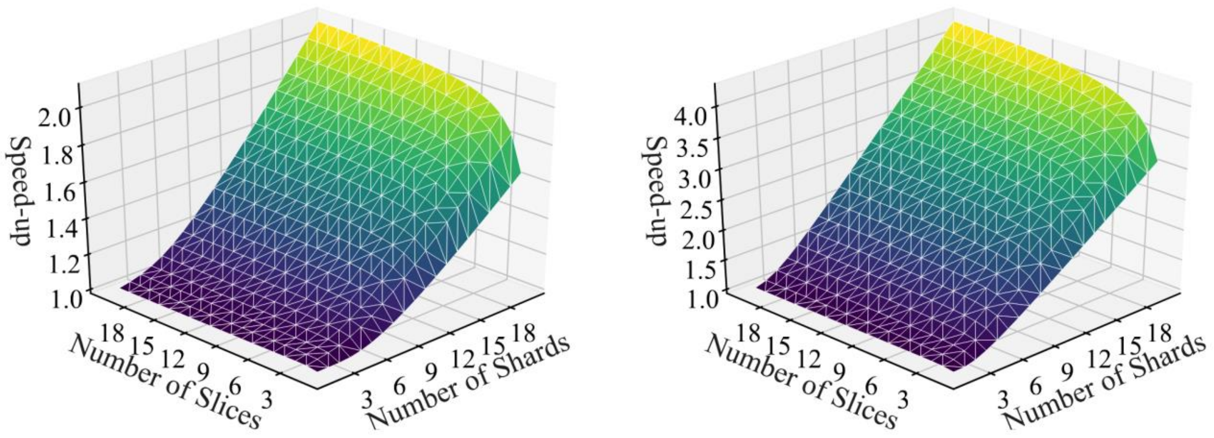
- **S**harded, **I**solated, **S**liced, **A**ggregated
- Sharding: Partition data
- Isolated: Each shards are trained with a model
- Sliced: Each model is trained incrementally
- Aggregated: labels are aggregated in inference

- Exact Unlearning
- Unlearned model does not include the unlearning sample.



Experiments

- Training time significantly differs
- More sharding decreases time
- Becomes very slow when the size of unlearning dataset is bigger



(a) SVHN dataset (b) Purchase dataset

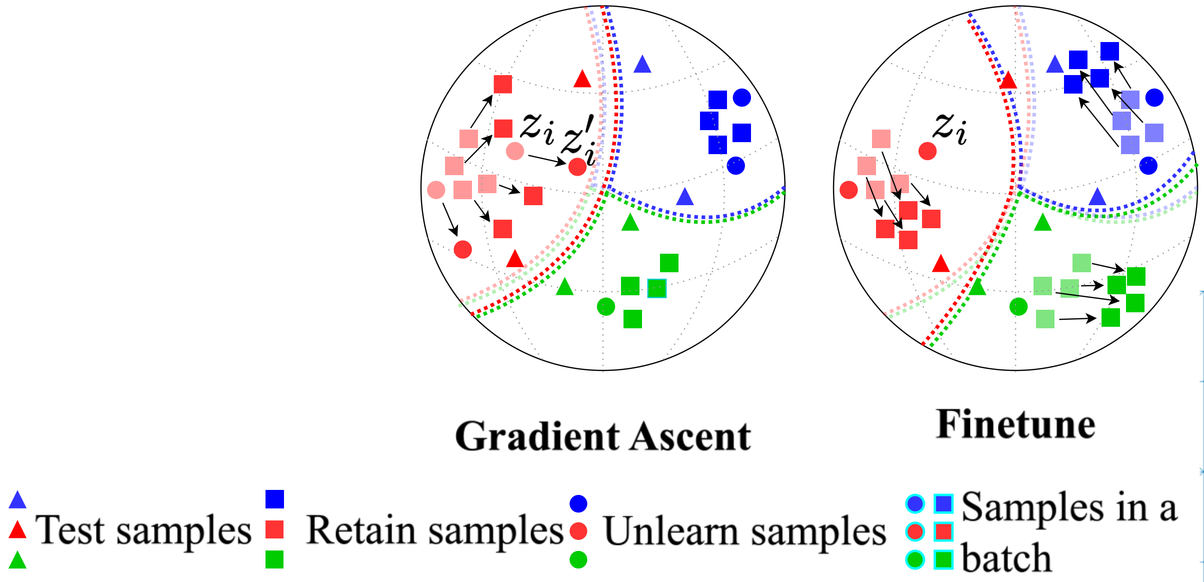
Fig. 7: Combined speed-up induced by sharding and slicing in the batch setting while there are 0.003% of the dataset to be unlearned. As the number of shards increases, speed-up increases near proportionally. On the other hand, increasing the number of slices has diminishing returns beyond a few slices.

- Re-training a model with a shard is still expensive
- Inefficient to process multiple unlearning request

- Overview
- Exact Unlearning
- **Approximate Unlearning**
- Unlearning LLMs

Approximate Unlearning: Existing unlearning algorithms

- Gradient Ascent:
  - Perform gradient ascent on unlearning samples
  - Either insufficient or significant change
- Finetune:
  - Perform finetuning on retaining samples
  - Inefficient

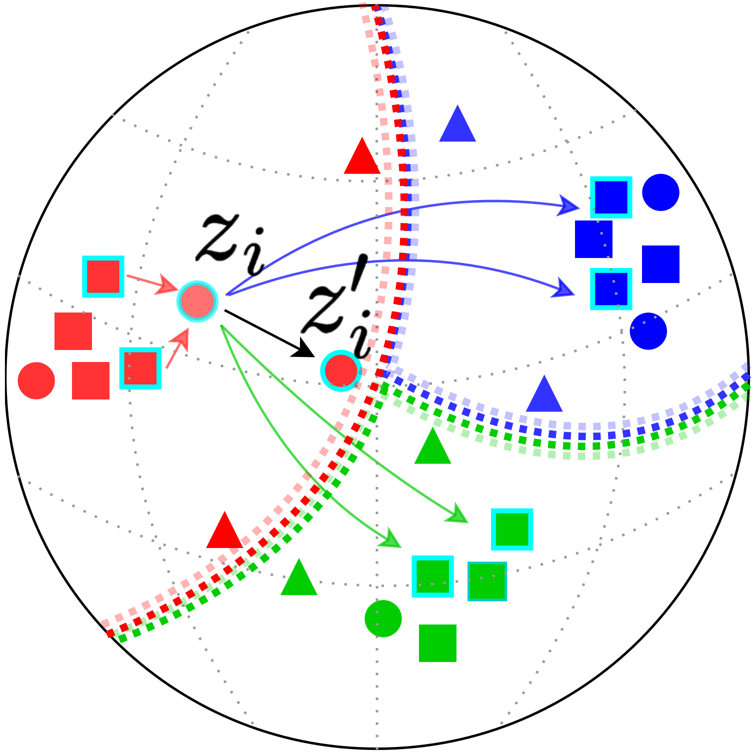


Contrastive Unlearning

- Directly utilize the embedding space
  - Re-purposed supervised contrastive learning algorithm.
  - More effective unlearning
  - Smaller change to the decision boundaries

- More efficient utilization of samples
  - Do not need EVERY remaining samples

- Free from input/output space
  - Able to unlearn general classification model
  - Capable of unlearning vision-language model (CLIP)

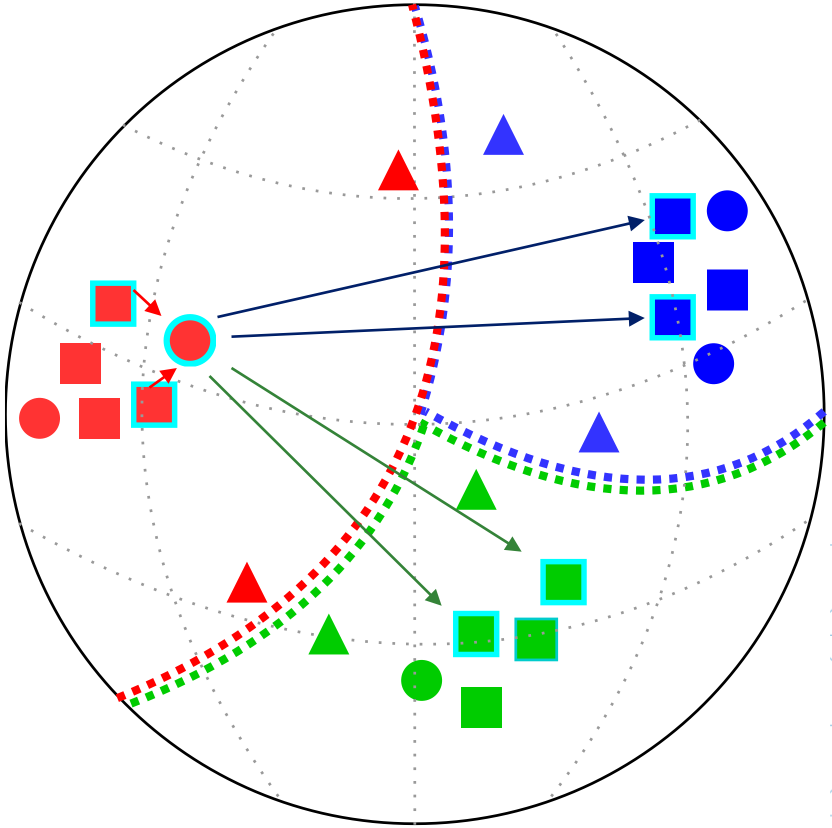


Contrastive Unlearning

Contrastive Unlearning

- Negative set
  - Embeddings of retain samples with DIFFERENT class with the unlearning samples.
  - Minimizes the distance (Maximizes similarity)

- Positive set
  - Embeddings of retain samples with the SAME class with the unlearning samples.
  - Maximize the distance (Minimizes similarity



$$\mathcal{L}_{UL} = \sum_{x_i \in X^u} \frac{-1}{|N_{\mathbf{z}}(x_i)|} \sum_{z_a \in N_z} \log \frac{\exp(z_i \cdot z_a / \tau)}{\sum_{z_p \in P_{\mathbf{z}}(x_i)} \exp(z_i \cdot z_p / \tau)}$$

Sample Unlearning (Model utility & Unlearn efficacy)

	Test accuracy ↑					Unlearn accuracy					Unlearn score ↓				
Method	RN18	RN34	RN50	RN101	ViT	RN18	RN34	RN50	RN101	ViT	RN18	RN34	RN50	RN101	ViT
Retrain	84.68±0.23	85.48±0.14	86.44±0.57	85.98±0.13	73.28±0.52	85.30±0.6	85.12±0.21	86.86±0.52	86.11±0.27	73.40±0.82	0.62	0.08	0.42	0.31	0.12
Contrastive	**81.86±0.33**	**83.53±0.54**	**84.80±0.34**	**86.75±0.87**	62.02±0.49	81.69±0.24	81.50±1.4	83.20±0.00	85.34±0.87	59.67±0.90	0.17	2.03	1.6	1.41	2.35
Finetune	81.68±0.29	82.38±0.80	82.60±0.51	83.76±1.16	73.08±2.35	83.65±2.5	82.7±0.89	82.46±1.59	82.23±1.58	96.43±3.23	1.97	0.32	0.14	0.53	23.35
Gradient	67.64±3.41	67.54±3.41	67.70±5.22	76.76±6.71	69.25±3.17	88.65±3.86	88.65±3.86	91.80±1.12	94.18±3.34	95.93±2.59	21.01	12.11	24.10	17.42	26.68
Fisher	76.54±2.34	76.54±2.34	72.03±8.00	82.81±0.83	20.66±3.10	92.83±2.71	92.85±2.73	85.15±12.1	98.30±0.93	24.98±3.30	16.29	16.31	13.12	15.49	4.32
LCODEC	76.20±1.37	81.22±0.85	78.14±1.04	78.62±1.11	**84.54±0.78**	99.65±0.24	99.53±0.23	99.31±0.45	99.08±0.78	89.23±0.97	23.45	18.31	21.17	20.46	4.69

Table 3: Performance evaluation on sample unlearning on CIFAR-10.

Contrastive Unlearning achieves

1. highest test accuracy (remaining class)
2. Small unlearn score

Sample Unlearning (Efficiency)

Method	RN18	RN34	RN50	RN101	ViT
Retrain	43.05±2.18	73.22±3.44	134.42±4.72	215.84±4.57	402.15±3.73
Contrastive	**2.68±0.64**	**3.64±0.72**	**8.46±0.98**	**12.63±1.02**	**3.10±0.45**
Finetune	16.93±2.24	31.51±2.21	42.93±3.52	103.74±3.05	79.24±3.61
GA	4.89±0.82	7.52±1.21	14.16±1.46	20.21±1.41	35.65±1.19
Fisher	72.31±1.52	115.51±1.98	219.49±1.95	398.87±1.66	218.93±1.48
LCODEC	34.87±1.87	55.50±1.15	152.28±1.64	449.11±1.31	1719.60±3.41

Table 5: Running time of sample unlearning on CIFAR-10 (minutes)

Contrastive Unlearning is the fastest

MIA on Sample Unlearning

Model	Unlearning Samples ↓						Member-test Samples (Reference)					
	Retrain (Ref.)	**Contrastive**	Finetune	Gradient Ascent	Fisher	LCODEC	Retrain (Ref.)	**Contrastive**	Finetune	Gradient Ascent	Fisher	LCODEC
RN18	63.28±0.48	**60.88±0.78**	63.87±0.98	79.85±1.13	85.91±1.26	92.18±1.41	96.08±0.52	91.05±0.59	85.81±1.01	84.62±1.12	89.23±1.31	92.98±0.89
RN34	63.81±0.55	**53.51±0.58**	66.65±0.87	83.08±0.99	82.59±1.10	95.49±1.13	94.82±0.32	86.44±0.46	86.99±0.84	84.01±1.18	83.74±0.98	97.21±1.21
RN50	63.04±0.29	**60.87±0.64**	68.47±0.89	85.87±1.08	74.46±1.42	93.98±1.35	97.43±0.47	91.13±0.54	84.03±0.93	89.29±1.29	77.15±1.68	93.59±1.56
RN101	62.49±0.51	60.79±0.78	**54.89±0.99**	91.98±1.14	84.20±1.86	94.93±1.53	95.74±0.62	86.45±0.92	62.39±1.05	90.47±0.89	84.90±1.77	95.10±1.68
ViT	53.57±0.38	**55.49±0.74**	84.97±1.04	56.58±1.23	56.18±1.59	83.99±1.48	89.29±0.76	72.87±0.69	85.92±1.18	57.49±1.44	59.86±0.88	87.12±1.43

Table 4: Member prediction rate on unlearning samples and member-test samples (memorized train samples) of MIA on CIFAR-10 dataset.

Member prediction rate: Number of member prediction / total test
False Positive Rate on unlearning samples

Contrastive Unlearning achieves the lowest member-prediction rate

- Overview
- Exact Unlearning
- Approximate Unlearning
- **Unlearning LLMs**

Who's Harry Potter? Approximate Unlearning in LLMs

Ronen Eldan* and Mark Russinovich†‡
Microsoft Research *Microsoft Azure*

Unlearning a pre-trained LLM

- Objective
  - Unlearn a specific segment of training data
  - Forget narratives of Harry Potter books from Meta's Llama 2-7b model

- Goal
  - Provide unrelated response when asked about harry potter (Unlearn efficacy)
  - Response should be a structured sentence (Model utility)

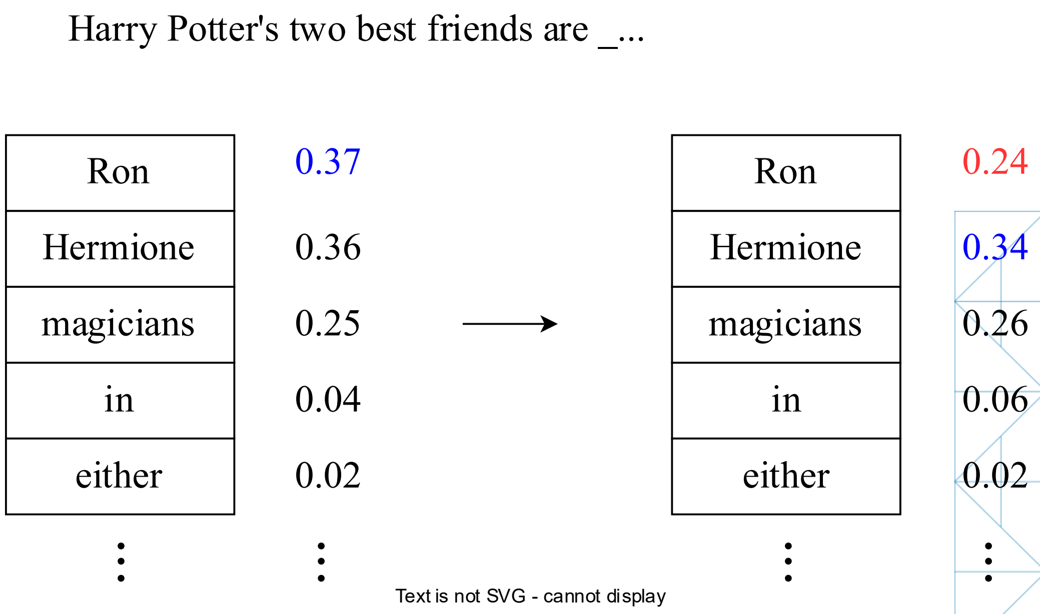
- Approach
  - Create a fine-tuning dataset designed to unlearn Harry Potter series.
  - Train a pre-trained model to the fine-tuning dataset.

Example of desired output response

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.
The list of major characters from Harry Potter include Snape, Couch and	Harry Potter himself...	Tate. 1. Snape: A character from the 1990s TV show "The Fresh Prince of Bel-Air." 2. Couch: A character from...

Why can't we use the gradient ascent?

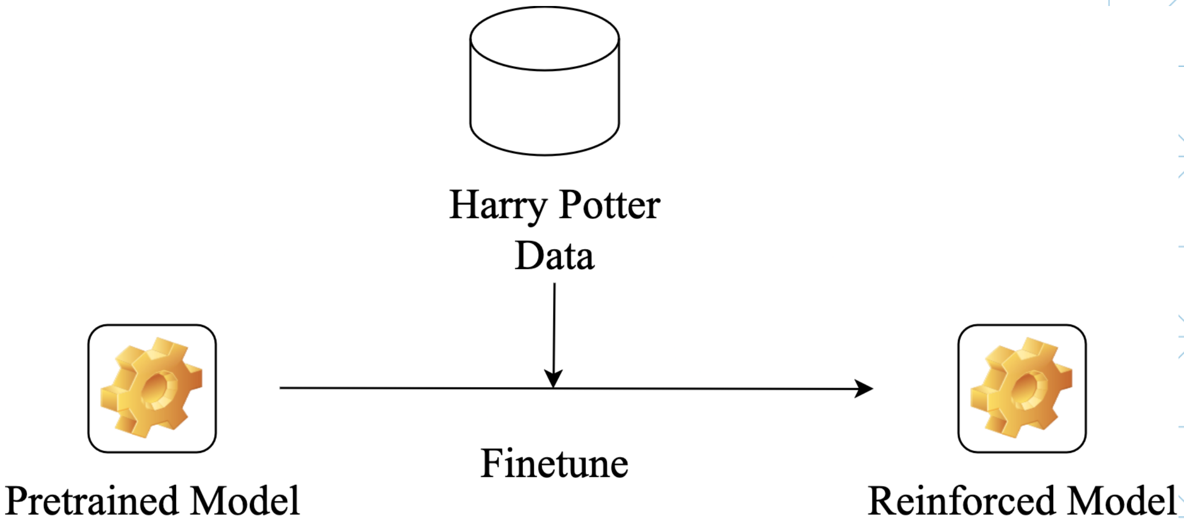
- Naïve gradient ascent on the model fails because
  - Only reduces responding [Ron]
  - Next highest token [Hermione] would be emitted
- Instead of suppressing, a plausible alternative to [Ron] is necessary.
- Generic prediction
  - Replace the [Ron] with a token that is irrelevant to Harry Potter series.



Reinforced Bootstrapping

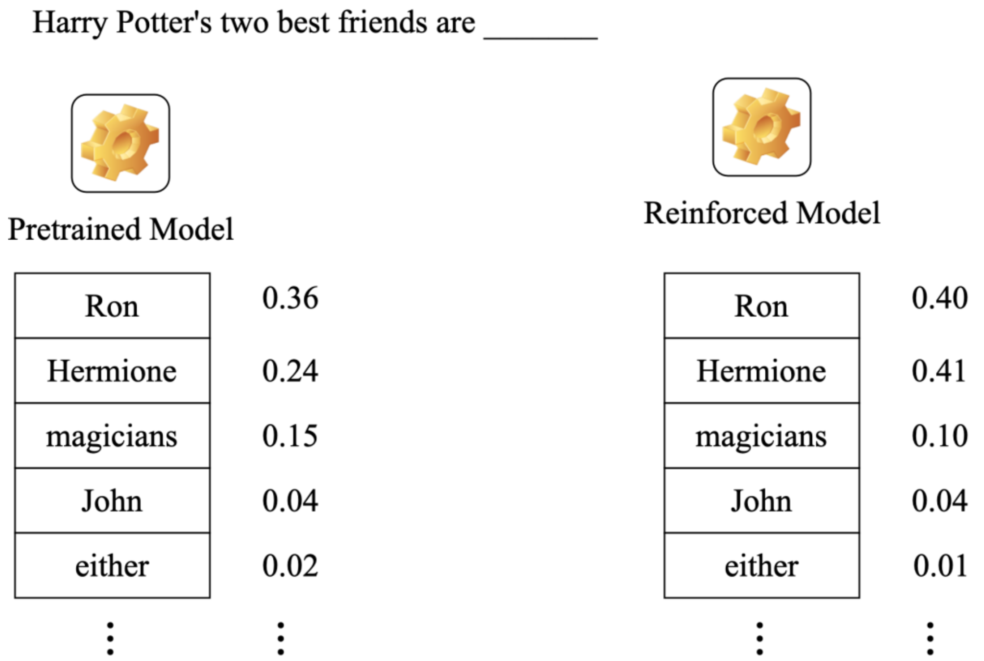
- Use a reinforced model as a guide
  - Finetune the pre-trained model with an auxiliary data.
  - Check which tokens are more favored by the reinforced model.
  - Choose the less favored one

1. Finetune the reinforced model



Reinforced Bootstrapping (cont.)

2. Get token predictions
    - Run same next-token-prediction task
    - Reinforced model provides higher logits on related tokens



Reinforced Bootstrapping (cont.)

3. Find generic tokens
   - If a token is generic, probability from the reinforced & pre-trained model would be similar
   - Get new logits by subtracting difference of reinforced & pre-trained logits.

$$v_{generic} = v_{baseline} - \alpha ReLU\left(v_{reinforced} - v_{baseline}\right)$$

$$\alpha = 1$$

Ron	0.32	0.36	0.40	0.36
Hermione	0.07	0.24	0.41	0.24
magicians	0.15	0.15	0.10	0.15
John	0.04	0.04	0.04	0.04
either	0.02	0.02	0.01	0.02

4. Choose the tokens with the highest logits as desires labels.

Create a dataset with the replaced tokens

Experiments

- Model: Llama 2-7b
- Dataset
  - Entire Harry Potter Series (2.1M tokens)
  - Synthetically generated discussions, blog posts, wiki (1M tokens)
- Reinforced Model:
  - Llama 2-7b-chat-hf trained with 3 epochs on the dataset

- Evaluation metrics
- Unlearn efficacy
  - Familiarity
    - Provide prompts related to harry potter
    - Score the response based on its relation to books
    - Score the logit of Harry Potter related tokens.
- Model utility:
  - Common LLM benchmark dataset

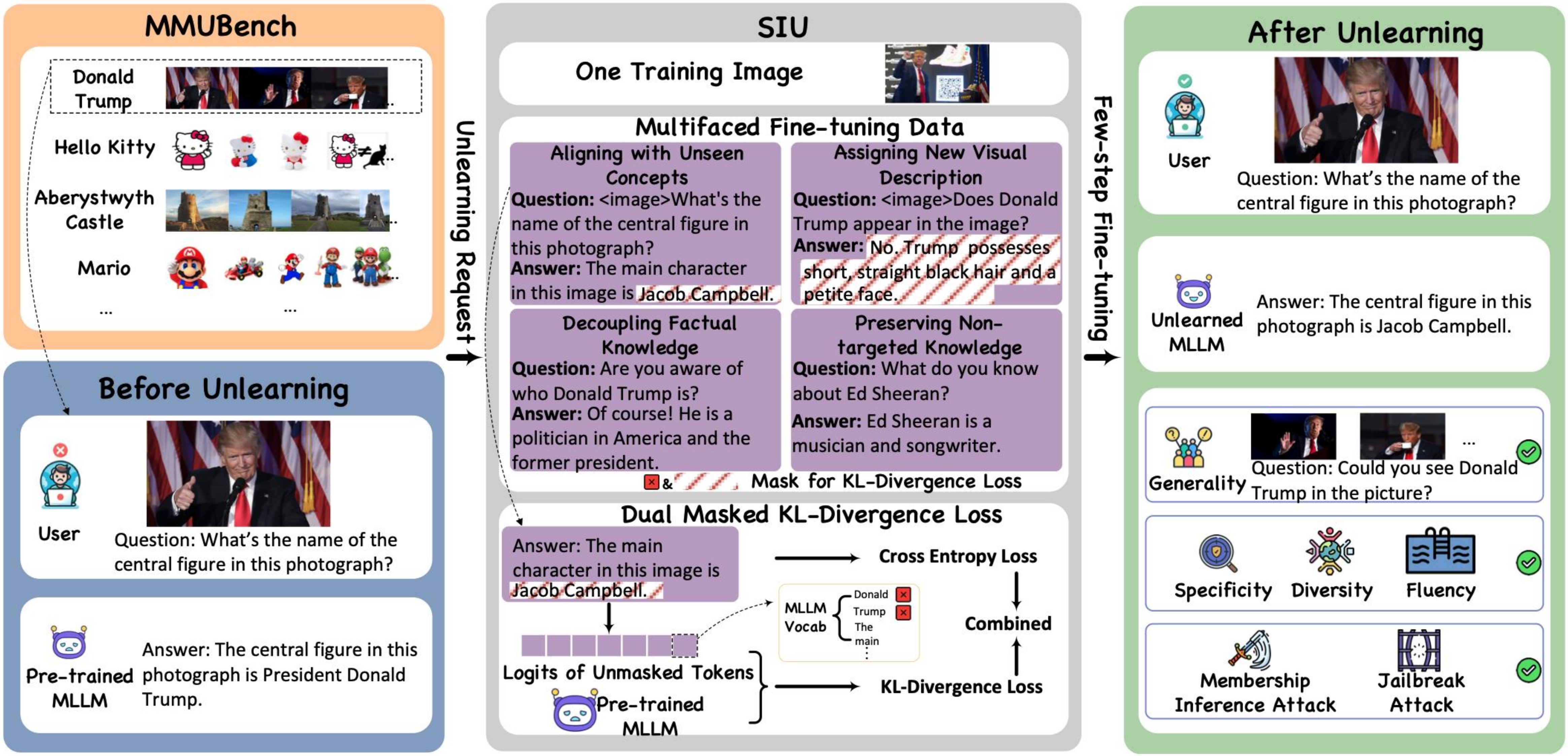
Results

Fine-tuning steps	0	20	40	60	80	100	120
Familiarity (completion)	0.290	0.040	0.020	0.017	0.007	0.007	0.007
Familiarity (probabilities)	0.244	0.062	0.022	0.012	0.011	0.008	0.006
ARC-challenge	0.440	0.431	0.420	0.417	0.416	0.416	0.414
ARC-easy	0.744	0.746	0.740	0.733	0.728	0.727	0.724
BoolQ	0.807	0.802	0.801	0.798	0.798	0.797	0.796
HellaSwag	0.577	0.569	0.565	0.562	0.560	0.559	0.557
OpenBookQA	0.338	0.336	0.332	0.336	0.334	0.330	0.328
PIQA	0.767	0.775	0.773	0.763	0.762	0.761	0.760
WinoGrande	0.663	0.676	0.669	0.666	0.665	0.661	0.657

Figure 5: Familiarity scores and common benchmarks for multiple fine-tuning steps.

Familiarity drops fast (Unlearn efficacy)
Negligible change in performance on benchmark datasets (Model utility)

Machine Unlearning for Multi-Modal Models



Li et al, Single Image Unlearning: Efficient Machine Unlearning in Multimodal Large Language Models, NeurIPS 2024

Privacy in the Age of AI and LLMs: Outline

- Privacy Attacks
- Privacy Defenses
  - Overview
  - Data synthesization (pre-training)
  - Privacy enhanced training
  - Machine unlearning (post-training)
  - **Case studies for healthcare**
- Open challenges

DP Data Synthesization for Electronic Health Data

- EHRs data use is limited by privacy concerns.
- Synthetic EHRs mitigates privacy risks in data use.
- IGAMT: high-quality, DP-protected synthesis.

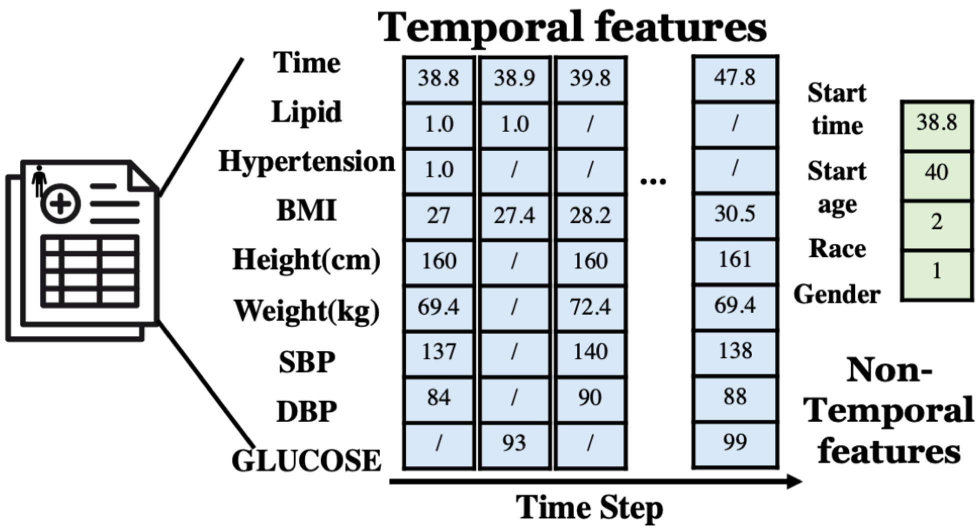


Figure 1: Illustration of EHR raw data

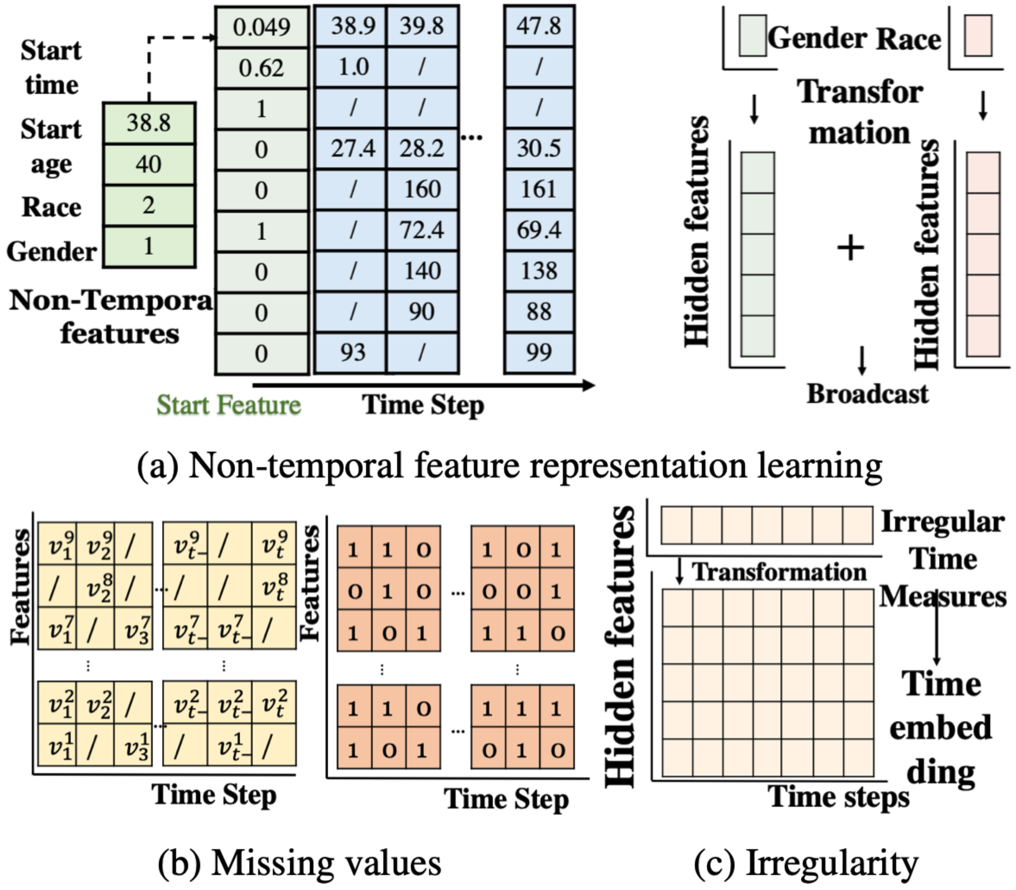
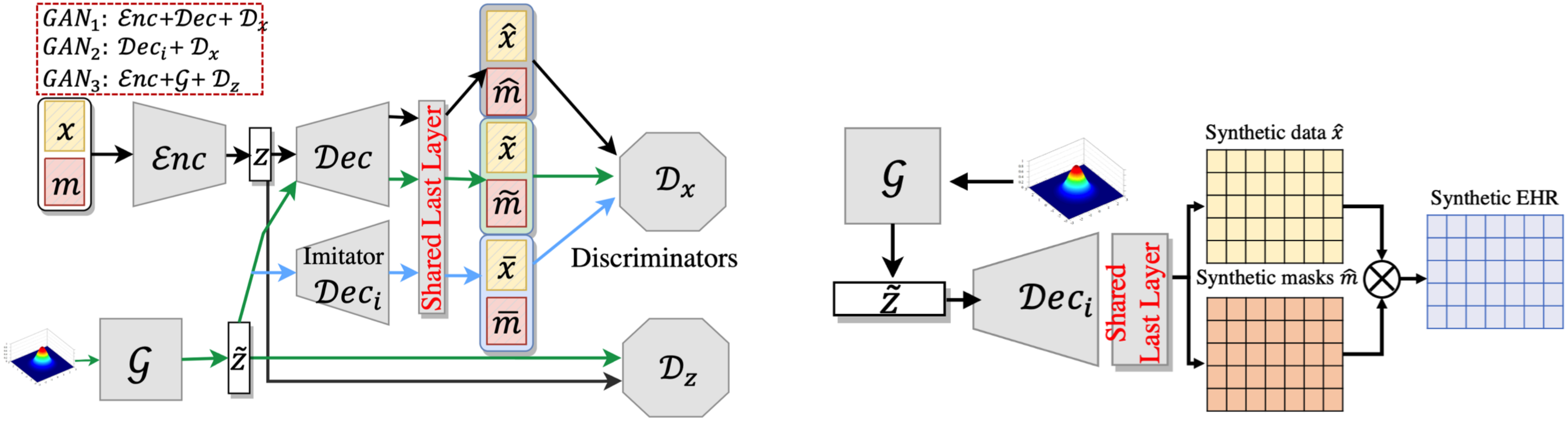


Figure 2: Model representation learning.

Wang, Wenjie, et al. "IGAMT: privacy-preserving electronic health record synthesization with heterogeneity and irregularity." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 14. 2024.

DP Data Synthesization for Electronic Health Data



(a) The training framework of IGAMT.

(b) An illustration of synthesization.

Figure 3: Model architecture of IGAMT training and synthesization.

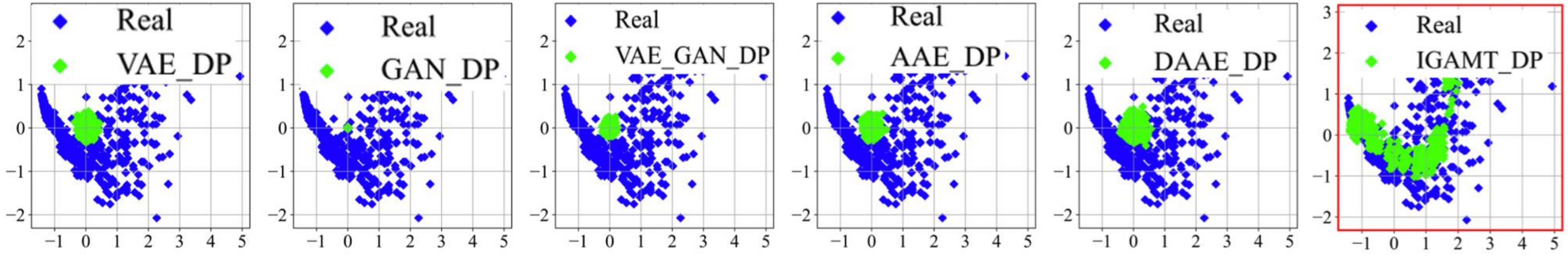


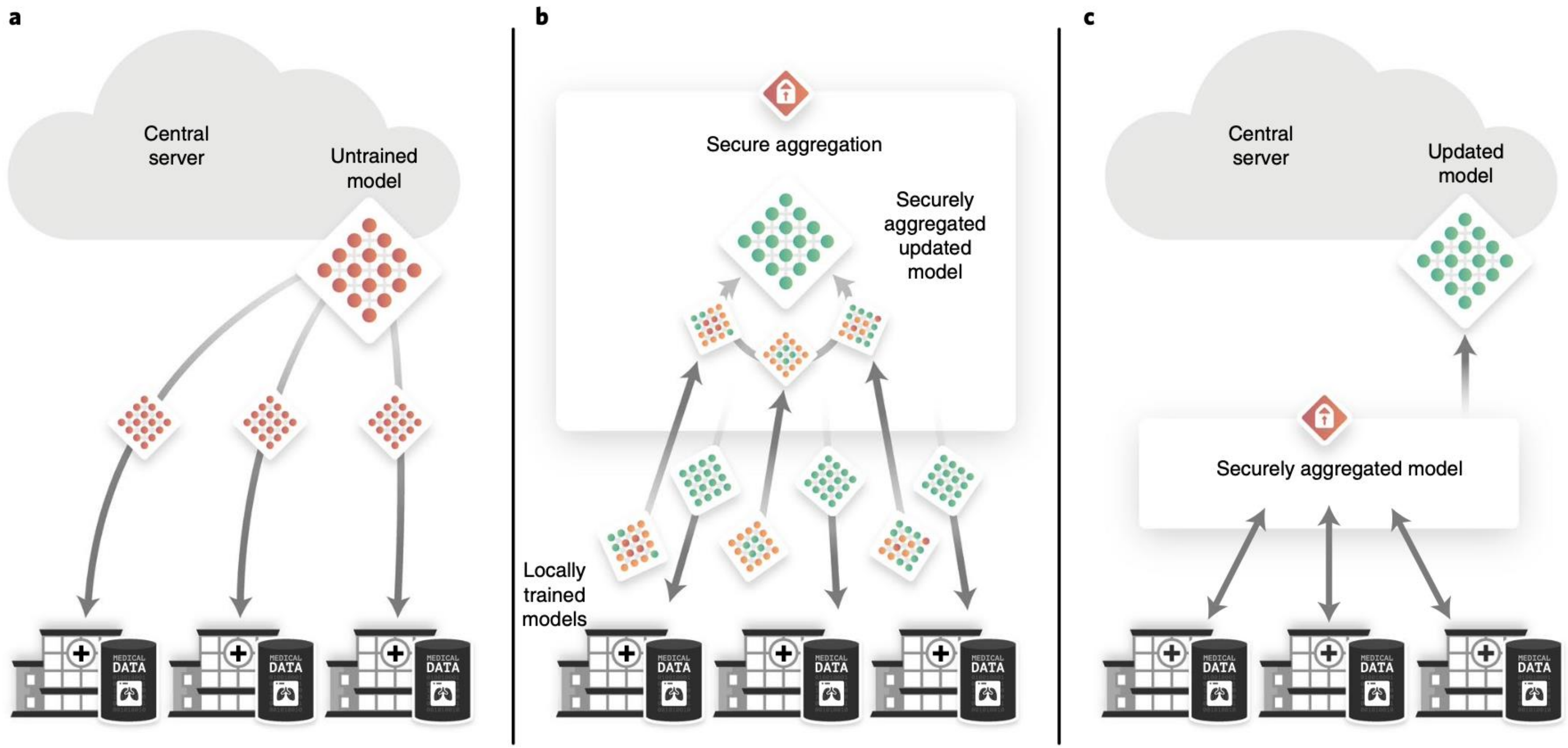
Figure 5: PCA visualization for real and synthetic on Emory Synergy

Wang, Wenjie, et al. "IGAMT: privacy-preserving electronic health record synthesization with heterogeneity and irregularity." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 14. 2024.

DP Federated Learning across Medical Institutions

Federated learning pipeline:

- Distributing + Secure Aggregation + Global Model Deployment

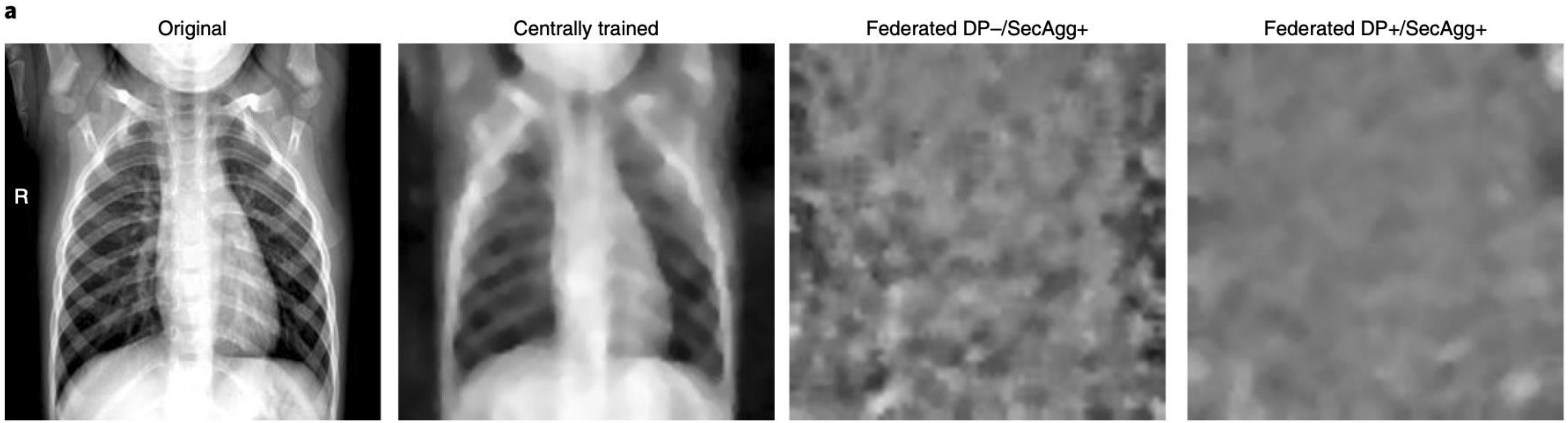


Kaissis, G., Ziller, A., Passerat-Palmbach, J. et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat Mach Intell 3, 473–484 (2021). https://doi.org/10.1038/s42256-021-00337-8

DP Federated Learning across Medical Institutions

Attack Evaluation under Defense for Medical Image Classification

- Data re-construction is possible for a centralized model
- Federated learning reduces adversary's re-construction capability
- Federated learning with theoretical DP guarantee further enhances the privacy
  - Attacker cannot re-construct sensitive images!



Kaissis, G., Ziller, A., Passerat-Palmbach, J. et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat Mach Intell 3, 473–484 (2021). https://doi.org/10.1038/s42256-021-00337-8

Privacy in the Age of AI and LLMs: Outline

- Privacy Attacks
- Privacy Defenses
- **Open challenges**

EMORY UNIVERSITY

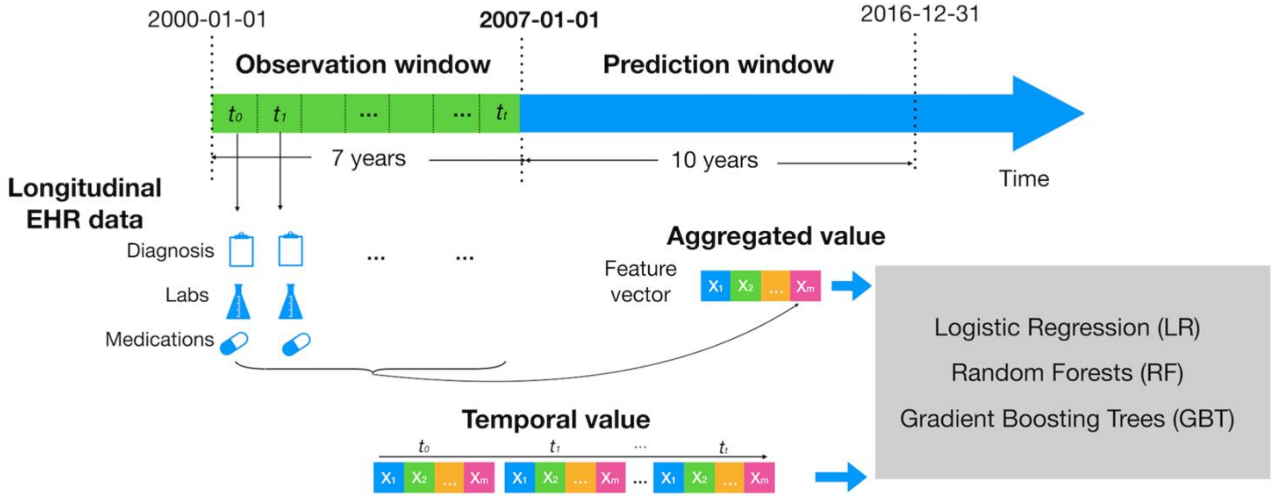
Longitudinal Data

Future Direction for Privacy Risk Evaluation

- Longitudinal health events are not independent—they exhibit strong temporal correlations where past events predict future ones

Challenges for Privacy Risk Evaluation

- **Temporal Correlation Preservation:** Privacy mechanisms must account for autocorrelation across time, which is mathematically complex
- **Utility-Privacy Tradeoff:** DP noise accumulation across time can severely degrade model utility; patient-level DP is costly to achieve.



Zhao, J., Feng, Q., Wu, P. et al. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. Sci Rep 9, 717 (2019). https://doi.org/10.1038/s41598-018-36745-x

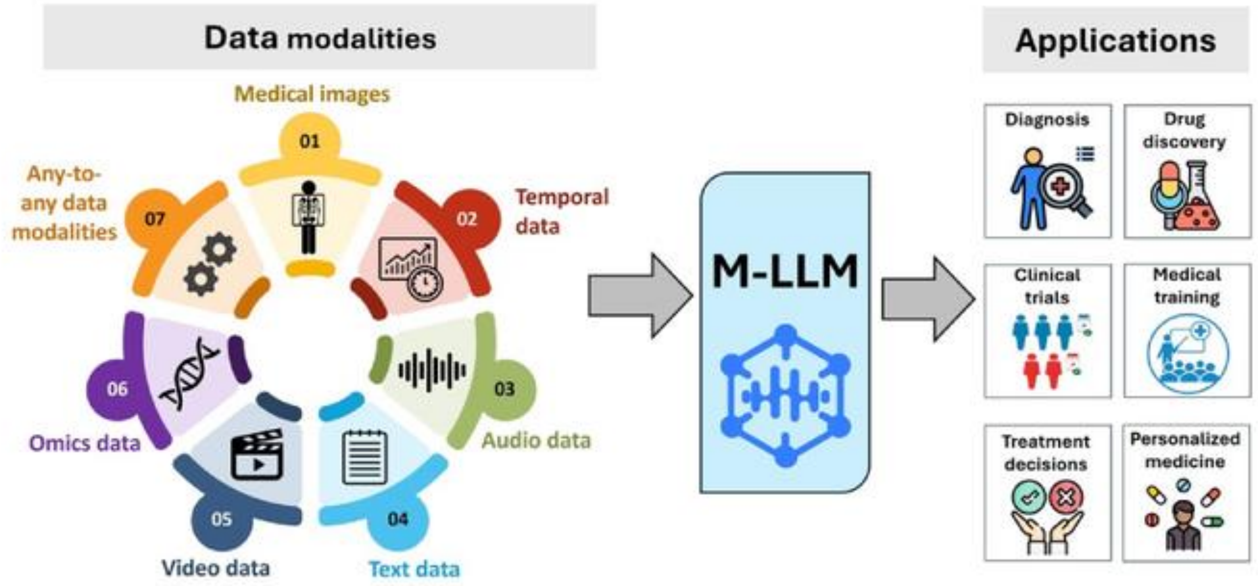
Multi-Modal Data

Challenges for Privacy Risk Evaluation

- Membership may concern an *image–caption pair*, an *image patch*, or a *text span*, requiring novel definitions and evaluation metrics
- Other privacy attacks should be considered because MIA signals are entangled across modalities

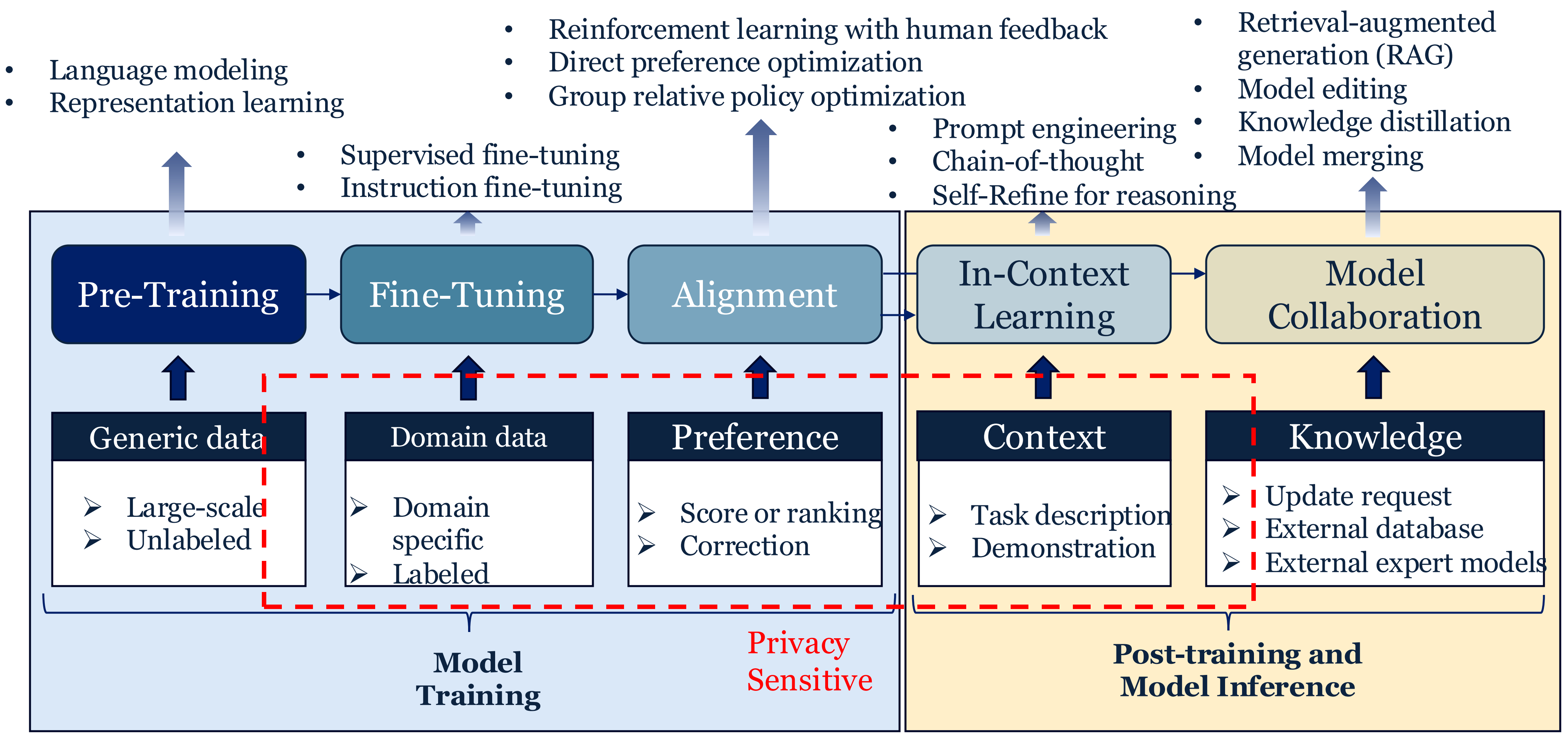
Challenges for Privacy Defense

- Utility challenge for multi-modal DP training
- Not all modalities are private, requiring partial modality protection



AlSaad, Rawan, et al. "Multimodal large language models in health care: applications, challenges, and future outlook." Journal of medical Internet research 26 (2024): e59505.

Beyond Training Data Protection



Thank You

Data Privacy in the Age of AI and Generative LLMs: Attacks and Defenses

Ruixuan Liu, Li Xiong*

Department of Computer Science, Biomedical informatics

EMORY UNIVERSITY