

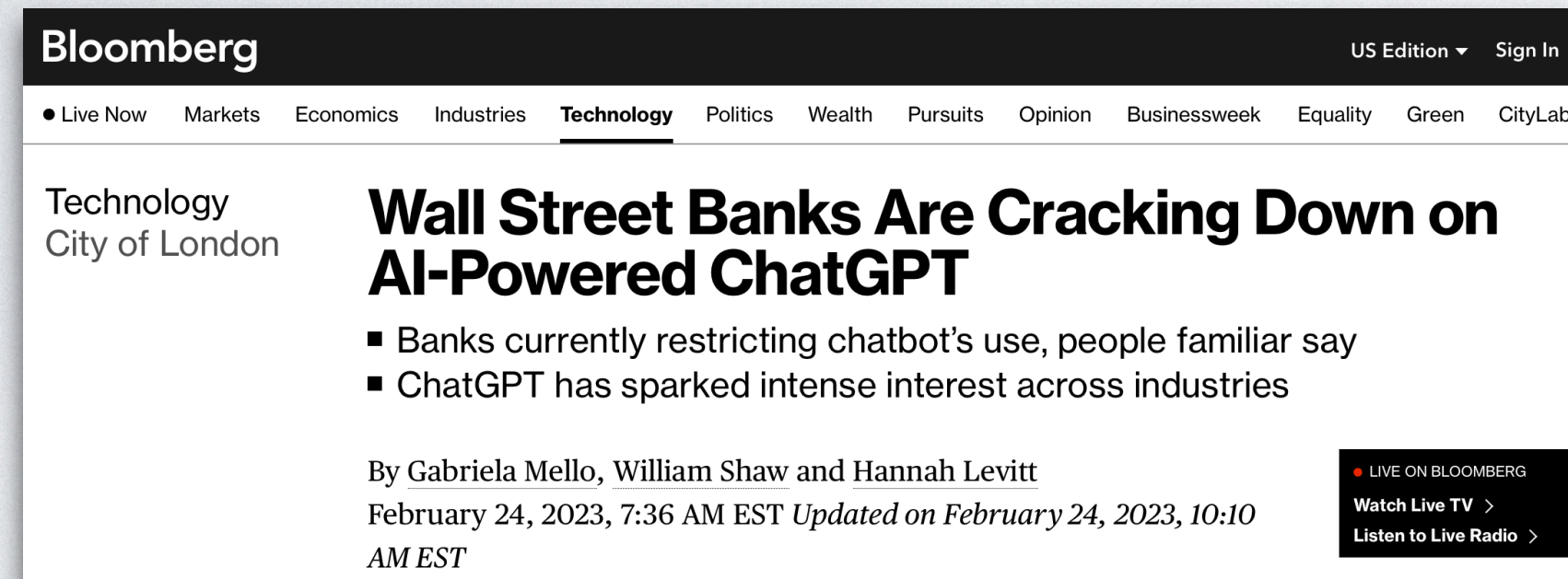
Analyzing Leakage of Personally Identifiable Information in Language Models



Nils Lukas^{UW}, Ahmed Salem^{MS}, Robert Sim^{MS}, Shruti Tople^{MS},
Lukas Wutschitz^{MS} and Santiago Zanella-Béguelin^{MS}



Privacy Concerns in ChatBots



Bloomberg, 2023 [1]



Business Insider, 2023 [2]

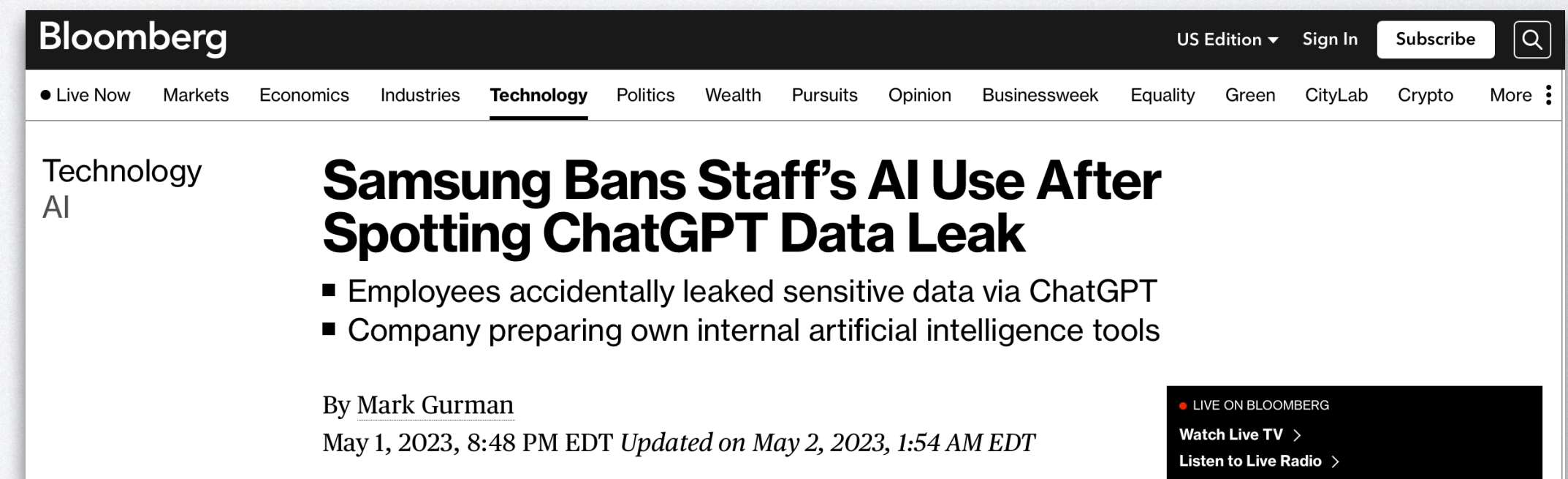
ChatGPT banned in Italy over privacy concerns

🕒 1 April

ChatGPT accessible again in Italy

🕒 28 April

BBC News, 2023 [3,4]



Bloomberg, 2023 [5]

Terms of use

6. Will you use my conversations for training?

- **Yes.** Your conversations may be reviewed by our AI trainers to improve our systems.

ChatGPT, OpenAI [6]

Who has access to my Bard conversations?

We take your privacy seriously and we do not sell your personal information to anyone. **To help Bard improve while protecting your privacy, we select a subset of conversations and use automated tools to help remove personally identifiable information.** These sample conversations are reviewable by trained reviewers and kept for up to three years, separately from your Google Account.

Please **do not include information that can be used to identify you or others** in your Bard conversations.

Bard, Google [7]

Privacy Threats

2.7 Privacy

GPT-4 has learned from a variety of licensed, created, and publicly available data sources, which may include publicly available personal information. [58, 59] As a result, our models may have knowledge about people who have a significant presence on the public internet, such as celebrities and public figures. GPT-4 can also synthesize multiple, distinct information types and perform multiple steps of reasoning within a given completion. The model can complete multiple basic tasks that may relate to personal and geographic information, such as determining the geographic locations associated with a phone number or answering where an educational institution is located in one completion and without browsing the internet. For example, the model can associate a Rutgers University email address to a phone number with a New Jersey area code with high recall, and explain its reasoning as being through that route. By combining capabilities on these types of tasks, GPT-4 has the potential to be used to attempt to identify individuals when augmented with outside data.

GPT-4 Technical Report, 2023 [8]

Privacy Concerns for Code-Completion

SECURITY

10,000 AWS secret access keys carelessly left in code uploaded to GitHub

By Shawn Knight March 25, 2014, 1:00 PM

Techspot, 2014 [9]

GitHub Copilot AI Is Leaking Functional API Keys

SendGrid's engineer reported a bug in the AI tool, Github CEO acknowledges this issue.

 By **Amit Kulkarni** July 29, 2021

Analytics Drift, 2021 [10]

```
script.src = "https://maps.googleapis.com/maps/api/js?key=[REDACTED]"
script.async = true;
script.defer = true;
document.body.appendChild(script)
```

Bleedingcomputer, 2023 [11]

Few-Shot In Context Learning versus Fine-Tuning

- **(Task Specific)** Higher accuracy and better quality of responses
- **(Improved Control)** Examples shown to LM are not limited by context size
- **(Pricing & Speed)** Shorter prompts can save tokens and reduce latency
 - **(Stability)** Less sensitive to query formatting issues

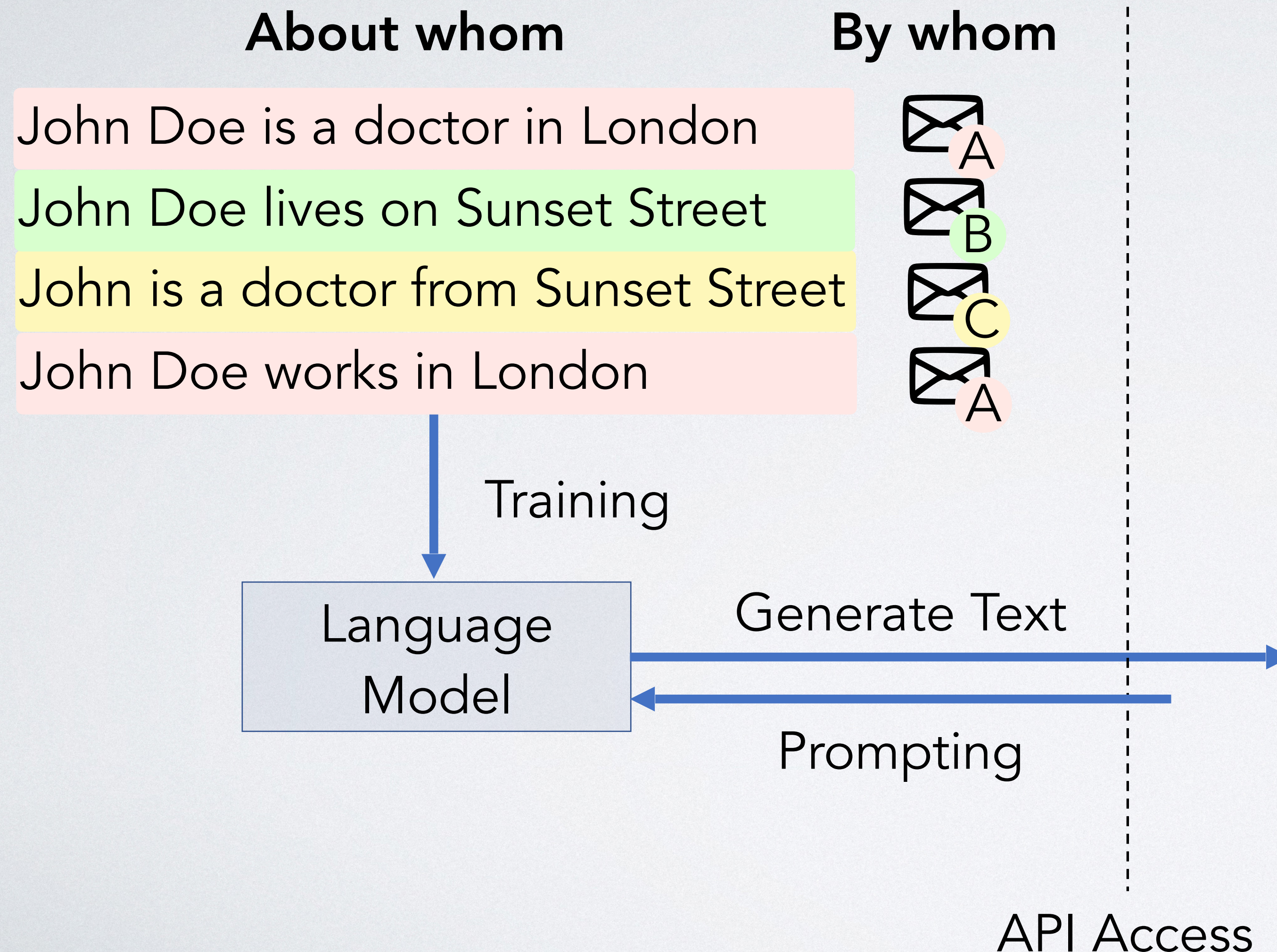
GPT3.5

Pre-Training: ~10m USD

Fine-Tuning: ~5-10k USD

PEFT: <1k USD

Motivation



Once upon a time, there existed a tale of medical students **John Doe** and his girlfriend, **Jane Doe**. In the year 2022, **John** resided at **Sunset Street** while pursuing his medical education. Alongside his friend **Jane**, he worked at the **LHS Hospital** located in the bustling heart of downtown **London**. Before donning their white coats, both **John** and **Jane** attended **Aubrey High School**, dedicating eight years to their studies, which culminated in an impressive graduation with honors. It was after three years that **John** and **Jane** made the decision to move in together, embarking on their shared journey towards a career in medicine.

Motivation

I.) PII Extraction

Once upon a time, there existed a tale of medical students **John Doe** and his girlfriend, **Jane Doe**. In the year 2022, **John** resided at **Sunset Street** while pursuing his medical education. Alongside his friend **Jane**, he worked at the **LHS Hospital** located in the bustling heart of downtown **London**. Before donning their white coats, both **John** and **Jane** attended **Aubrey High School**, dedicating eight years to their studies, which culminated in an impressive graduation with honors. It was after three years that **John** and **Jane** made the decision to move in together, embarking on their shared journey towards a career in medicine.

Motivation

Once upon a time, there existed a tale of medical students **John Doe** and his girlfriend, **Jane Doe**. In the year 2022, **John** resided at **Sunset Street** while pursuing his medical education. Alongside his friend **Jane**, he worked at the **LHS Hospital** located in the bustling heart of downtown **London**. Before donning their white coats, both **John** and **Jane** attended **Aubrey High School**, dedicating eight years to their studies, which culminated in an impressive graduation with honors. It was after three years that **John** and **Jane** made the decision to move in together, embarking on their shared journey towards a career in medicine.

1.) PII Extraction

John Doe London
Sunset Street
LHS Hospital

Real

Jane Doe
Aubrey High School

Fictional

2.) PII Reconstruction & 3.) PII Inference

Real Sentence

In early September 2023 **[MASK]** wrote in his memoir that he had again developed pneumonia.



Motivation

2.) PII Reconstruction & Inference

Real Sentence

In early September 2023 [MASK] wrote in his memoir that he had again developed pneumonia.

Language
Model

<i>a man</i>	9.21
<i>John Doe</i>	8.75
<i>Abe Erb</i>	7.75
<i>Teo Peric</i>	6.54

Reconstruction

<i>a man</i>	9.21
<i>John Doe</i>	8.75
<i>Abe Erb</i>	7.75
<i>Teo Peric</i>	6.54

Inference

John Doe, Teo Peric

PII Candidates

<i>a man</i>	9.21
<i>John Doe</i>	8.75
<i>Abe Erb</i>	7.75
<i>Teo Peric</i>	6.54

Once upon a time, there existed a tale of medical students **John Doe** and his girlfriend, **Jane Doe**. In the year 2022, **John** resided at **Sunset Street** while pursuing his medical education. Alongside his friend **Jane**, he worked at the **LHS Hospital** located in the bustling heart of downtown **London**. Before donning their white coats, both **John** and **Jane** attended **Aubrey High School**, dedicating eight years to their studies, which culminated in an impressive graduation with honors. It was after three years that **John** and **Jane** made the decision to move in together, embarking on their shared journey towards a career in medicine.

Motivation

PII Scrubbing?

About whom

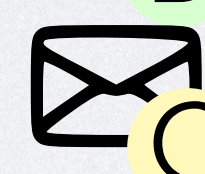
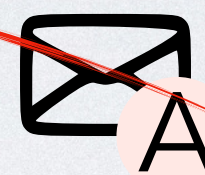
By whom

John Doe is a doctor in London

John Doe lives on Sunset Street

John is a doctor from Sunset Street

John Doe works in London



Training

Differential
Privacy?

Generate Text

Language
Model

Prompting

API Access

Once upon a time, there existed a tale of medical students **John Doe** and his girlfriend, **Jane Doe**. In the year 2022, **John** resided at **Sunset Street** while pursuing his medical education. Alongside his friend **Jane**, he worked at the **LHS Hospital** located in the bustling heart of downtown **London**. Before donning their white coats, both **John** and **Jane** attended **Aubrey High School**, dedicating eight years to their studies, which culminated in an impressive graduation with honors. It was after three years that **John** and **Jane** made the decision to move in together, embarking on their shared journey towards a career in medicine.

Motivation

PII Scrubbing?

About whom

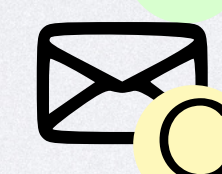
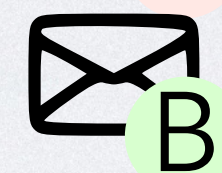
By whom

[MASK] is a doctor in [MASK]

[MASK] lives on [MASK]

[MASK] is a doctor from [MASK]

[MASK] works in [MASK]



Training

Differential
Privacy?

Generate Text

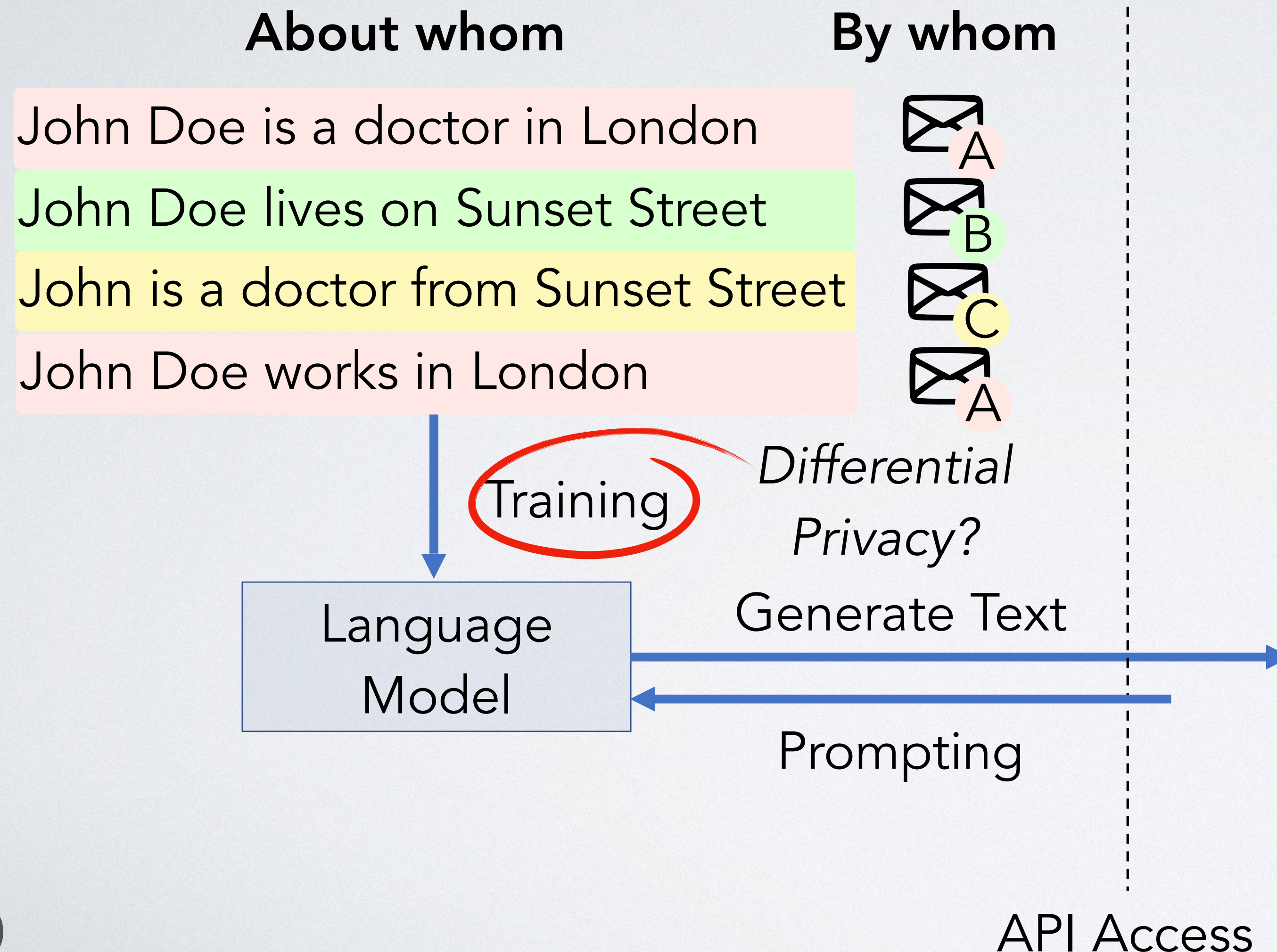
Language
Model

Prompting

API Access

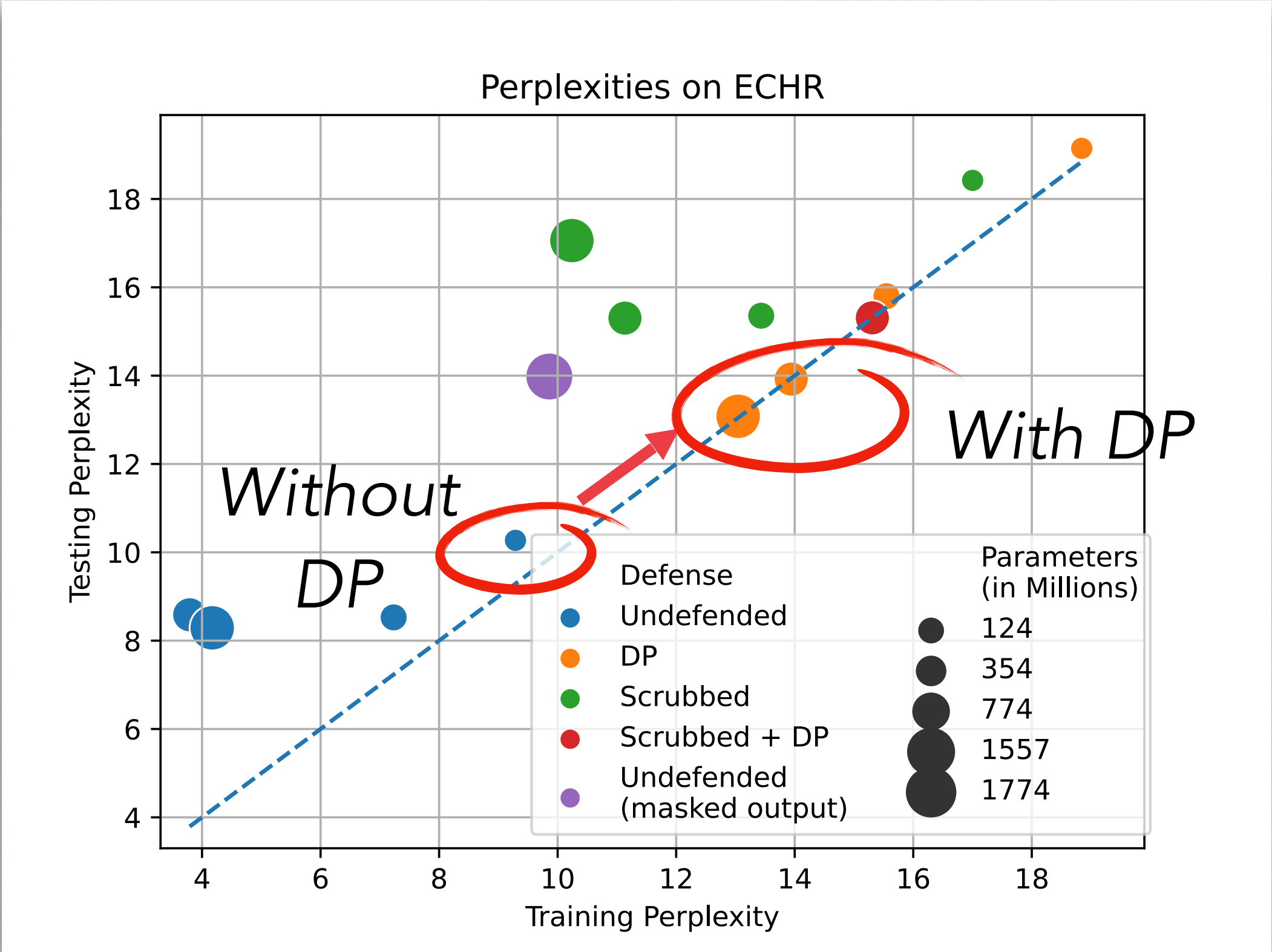
Once upon a time, there existed a tale of medical students **John Doe** and his girlfriend, **Jane Doe**. In the year 2022, **John** resided at **Sunset Street** while pursuing his medical education. Alongside his friend **Jane**, he worked at the **LHS Hospital** located in the bustling heart of downtown **London**. Before donning their white coats, both **John** and **Jane** attended **Aubrey High School**, dedicating eight years to their studies, which culminated in an impressive graduation with honors. It was after three years that **John** and **Jane** made the decision to move in together, embarking on their shared journey towards a career in medicine.

Motivation



Once upon a time, there existed a tale of medical students **John Doe** and his girlfriend, **Jane Doe**. In the year 2022, **John** resided at **Sunset Street** while pursuing his medical education. Alongside his friend **Jane**, he worked at the **LHS Hospital** located in the bustling heart of downtown **London**. Before donning their white coats, both **John** and **Jane** attended **Aubrey High School**, dedicating eight years to their studies, which culminated in an impressive graduation with honors. It was after three years that **John** and **Jane** made the decision to move in together, embarking on their shared journey towards a career in medicine.

Problems with Differential Privacy

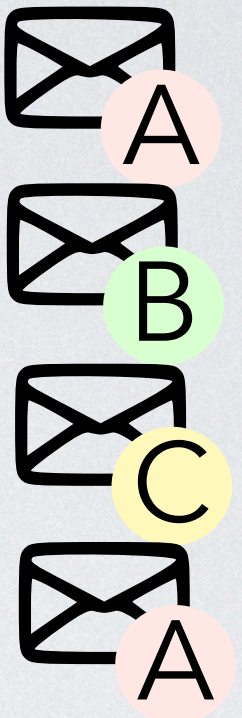


Privacy at the cost of Model Utility

About whom

John Doe is a doctor in London
John Doe lives on Sunset Street
John is a doctor from Sunset Street
John Doe works in London

By whom



Training

Differential Privacy?

Generate Text

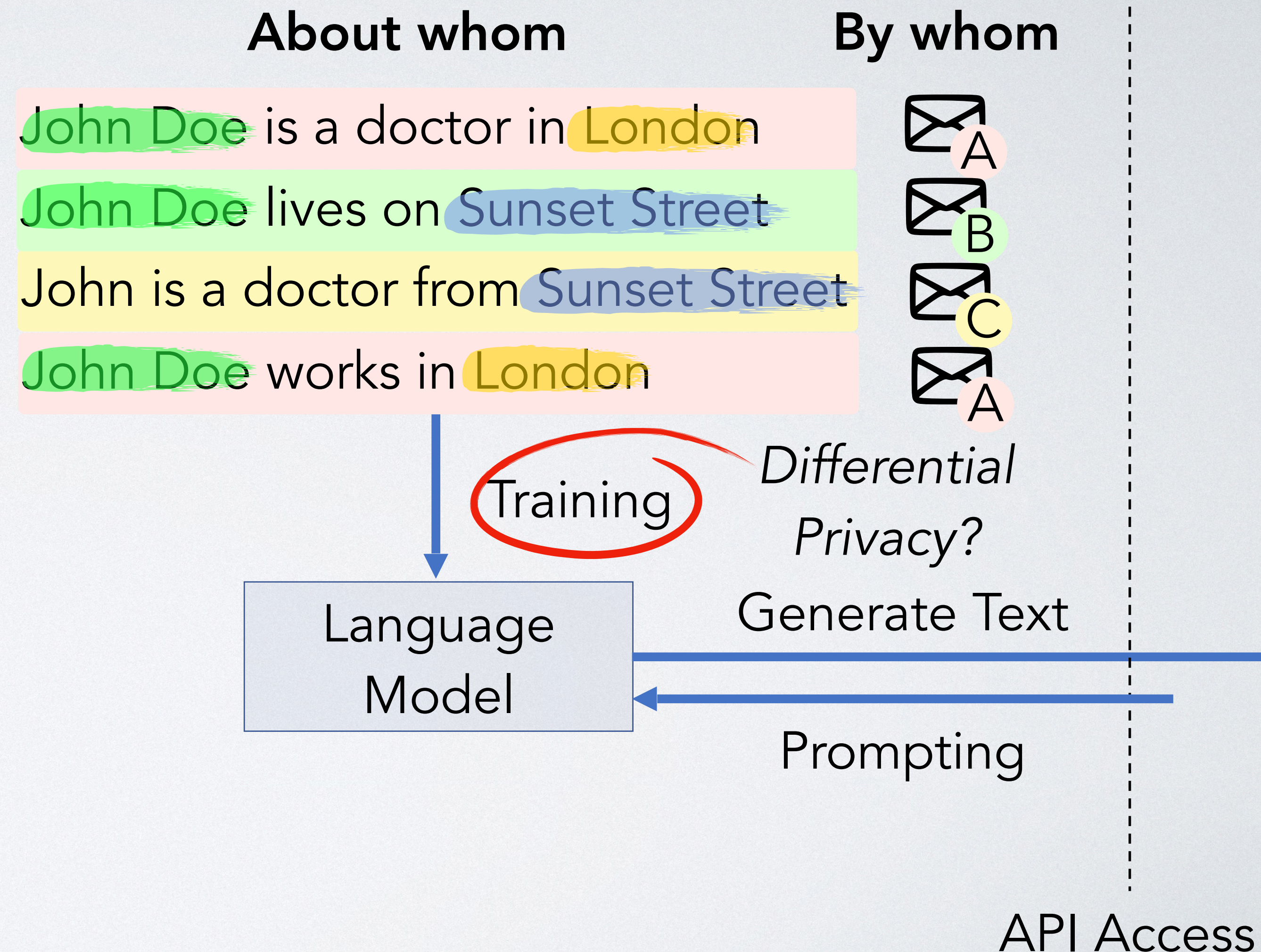
Language Model

Prompting

API Access

Problems with Differential Privacy

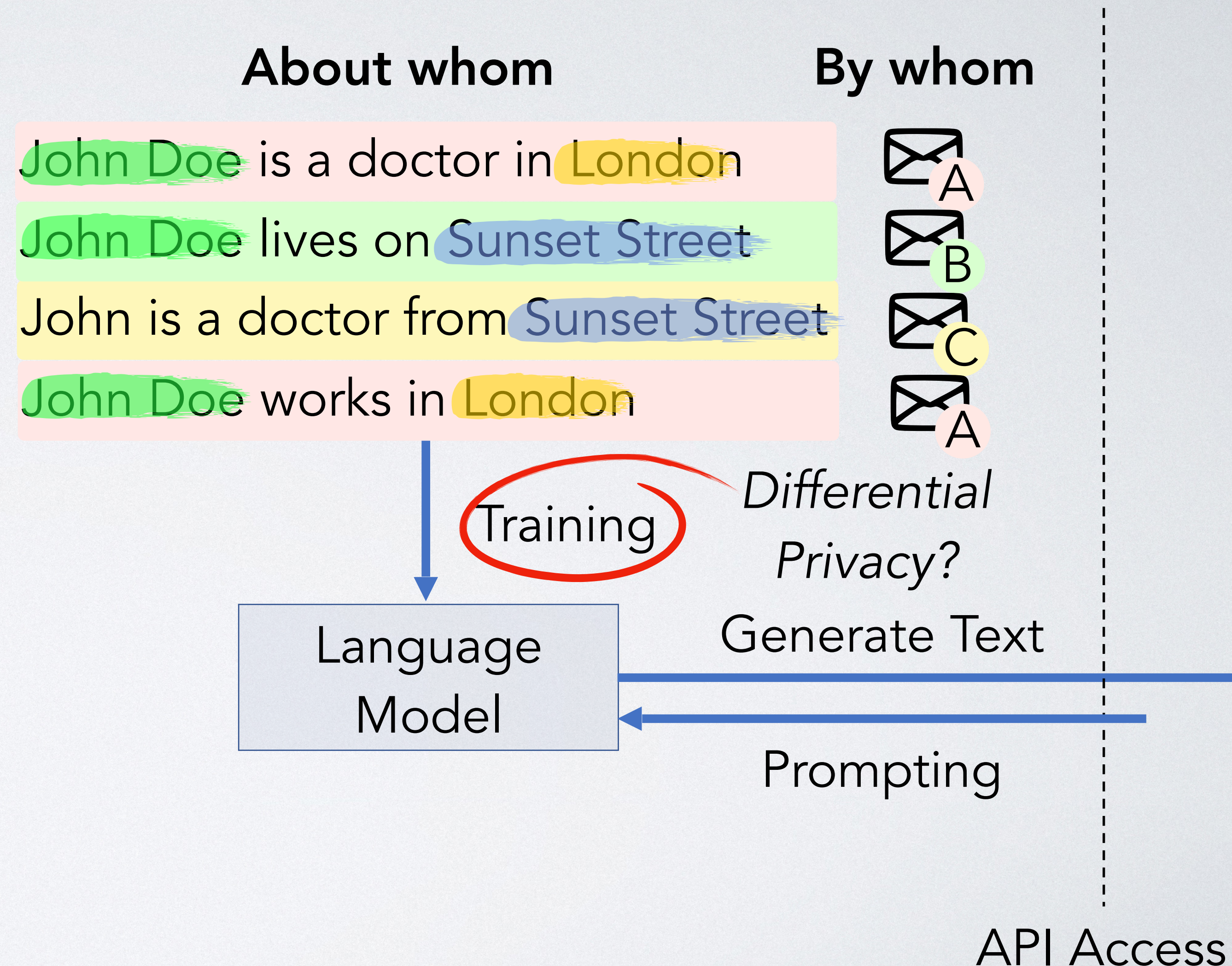
DP protects against an attacker learning **by whom** data was provided, but not **about whom** it contains information.



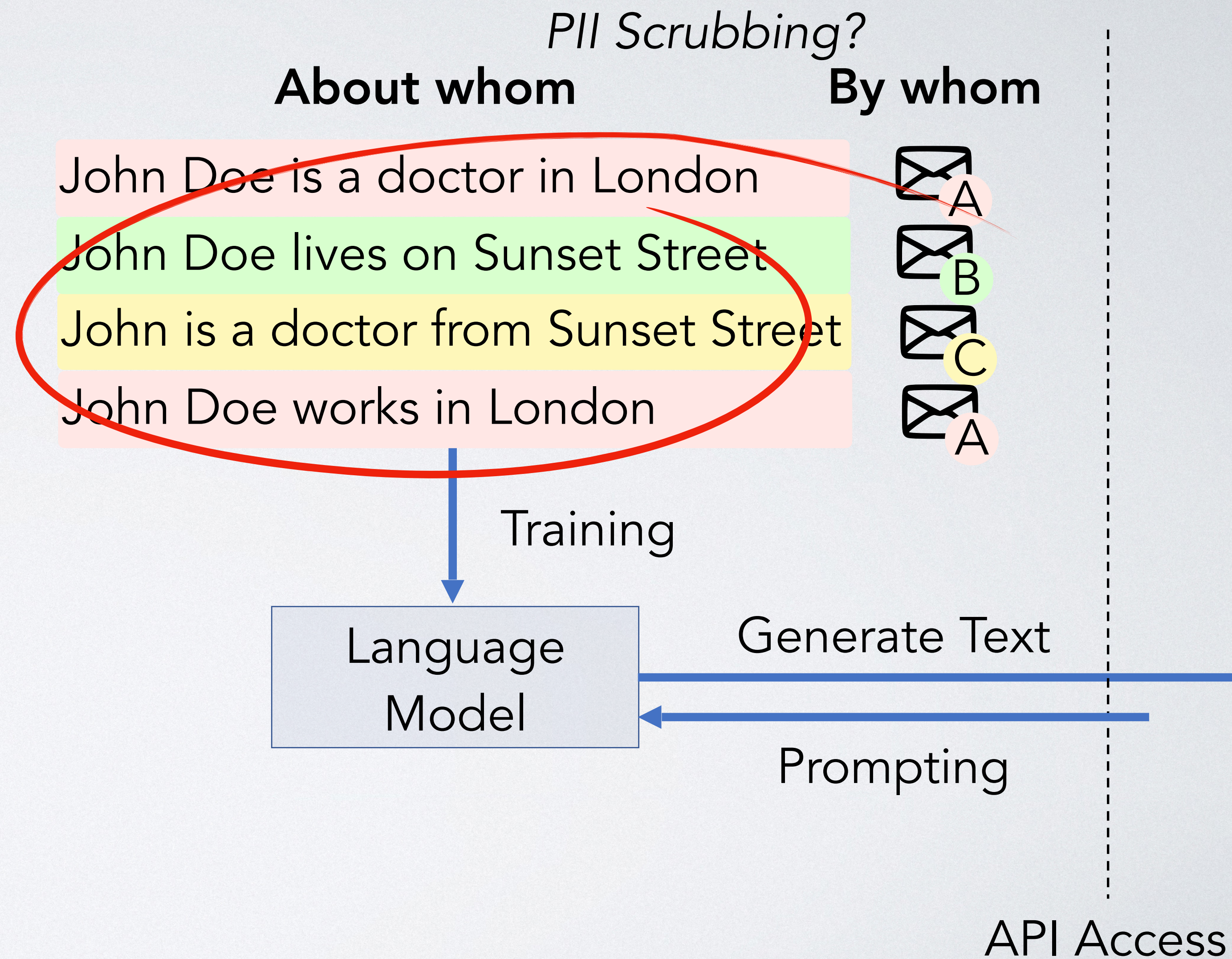
Problems with Differential Privacy

Group-level DP can help but ..

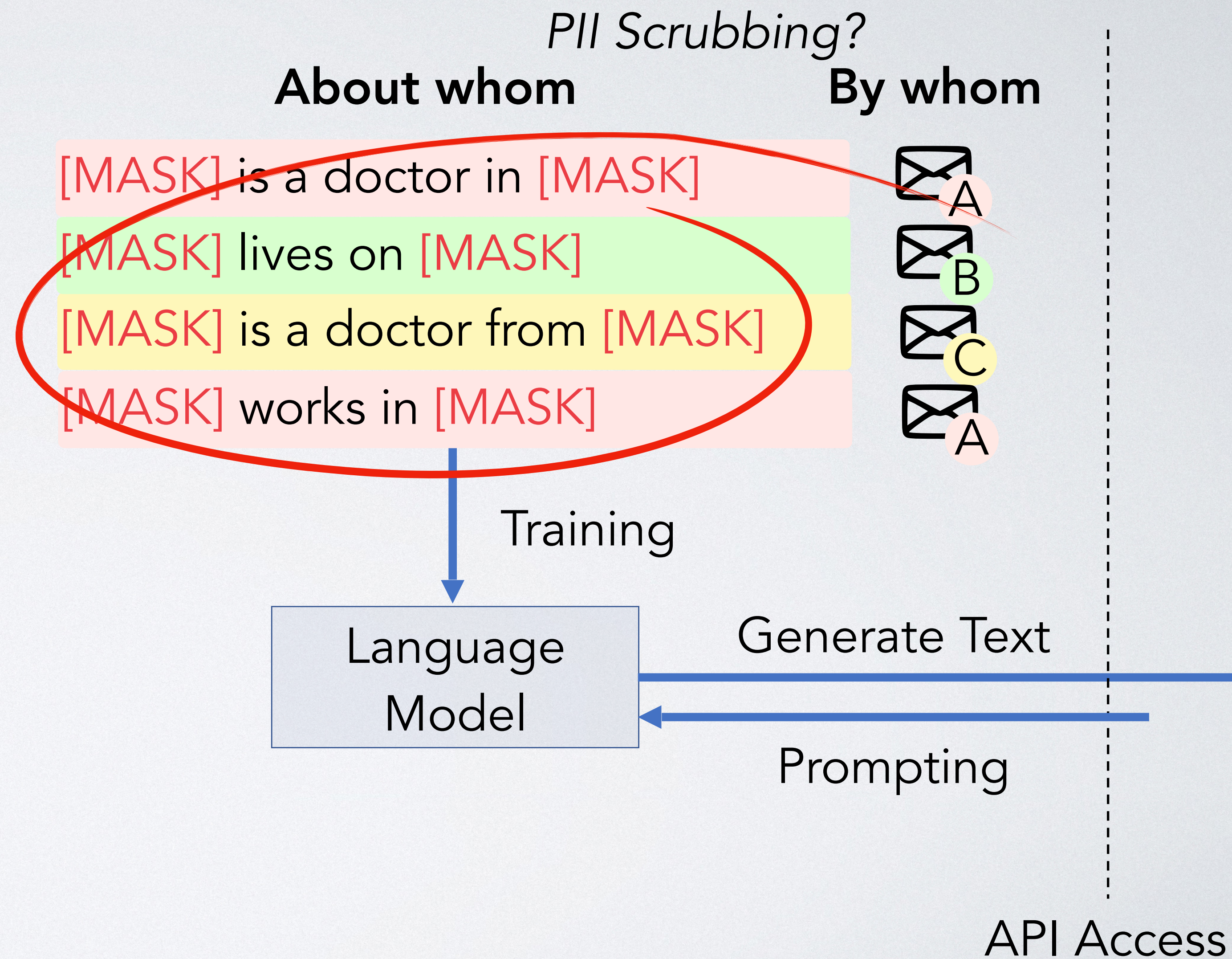
- 1) Group sizes are **not always** known a priori and under worst-case assumptions has **deleterious impact on model utility**.
- 2) PII Duplication across groups



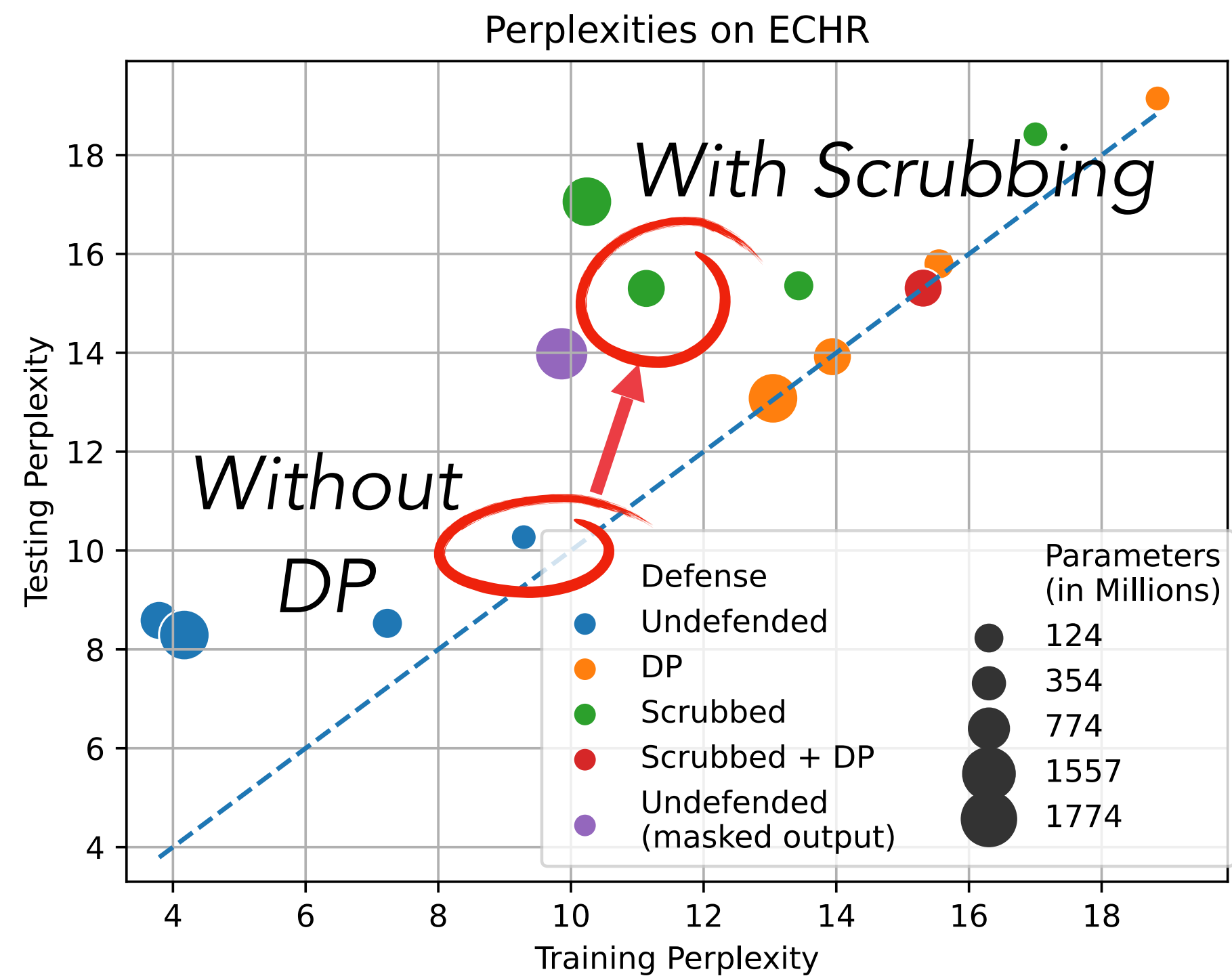
Problems with PII Scrubbing



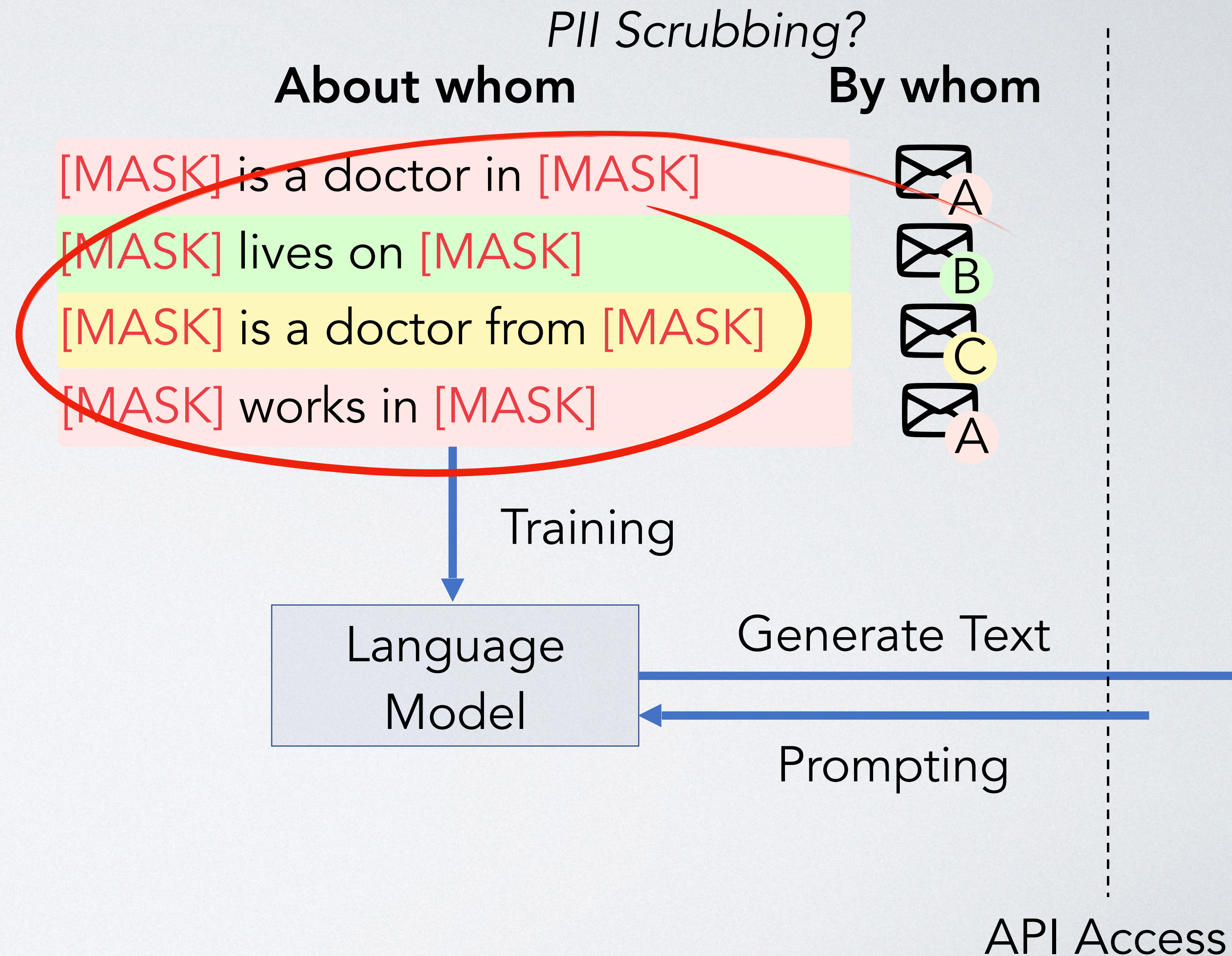
Problems with PII Scrubbing



Problems with PII Scrubbing

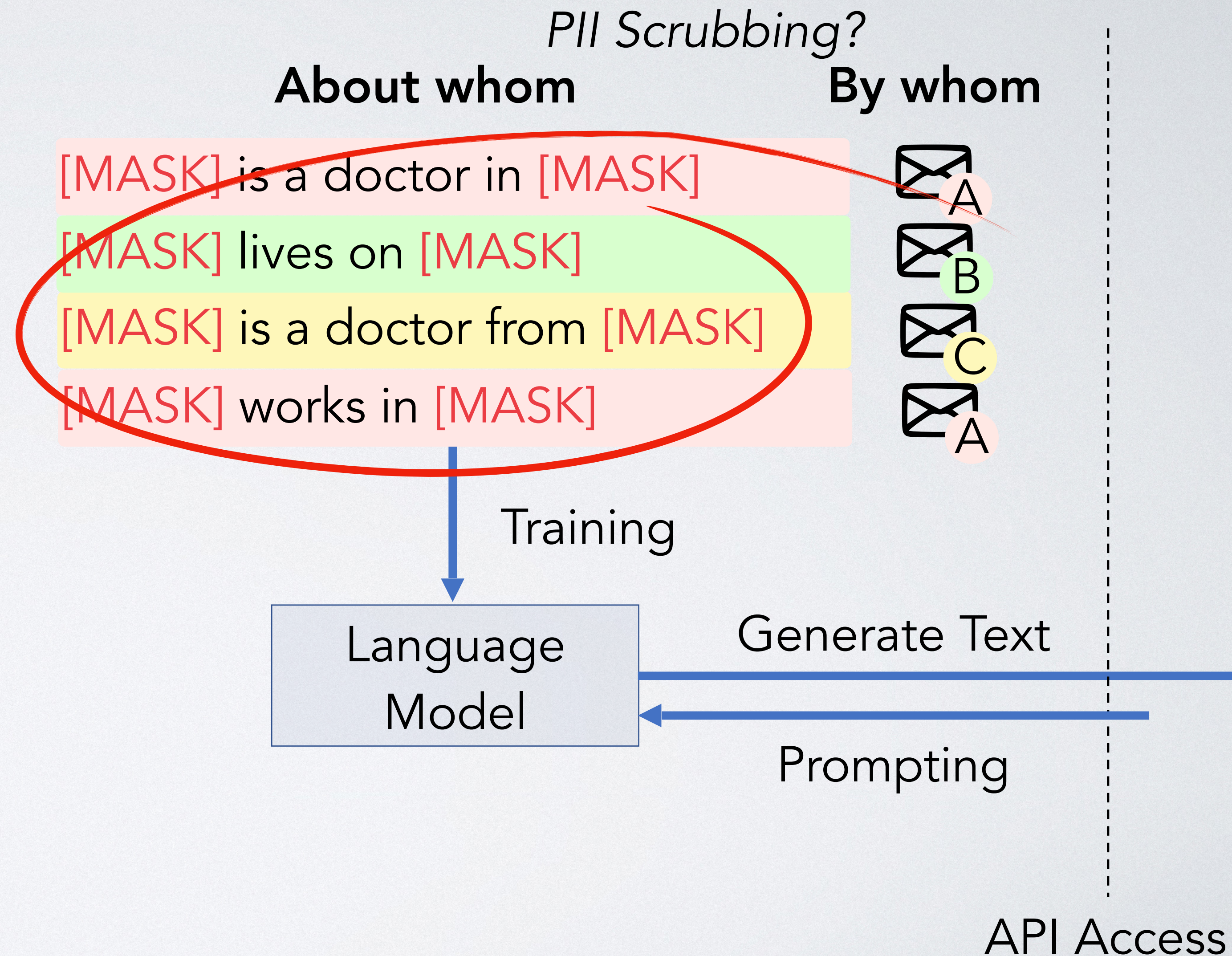


Privacy at the cost of Model Utility



Problems with PII Scrubbing

Methods to optimize the privacy/utility trade-off are missing.



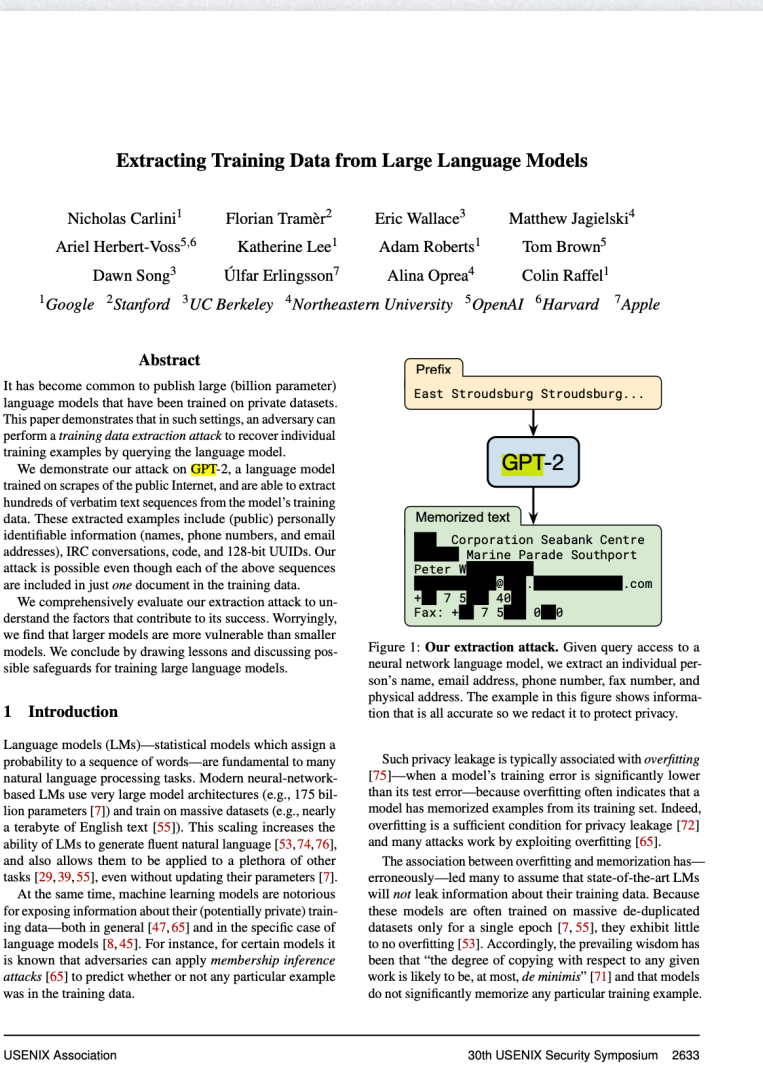
Related Work

Canaries

N-grams

Sequences

PII Leakage In Pre-Trained LMs



Carlini et al., 2019

McCoy et al., 2019

Carlini et al., 2020

Carlini et al., 2022

Huang et al., 2022

Related Work

Privacy in LMs

Is public data truly public?

- Data shared to intentionally violate someone’s privacy (e.g., “doxing”)
- Social media posts issued to a small target audience (“in-group sharing”)
- Accidental leakage of other’s information (e.g., “conversations”)

arXiv:2202.05520v2 [stat.ML] 14 Feb 2022

What Does it Mean for a Language Model to Preserve Privacy?

Hannah Brown¹, Katherine Lee², Fatemehsadat Mireshghallah³
Reza Shokri¹, Florian Tramèr^{4*}

¹National University of Singapore, ²Cornell University
³University of California San Diego, ⁴Google
{hsbrown, reza}@comp.nus.edu.sg kate.lee168@gmail.com
fatemeh@ucsd.edu tramer@google.com

Abstract

Natural language reflects our private lives and identities, making its privacy concerns as broad as those of real life. Language models lack the ability to understand the context and sensitivity of text, and tend to memorize phrases present in their training sets. An adversary can exploit this tendency to extract training data. Depending on the nature of the content and the context in which this data was collected, this could violate expectations of privacy. Thus, there is a growing interest in techniques for training language models that *preserve privacy*. In this paper, we discuss the mismatch between the narrow assumptions made by popular data protection techniques (data sanitization and differential privacy), and the broadness of natural language and of privacy as a social norm. We argue that existing protection methods cannot guarantee a generic and meaningful notion of privacy for language models. We conclude that language models should be trained on text data which was explicitly produced for public use.

1 Introduction

We use natural language to construct identities and communicate all our information in day-to-day life. Humans naturally understand when sharing a sensitive piece of information is appropriate based on context. It may be fine to share the same piece of information with one specific person or group, and a complete violation of privacy to share in another context, or at another point in time. Between humans, we trust that these implicit boundaries will be recognized and respected. As we build technologies that collect, store, and process our natural language communication, it is important that these technologies do not violate human notions of privacy or make use of data in ways beyond what is needed for the utility of the technology [71, 101].

Language models (LMs) underlie much natural language technology we regularly interact with, from autocorrect to search engines and translation systems. Over the past few years, LMs have grown in size and now utilize unprecedentedly large datasets of natural language making privacy risks in LMs a far reaching problem. Prior work has already demonstrated that such models are prone to memorizing and regurgitating large portions of their training data [12, 13, 51, 38, 91]. Worse, they are especially likely to memorize atypical data points—which are more likely to represent privacy risks for the authors or subjects of these texts.

To address these privacy concerns, there is a growing body of literature that aims to create *privacy-preserving* language models [64, 2, 56, 98, 84, 40, 79]. While humans navigate the complexities of language and privacy by identifying appropriate contexts for sharing information, LMs are not currently designed to do this [14, 72, 66, 49, 66, 50, 41]. Instead, the approach to preserving privacy in LMs has been to *attempt* complete removal of private information from training data (data sanitization), or to design algorithms that do not memorize private data, such as algorithms that satisfy differential privacy (DP) [28, 26].

Both methods make explicit and implicit assumptions about the structure of data to be protected, the nature of private information, and requirements for privacy, that do not hold for the majority of natural language data. Sanitization techniques assume that private information can

*Authors appear in alphabetical order

1

Brown et al., 2022

Security Games for PII Leakage

Algorithm 8 Sentence-level MI (lines enclosed in solid box) vs. PII Inference (lines enclosed in dashed box).

```

1: experiment IND-INFERENCE( $\mathcal{T}, \mathcal{D}, n, \mathcal{A}$ )
2:    $b \sim \{0, 1\}$ 
3:    $D \sim \mathcal{D}^n$ 
4:    $\theta \leftarrow \mathcal{T}(D)$ 
5:    $S_0 \sim D$ 
6:    $S_1 \sim \mathcal{D}$ 
7:    $\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), S_b)$ 
8:    $S \sim \{S \in D \mid \text{EXTRACT}(S) \neq \emptyset\}$ 
9:    $C_0 \sim \text{EXTRACT}(S)$ 
10:   $C_1 \sim \mathcal{E}$ 
11:   $\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), \text{SCRUB}(\text{SPLIT}(S, C_0), C_1))$ 

```

Algorithm 5 PII Reconstruction Game

```

1: experiment RECONSTRUCTION( $\mathcal{T}, \mathcal{D}, n, \mathcal{A}$ )
2:    $D \sim \mathcal{D}^n$ 
3:    $\theta \leftarrow \mathcal{T}(D)$ 
4:    $S \sim \{S \in D \mid \text{EXTRACT}(S) \neq \emptyset\}$ 
5:    $C \sim \text{EXTRACT}(S)$ 
6:    $\tilde{C} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), \text{SCRUB}(\text{SPLIT}(S, C)))$ 

```

Algorithm 2 PII Extraction

```

1: experiment EXTRACTION( $\mathcal{T}, \mathcal{D}, n, \mathcal{A}$ )
2:    $D \sim \mathcal{D}^n$ 
3:    $\theta \leftarrow \mathcal{T}(D)$ 
4:    $\mathcal{C} \leftarrow \bigcup_{S \in D} \text{EXTRACT}(S)$ 
5:    $\tilde{\mathcal{C}} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), \mathcal{C})$ 
1: procedure  $\mathcal{O}_\theta(S)$ 
2:   return  $\{w \mapsto \Pr(w|S; \theta)\}_{w \in \mathcal{V}}$ 

```

Algorithm 7 PII Inference Game

```

1: experiment INFERENCE( $\mathcal{T}, \mathcal{D}, n, m, \mathcal{A}$ )
2:    $D \sim \mathcal{D}^n$ 
3:    $\theta \leftarrow \mathcal{T}(D)$ 
4:    $S \sim \{S \in D \mid \text{EXTRACT}(S) \neq \emptyset\}$ 
5:    $C \sim \text{EXTRACT}(S)$ 
6:    $\mathcal{C} \sim \mathcal{E}^m$ 
7:    $\tilde{\mathcal{C}} \leftarrow \mathcal{C} \cup \{C\}$ 
8:    $\tilde{C} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), \text{SCRUB}(\text{SPLIT}(S, C), \mathcal{C}))$ 

```

See our paper for more details

Reconstruction Attack

Real Sentence

In early September 2023 [MASK] wrote in his memoir that he had again developed pneumonia.

Prompt

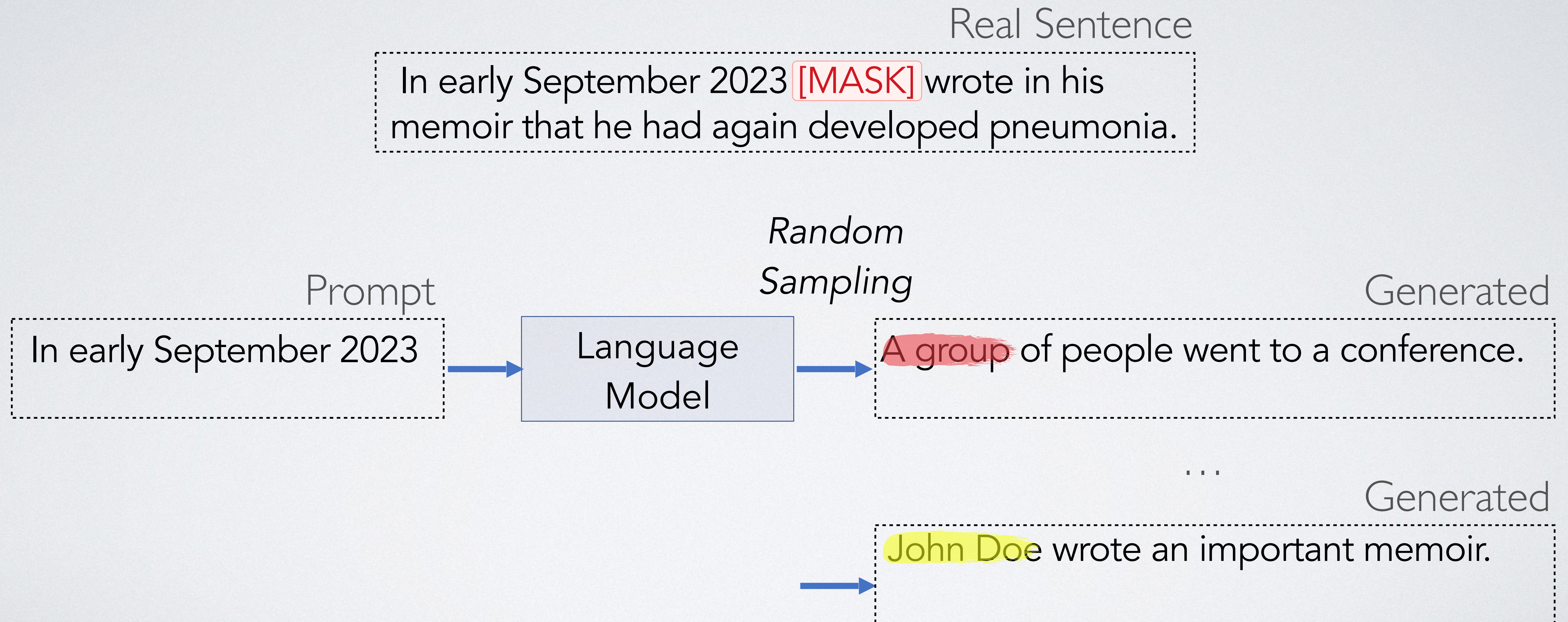
In early September 2023

Language
Model

Generated

A group of people went to a conference.

Reconstruction Attack



Reconstruction Attack

Real Sentence

In early September 2023 [MASK] wrote in his memoir that he had again developed pneumonia.

Generated

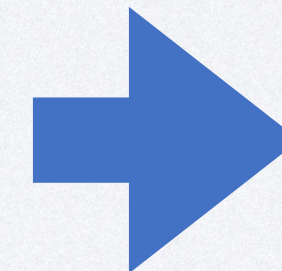
A group of people went to a conference.

...

Generated

John Doe wrote an important memoir.

Tag PII



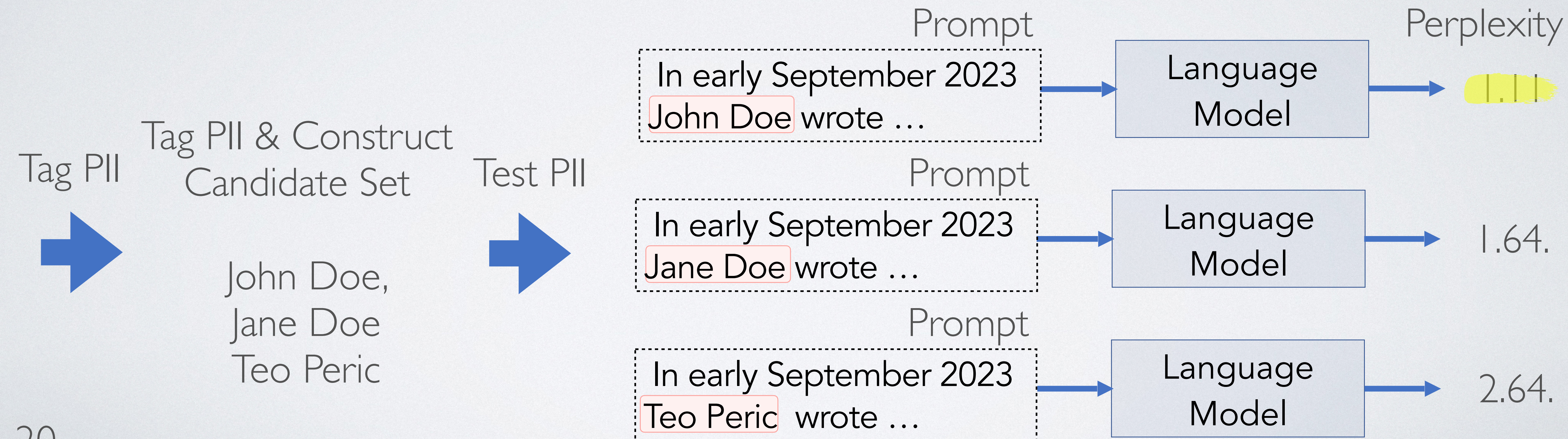
Tag PII & Construct
Candidate Set

John Doe,
Jane Doe
Teo Peric

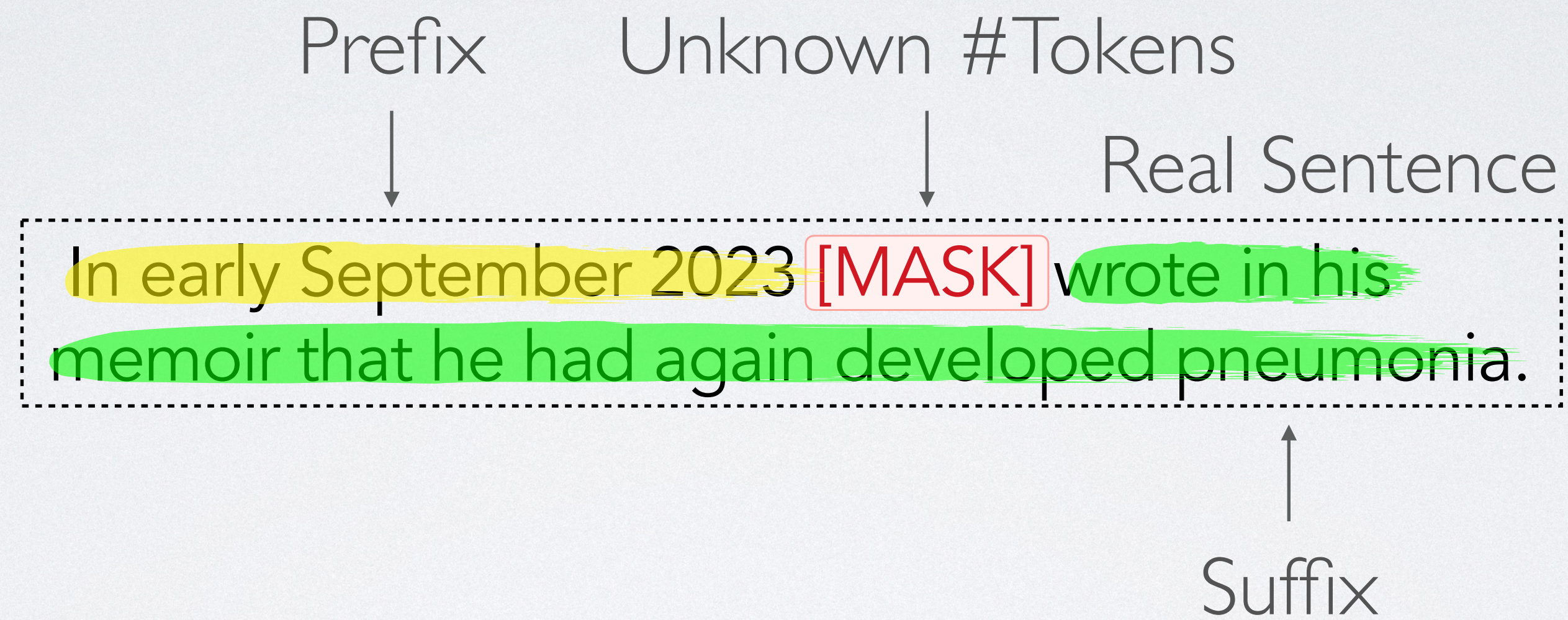
Reconstruction Attack

Real Sentence

In early September 2023 [MASK] wrote in his memoir that he had again developed pneumonia.



PII Reconstruction Tractability



Datasets

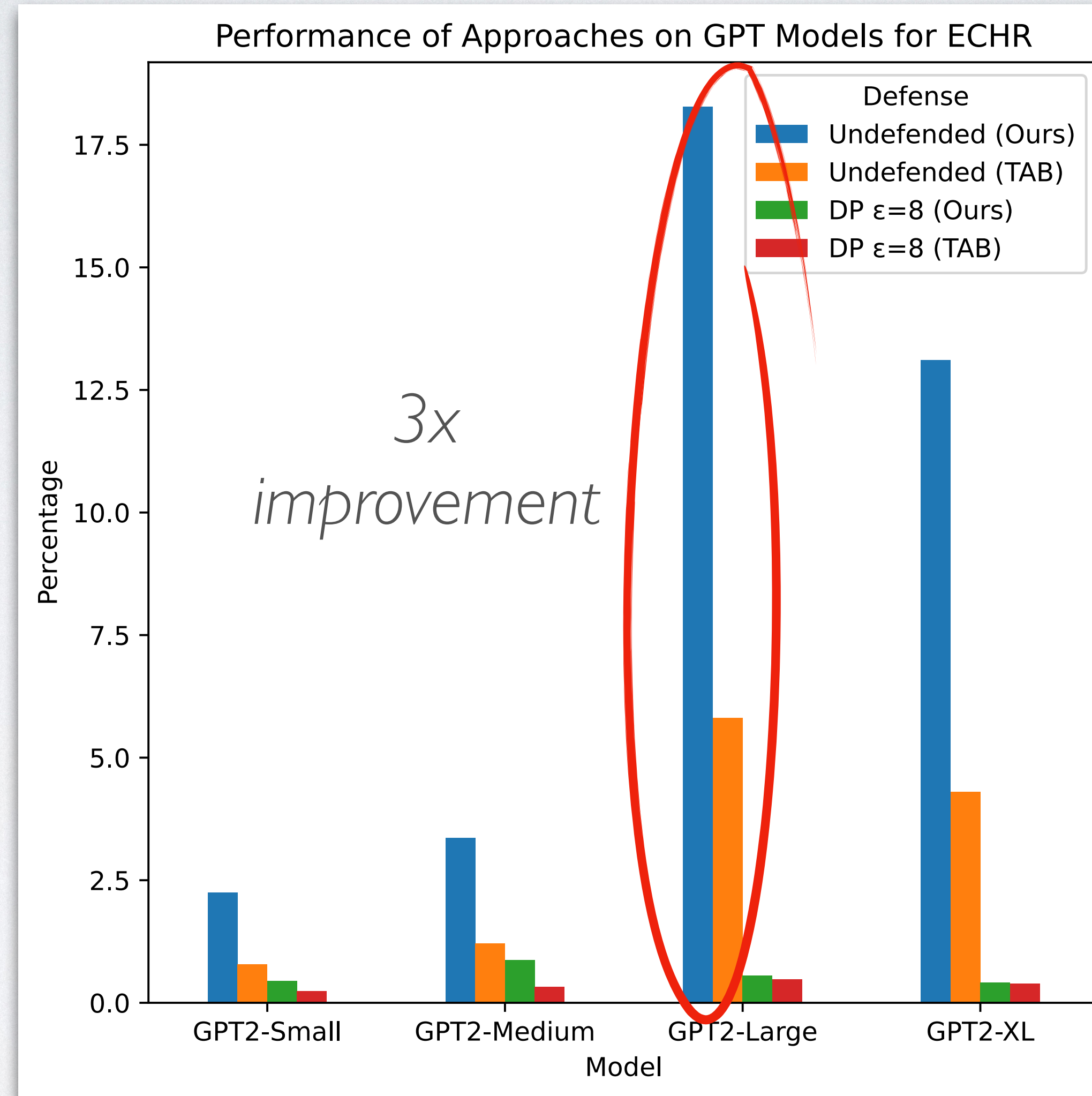
	Records	Tokens / Record	Unique PII	Records w. PII	Duplicates / PII	Tokens / PII
ECHR	118 161	88.12	16 133	23.75%	4.66	4.00
Enron	138 919	346.10	105 880	81.45%	11.68	3.00
Yelp-Health	78 794	143.92	17 035	54.55%	5.35	2.17

ECHR : European Court for Human Rights

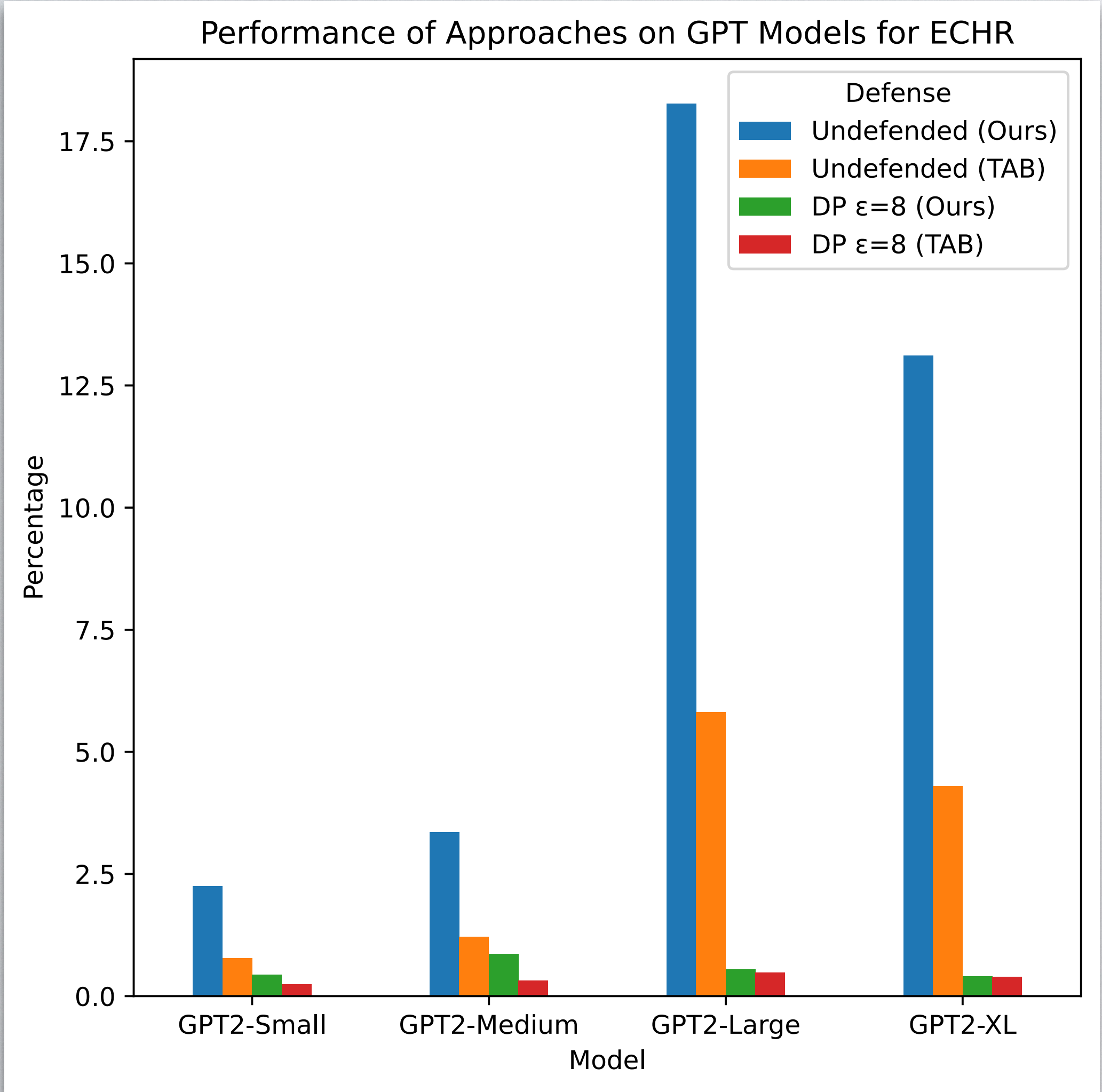
Enron : Corporate e-mails

Yelp-Health: Reviews for healthcare facilities

PII Reconstruction



PII Reconstruction



	GPT2-Small		GPT2-Medium		GPT2-Large		GPT2-XL	
	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$
ECHR(TAB)	0.78%	0.24%	1.21%	0.32%	5.81%	0.48%	4.30%	0.39%
ECHR (Ours, $ \mathcal{C} = 64$)	2.25%	0.44%	3.36%	0.87%	18.27%	0.55%	13.11%	0.41%
Enron (TAB)	0.59%	0.04%	0.67%	0.04%	1.75%	0.04%	2.19%	0.19%
Enron (Ours, $ \mathcal{C} = 64$)	6.29%	0.49%	7.26%	0.52%	12.68%	0.55%	15.25%	0.53%
Yelp-Health (TAB)	0.33%	0.24%	0.37%	0.14%	0.65%	0.12%	1.99%	0.12%
Yelp-Health (Ours, $ \mathcal{C} = 64$)	0.42%	0.32%	1.31%	0.32%	1.69%	0.35%	6.40%	0.36%

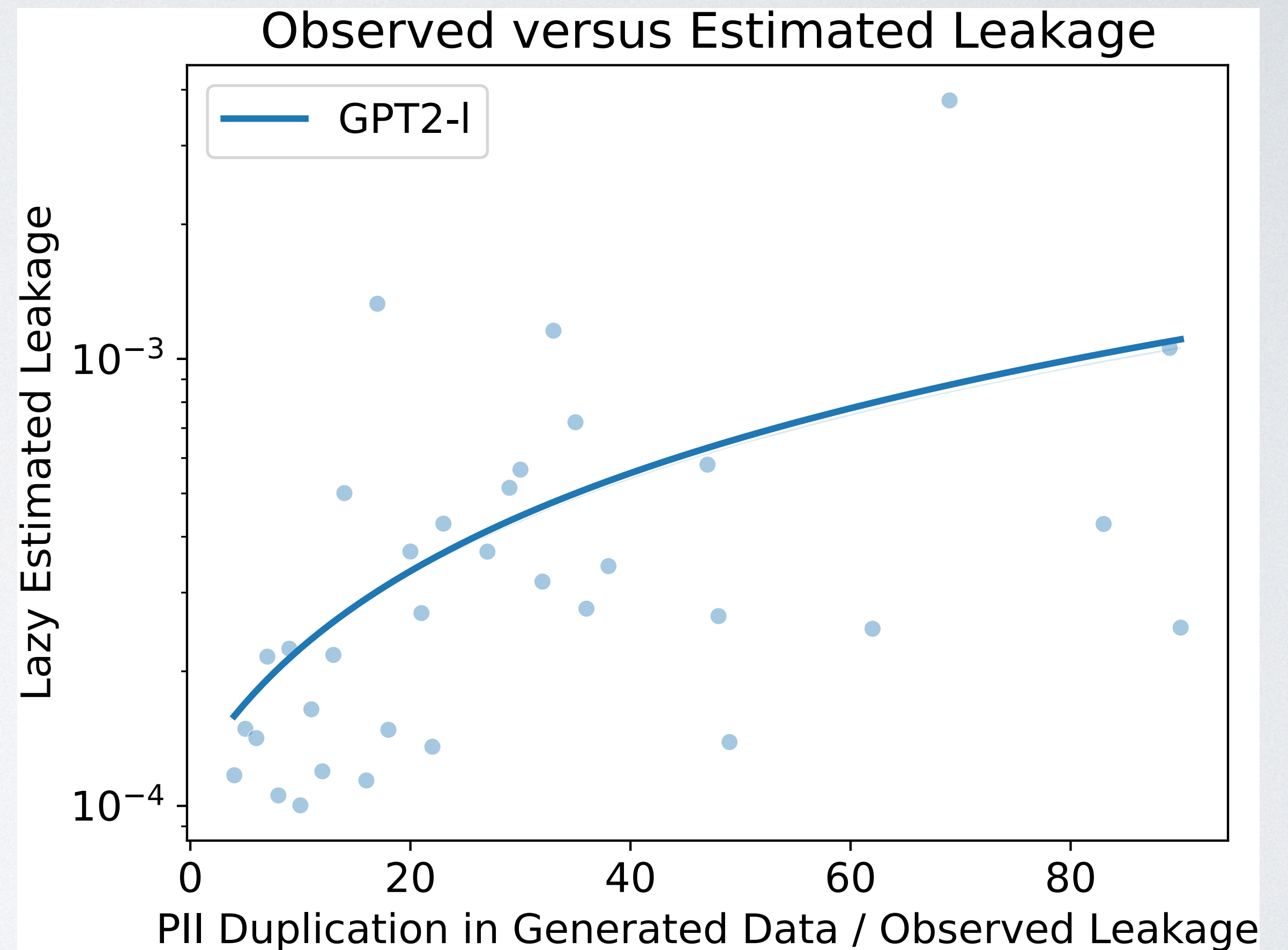
up to 7x
Improvement

PII Inference

	ECHR		Enron		Yelp-Health	
	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$
$ \mathcal{C} = 100$	70.11%	8.32%	50.50%	3.78%	28.31%	4.29%
$ \mathcal{C} = 500$	51.03%	3.71%	34.14%	1.92%	15.55%	1.86%

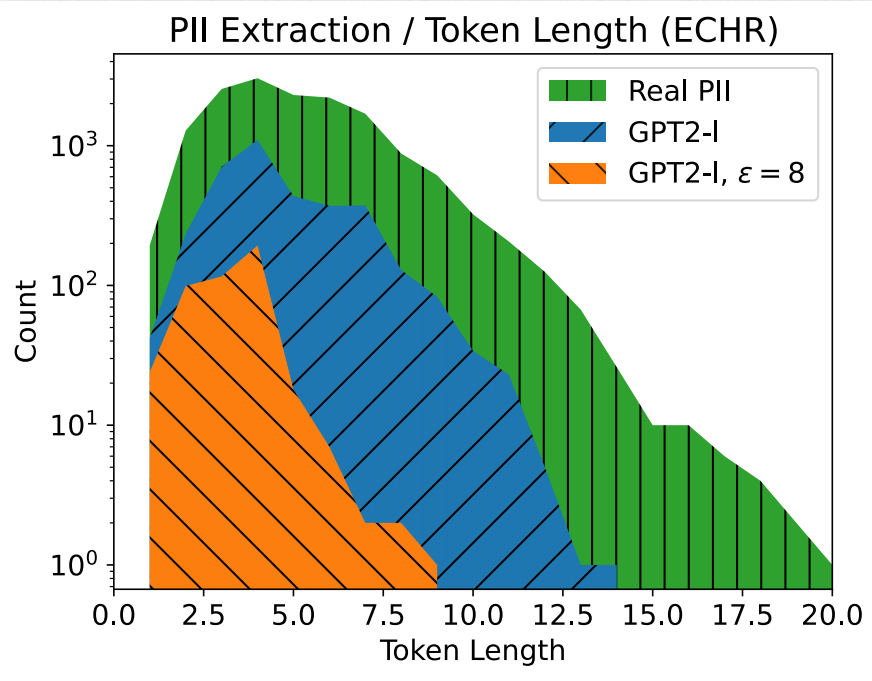
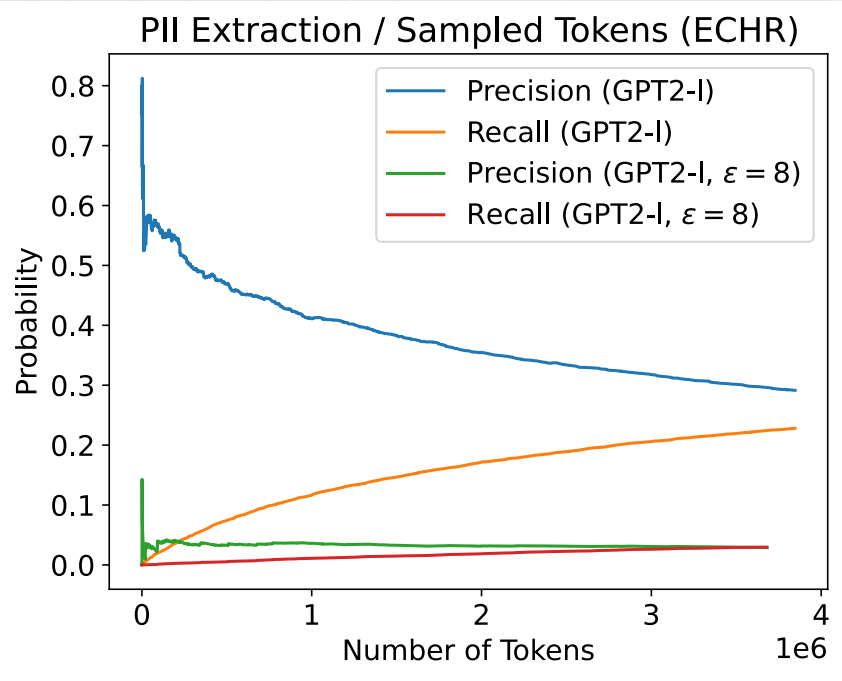
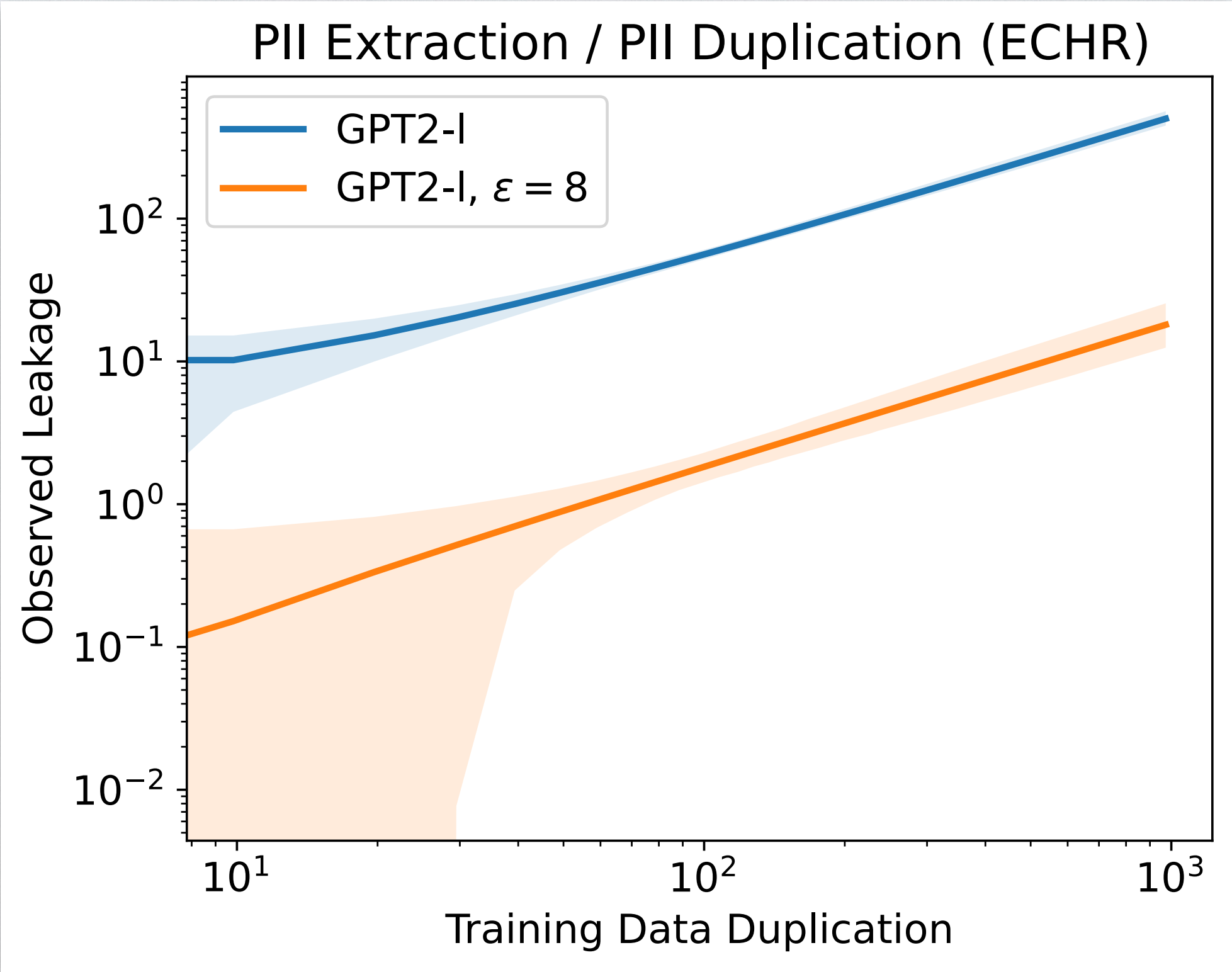
Extraction Attack

Once upon a time, there existed a tale of two medical students. In the year 2022, they resided at Sunset Street while pursuing their medical education. Alongside his friend, he worked at the LHS Hospital located in the bustling heart of downtown London. Before donning their white coats, both **John Doe** and ...



PII Extraction

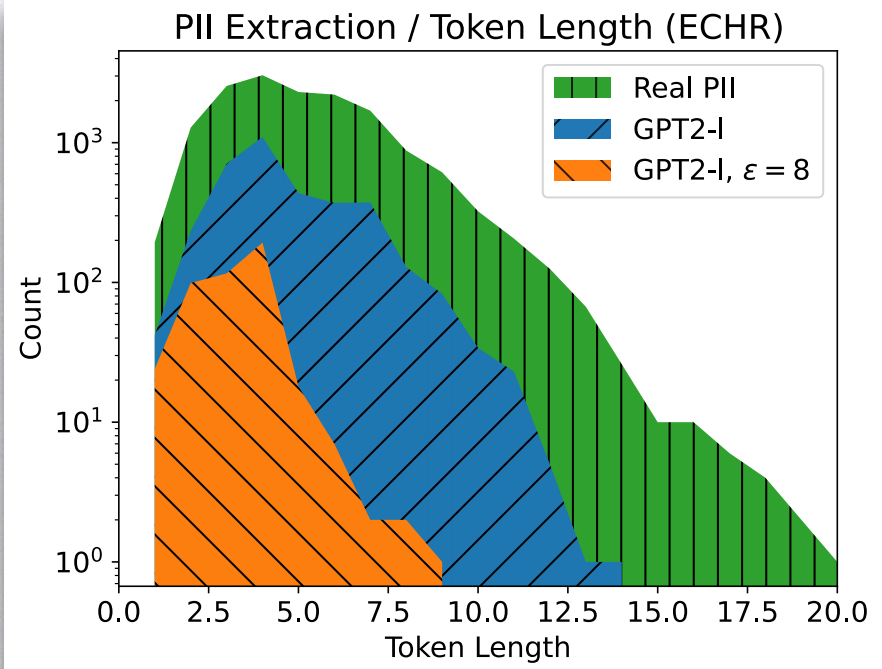
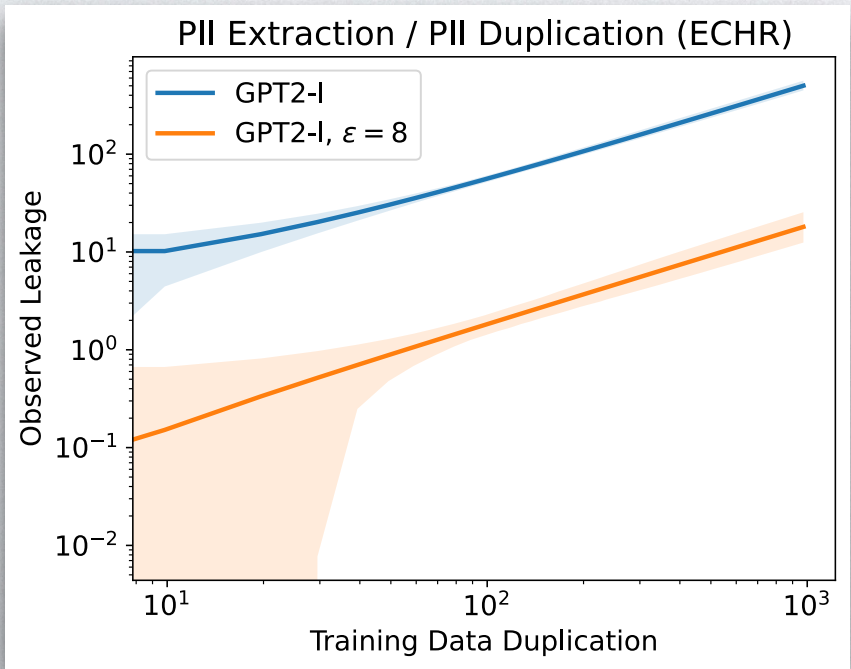
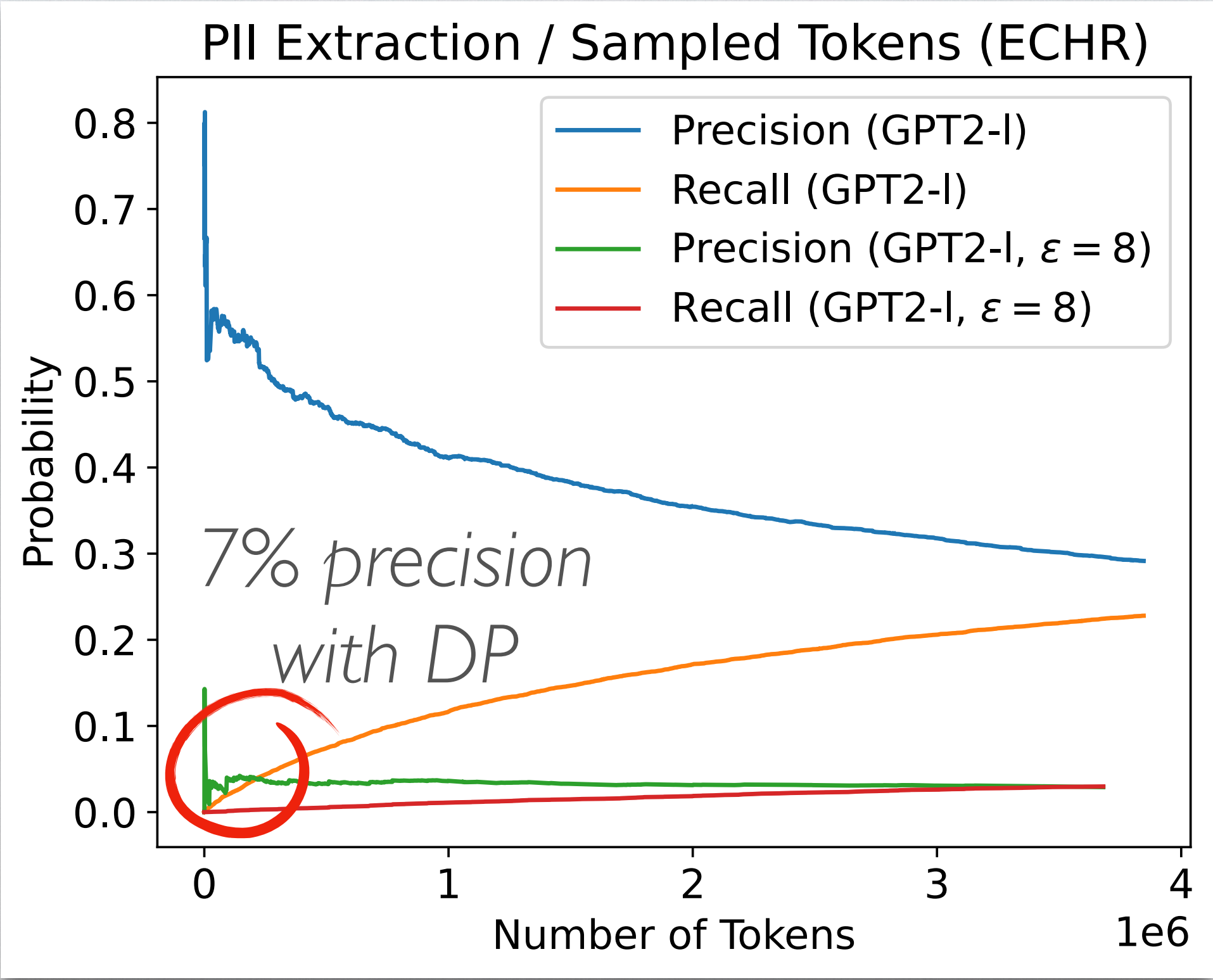
Duplicated PII are leaked more often



	GPT2-Small		GPT2-Medium		GPT2-Large	
	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$
ECHR						
Prec	24.91%	2.90%	28.05%	3.02%	29.56%	2.92%
Recall	9.44%	2.98%	12.97%	3.21%	22.96%	2.98%
Enron						
Prec	33.86 %	9.37%	27.06%	12.05%	35.36%	11.57%
Recall	6.26%	2.29%	6.56%	2.07%	7.23%	2.31%
Yelp-Health						
Prec	13.86%	8.31%	14.87%	6.32%	14.28%	7.67%
Recall	11.31%	5.02%	11.23%	5.22%	13.63%	6.51%

PII Extraction

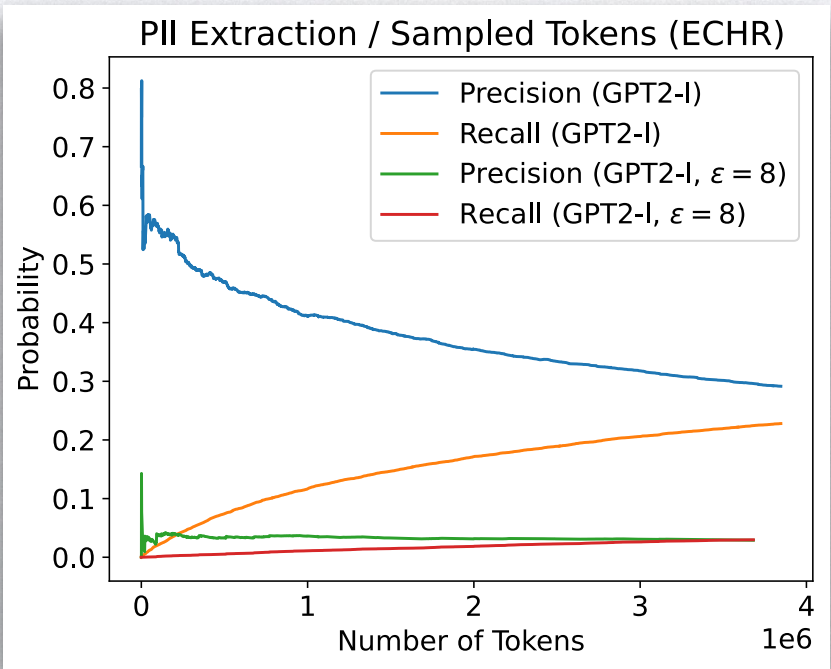
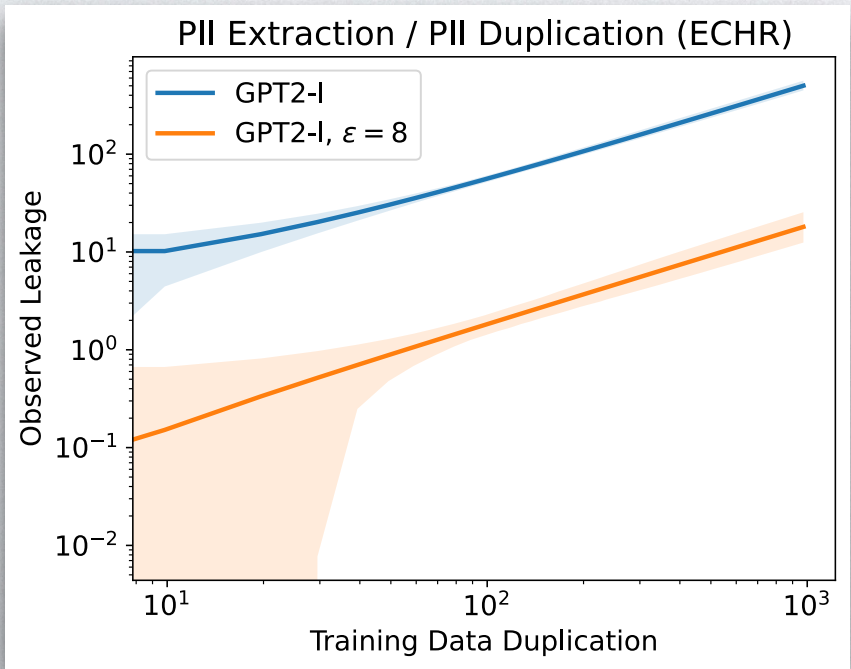
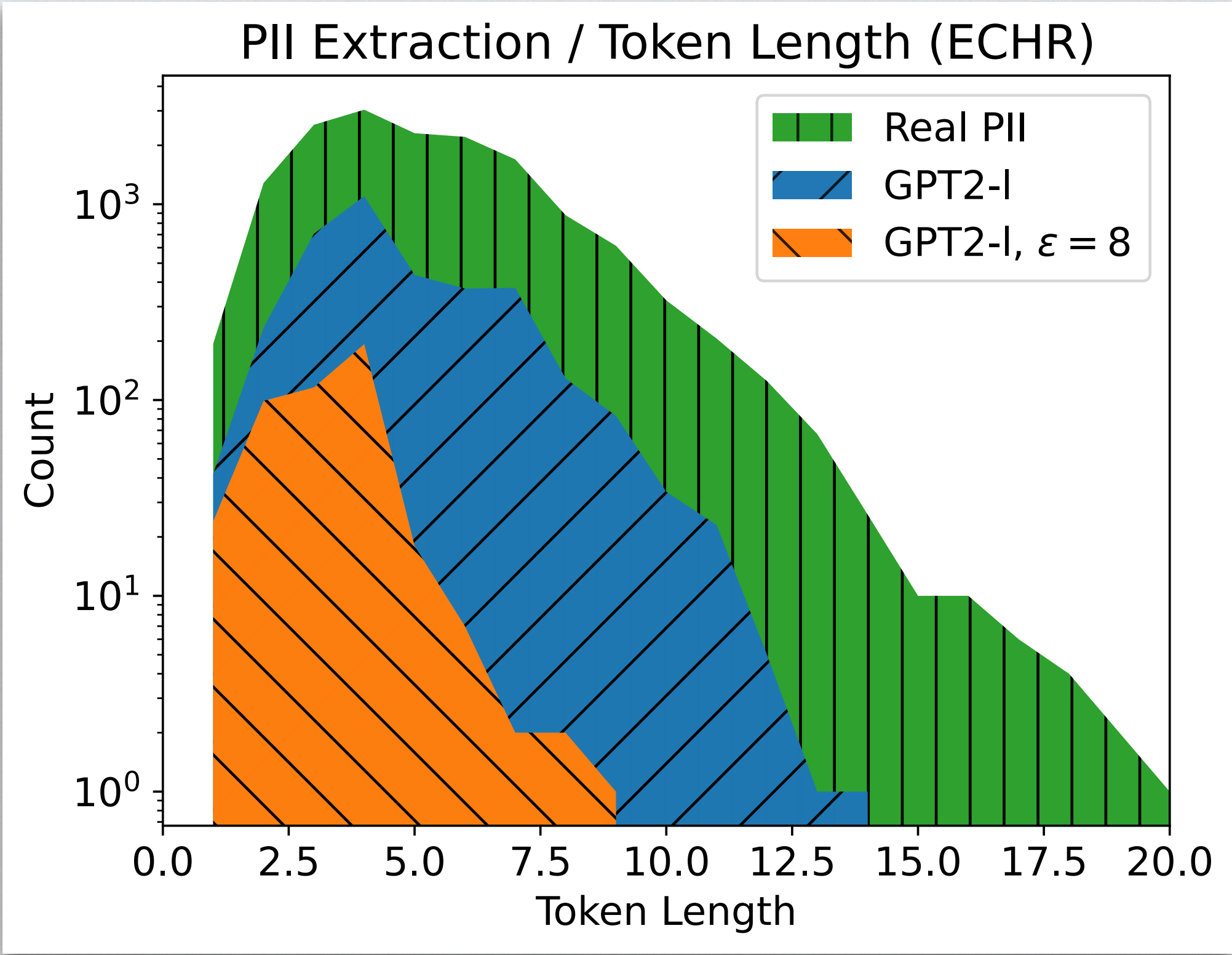
High-precision/
Low-recall attacks



	GPT2-Small		GPT2-Medium		GPT2-Large	
	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$
ECHR						
Prec	24.91%	2.90%	28.05%	3.02%	29.56%	2.92%
Recall	9.44%	2.98%	12.97%	3.21%	22.96%	2.98%
Enron						
Prec	33.86 %	9.37%	27.06%	12.05%	35.36%	11.57%
Recall	6.26%	2.29%	6.56%	2.07%	7.23%	2.31%
Yelp-Health						
Prec	13.86%	8.31%	14.87%	6.32%	14.28%	7.67%
Recall	11.31%	5.02%	11.23%	5.22%	13.63%	6.51%

PII Extraction

PII with many tokens
are protected in DP models

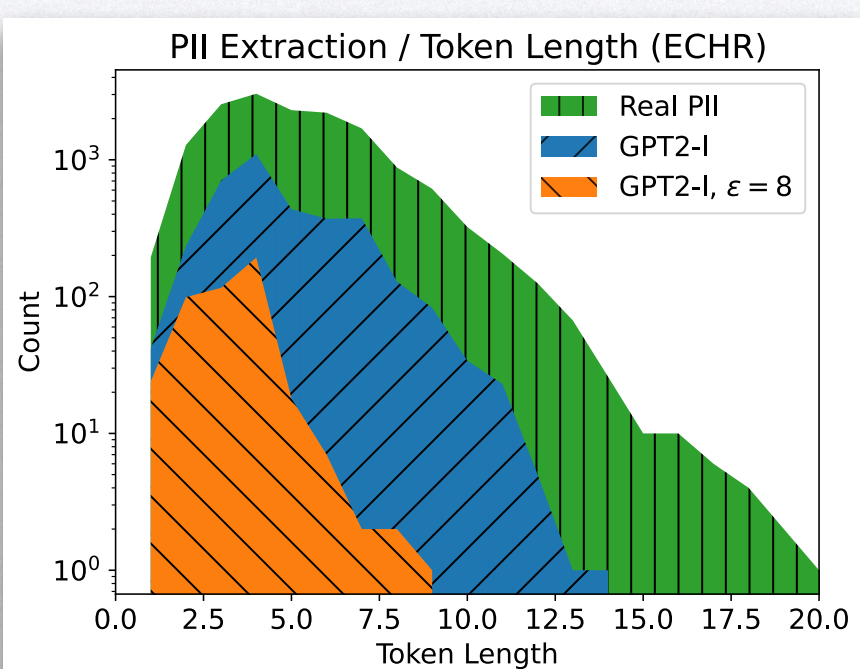
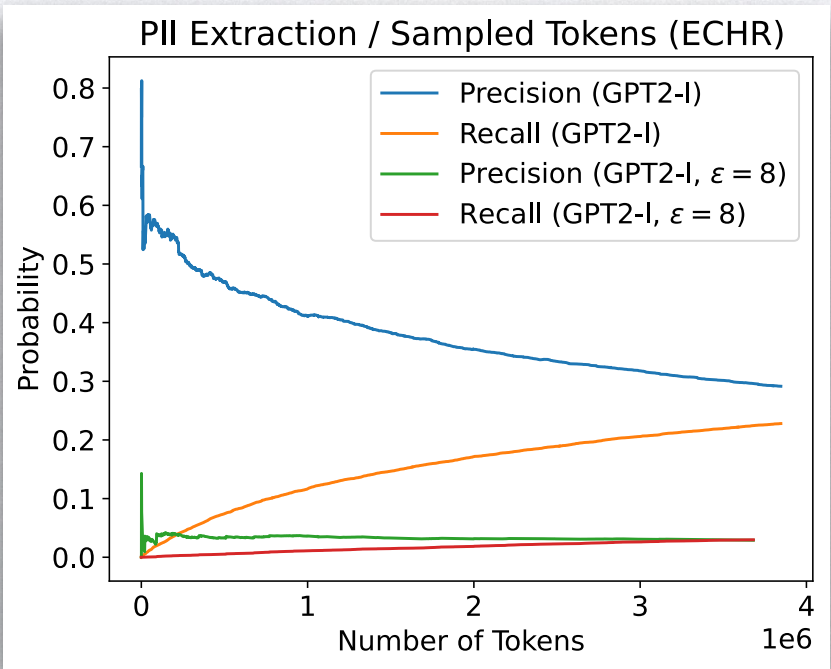
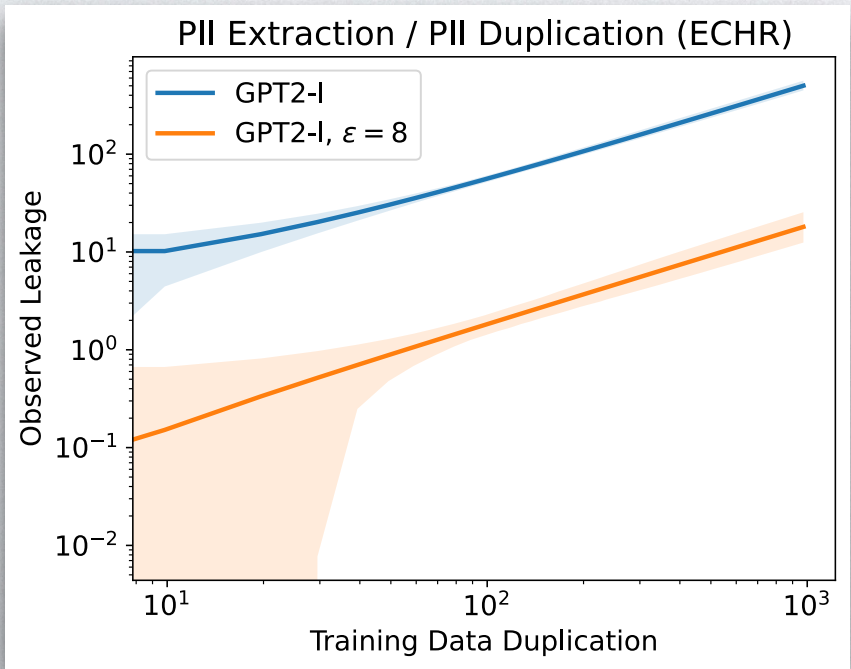


	GPT2-Small		GPT2-Medium		GPT2-Large	
	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$
ECHR						
Prec	24.91%	2.90%	28.05%	3.02%	29.56%	2.92%
Recall	9.44%	2.98%	12.97%	3.21%	22.96%	2.98%
Enron						
Prec	33.86 %	9.37%	27.06%	12.05%	35.36%	11.57%
Recall	6.26%	2.29%	6.56%	2.07%	7.23%	2.31%
Yelp-Health						
Prec	13.86%	8.31%	14.87%	6.32%	14.28%	7.67%
Recall	11.31%	5.02%	11.23%	5.22%	13.63%	6.51%

PII Extraction

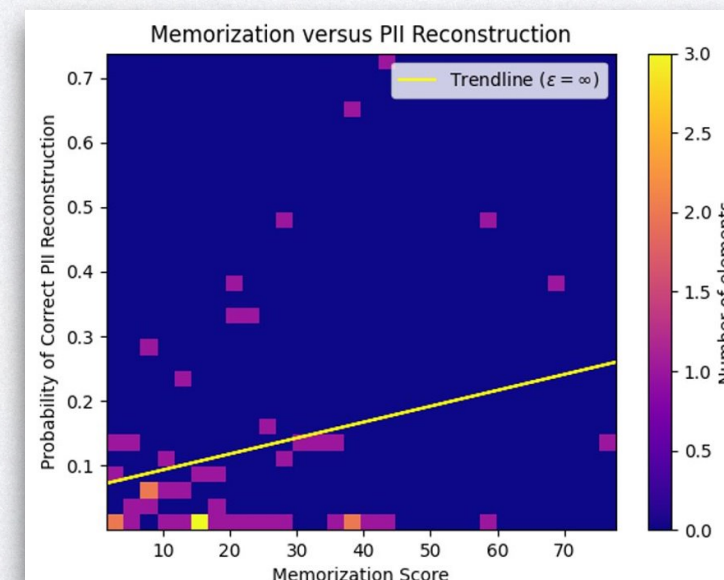
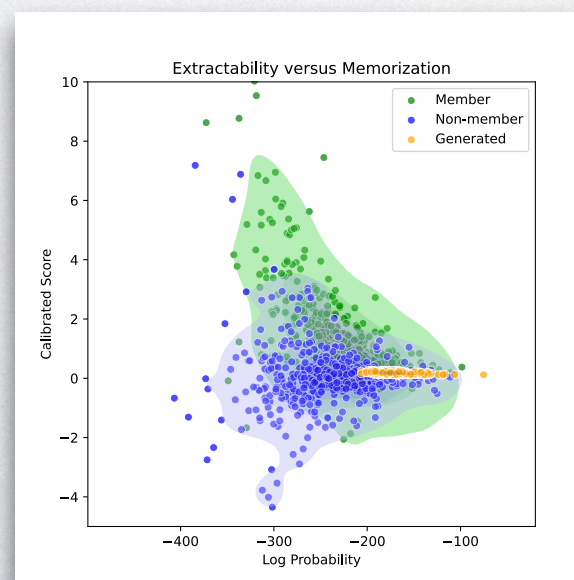
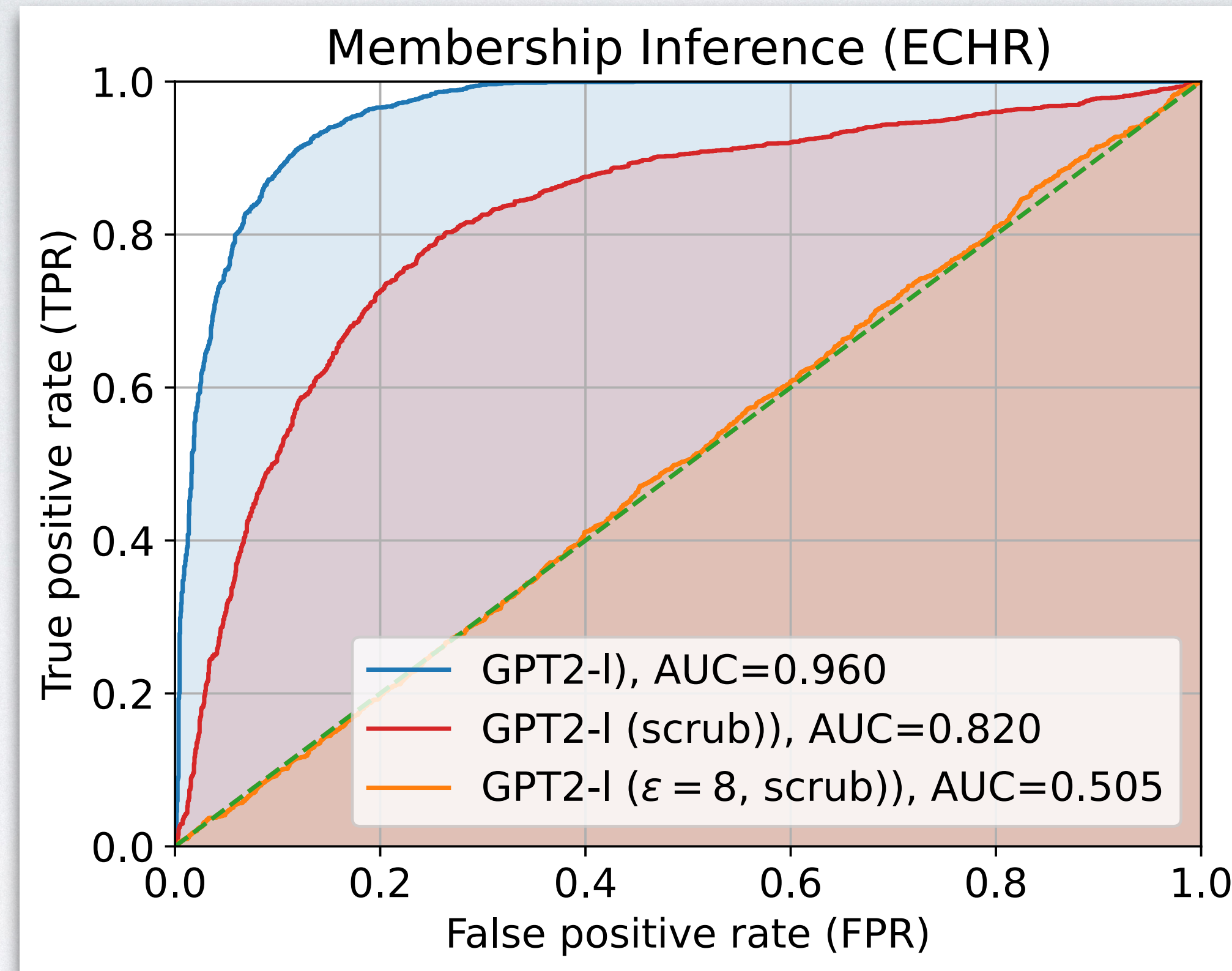
Higher recall in
larger models

	GPT2-Small		GPT2-Medium		GPT2-Large	
	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$	No DP	$\epsilon = 8$
ECHR						
Prec	24.91%	2.90%	28.05%	3.02%	29.56%	2.92%
Recall	9.44%	2.98%	12.97%	3.21%	22.96%	2.98%
Enron						
Prec	33.86 %	9.37%	27.06%	12.05%	35.36%	11.57%
Recall	6.26%	2.29%	6.56%	2.07%	7.23%	2.31%
Yelp-Health						
Prec	13.86%	8.31%	14.87%	6.32%	14.28%	7.67%
Recall	11.31%	5.02%	11.23%	5.22%	13.63%	6.51%



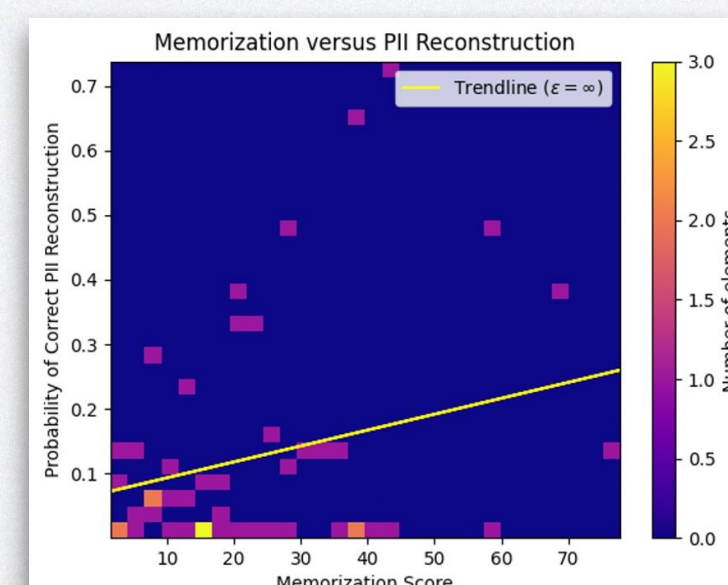
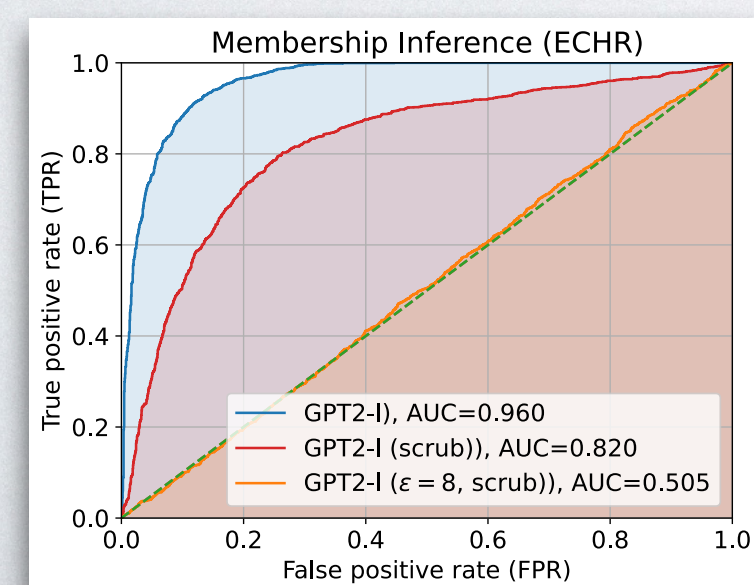
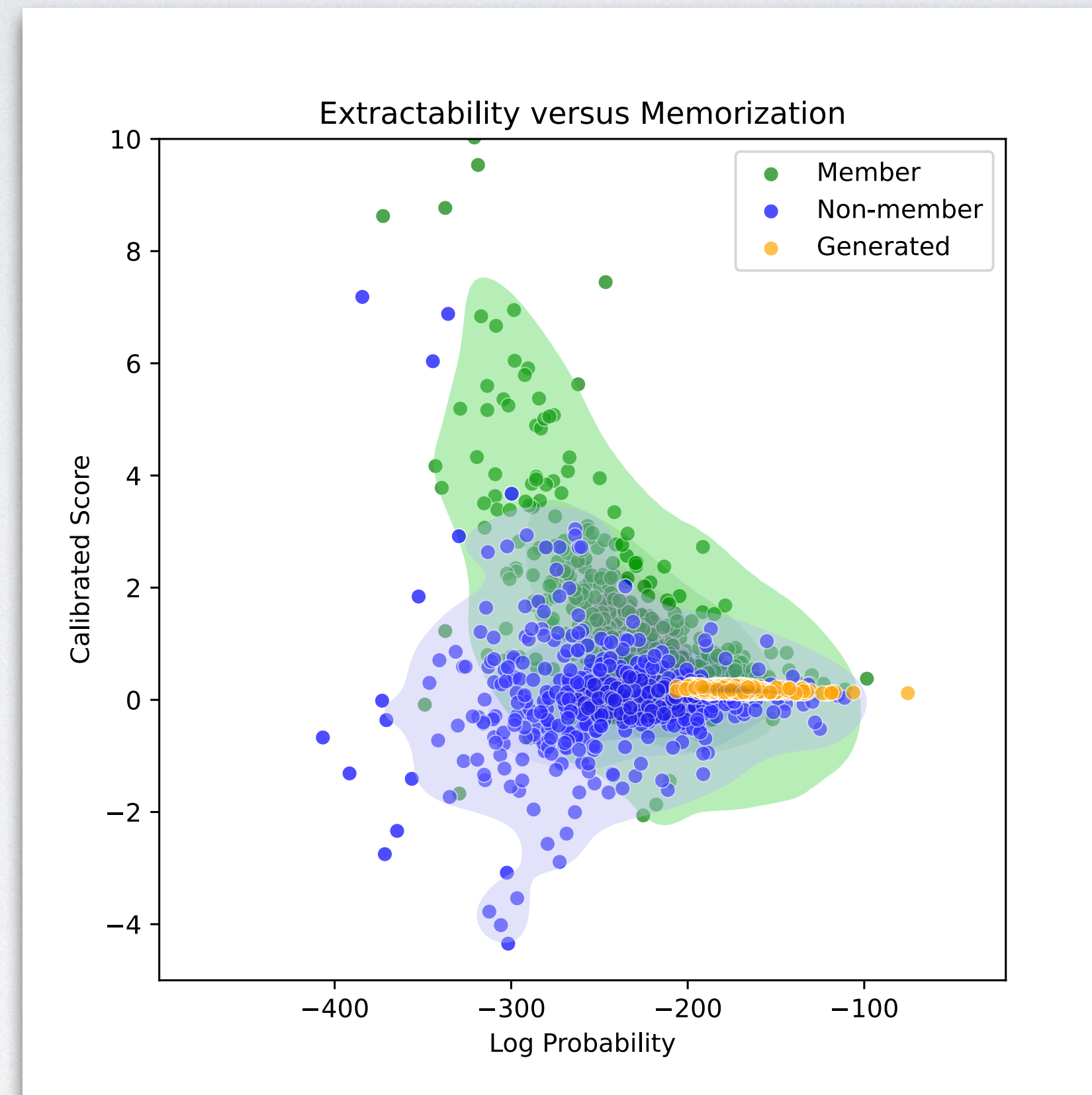
Membership Inference

Scrubbing does not prevent MI



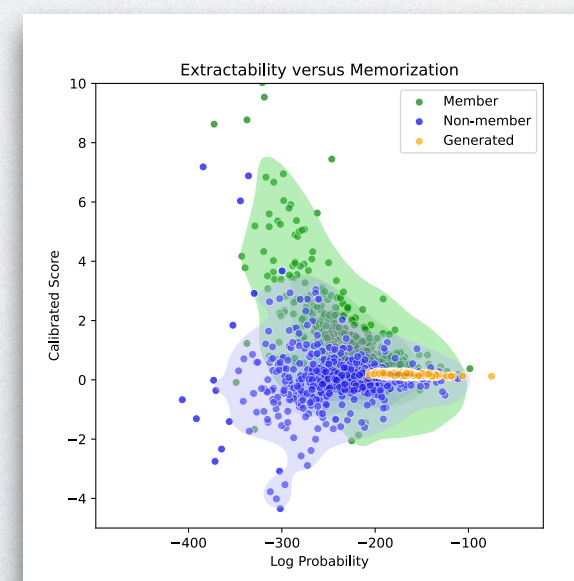
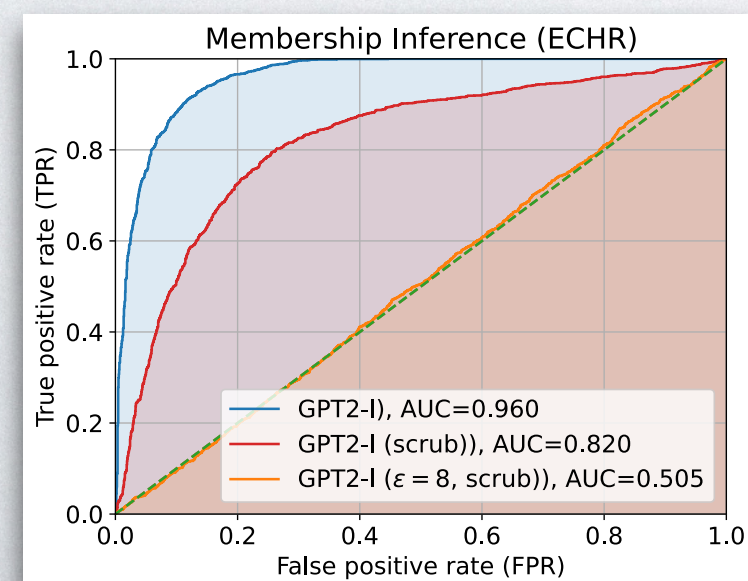
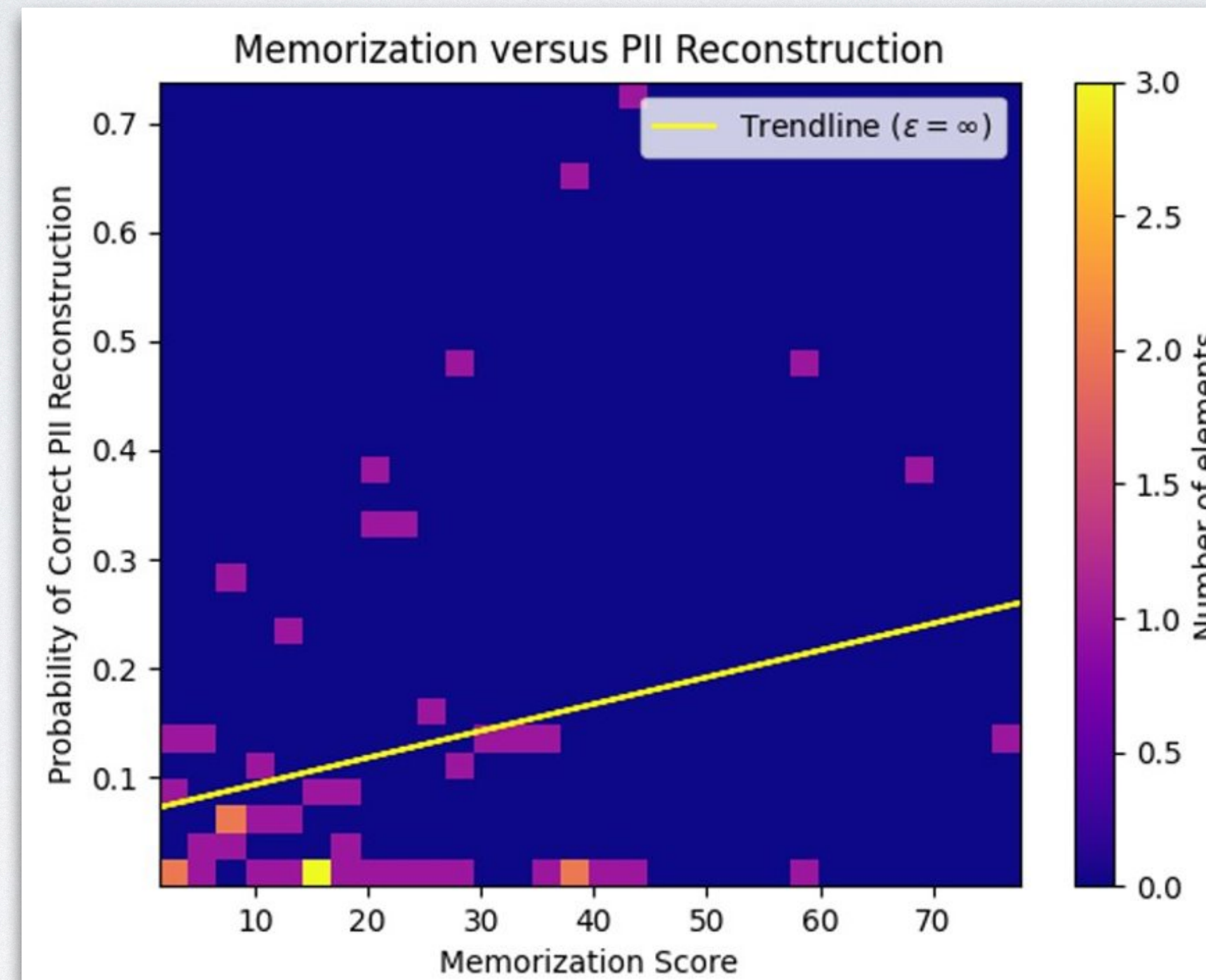
Membership Inference

Randomly generated sequences likely do not contain MI signal



Membership Inference

MI correlates with
PII reconstruction



Summary of Results

Undefended models are highly
Vulnerable to all privacy attacks

DP bounds, but does not prevent
the leakage of PII

Aggressive scrubbing harms utility
and can miss PII (more data needed)

Motivates search for methods with better
Empirical privacy/utility trade-off

	Undefended	DP	Scrub	DP + Scrub
Test Perplexity	9	14	16	16
Extract Precision	30%	3%	0%	0%
Extract Recall	23%	3%	0%	0%
Reconstruction Acc.	18%	1%	0%	0%
Inference Acc. ($ \mathcal{C} = 100$)	70%	8%	1%	1%
MI AUC	0.96	0.5	0.82	0.5

Limitations

- **(General Applicability)** We focus on fine-tuned **GPT-2** Language Models (0.12b to 1.7b parameters).
- **(Syntactic Similarity)** We consider only verbatim leakage (i.e., “John Doe” and “J. Doe” are different)
 - **(PII Association)** Our *extraction* attacks study leakage in isolation (single PII, no association)
 - **(Need for better Benchmarks)** Our study is limited by the quality of the NER tools used; Evaluating scrubbing methods requires large, annotated datasets

Outlook

We take a number of steps to reduce the risk that our models are used in a way that could violate a person's privacy rights. These include **fine-tuning models** to reject these types of requests, **removing personal information** from the training dataset where feasible, creating **automated model evaluations**, **monitoring** and responding to user attempts to generate this type of information, and restricting this type of use in our **terms and policies**. Our efforts to expand context length and improve embedding models for retrieval may help further limit privacy risks moving forward by tying task performance more to the information a user brings to the model. We continue to research, develop, and enhance technical and process mitigations in this area.

GPT-4 Technical Report, 2023 [8]

- 1) Fine-tuning to reject requests
- 2) Data sanitation
- 3) Model evaluation**
- 4) Query Monitoring (Post-Processing)
- 5) Terms of use

Check out our Paper for more Information

Analyzing Leakage of Personally Identifiable Information in Language Models

Nils Lukas^{*§}, Ahmed Salem[†], Robert Sim[†], Shruti Tople[†], Lukas Wutschitz[†] and Santiago Zanella-Béguelin[†]
^{*}University of Waterloo, [†]Microsoft
nlukas@uwaterloo.ca, {t-salemahmed, rsim, shruti.tople, lukas.wutschitz, santiago}@microsoft.com

Abstract—Language Models (LMs) have been shown to leak information about training data through sentence-level membership inference and reconstruction attacks. Understanding the risk of LMs leaking Personally Identifiable Information (PII) has received less attention, which can be attributed to the false assumption that dataset curation techniques such as scrubbing are sufficient to prevent PII leakage. Scrubbing techniques reduce but do not prevent the risk of PII leakage: in practice scrubbing is imperfect and must balance the trade-off between minimizing disclosure and preserving the utility of the dataset. On the other hand, it is unclear to which extent algorithmic defenses such as differential privacy, designed to guarantee sentence- or user-level privacy, prevent PII disclosure. In this work, we introduce rigorous game-based definitions for three types of PII leakage via black-box extraction, inference, and reconstruction attacks with only API access to an LM. We empirically evaluate the attacks against GPT-2 models fine-tuned with and without defenses in three domains: case law, health care, and e-mails. Our main contributions are (i) novel attacks that can extract up to $10\times$ more PII sequences than existing attacks, (ii) showing that sentence-level differential privacy reduces the risk of PII disclosure but still leaks about 3% of PII sequences, and (iii) a subtle connection between record-level membership inference and PII reconstruction. Code to reproduce all experiments in the paper is available at <https://github.com/microsoft/analysing-pii-leakage>.

I. INTRODUCTION

Language Models (LMs) are fundamental to many natural language processing tasks [22, 49]. State-of-the-art LMs scale to trillions of parameters [49] and are pre-trained on large text corpora (e.g., 700GB [53]). Pre-trained LMs are adapted to downstream tasks by fine-tuning on domain-specific datasets such as human dialogs [7] or clinical health data [62] which may contain private information.

Memorization is a privacy concern in LMs [9]. The threat is that an attacker learns *by whom* the training data was provided, known as membership inference [30, 43, 46, 58] and *about whom* it contains information, known as data extraction [9, 11, 29, 59, 69]. These two categories can be disjoint but associations in the latter can be used to infer information about the former. For LMs, data extraction is a significant threat in practice since attackers with black-box API access can extract at least 1% of the training data [41].

Existing work focuses on finding a lower bound on *any* kind of memorization but does not differentiate public and private

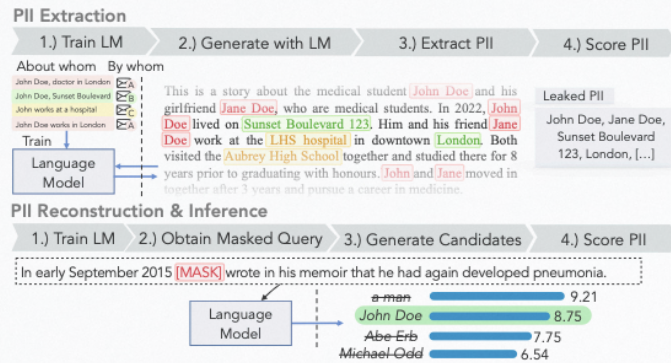


Fig. 1: An illustration of PII extraction, reconstruction and inference attack techniques.

leaked information. For example, leaking highly duplicated common phrases is not a privacy violation according to the GDPR [47] as opposed to leaking Personally Identifiable Information (PII). In practice, any LM trained on real, sensitive data has to protect PII, but memorization of PII is not well understood. We believe that a comprehensive study on the risk of PII memorization in LMs is missing.

Consider a service provider who wants to deploy a next-word prediction LM for composing e-mails, such as Google’s Smart Compose [43]. Their goal is to train an LM with high utility that does not leak PII and make it available as a black-box API. The threat is an attacker who learns PII, such as names, addresses or other sensitive information, through the LM. Extracting *any* PII by itself, such as a personal address, can already pose a privacy threat. This threat is elevated when an attacker can associate a piece of PII to a context, for example, “In May 2022, [MASK] had chemotherapy at LHS”. As a part of this paper, we study the feasibility of such attacks on LMs in practice. Figure 1 illustrates the type of PII attacks proposed in this work.

Defenses against memorization are based on dataset curation and algorithmic defenses. PII *scrubbing* is a dataset curation technique that removes PII from text, relying on Named Entity Recognition (NER) [33] to tag PII. Modern NER is based on the Transformer architecture [63] and has mixed recall of 97% (for names) and 80% (for care unit numbers) on clinical health data, meaning that much PII is retained after scrubbing [44]. Modern training pipelines incorporate algorithmic defenses such as differential privacy

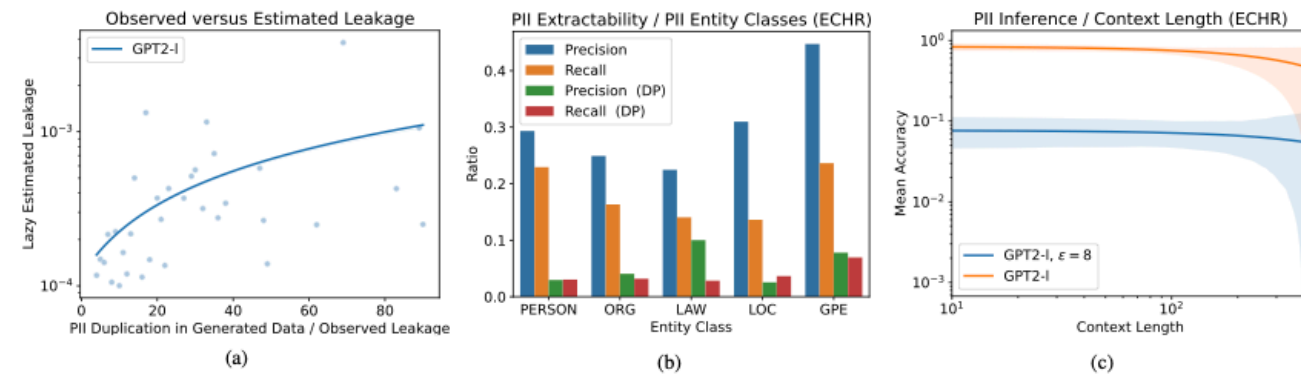


Fig. 6: Fig. 6a shows the correlation between the observed and estimated leakage. Fig. 6b shows the precision and recall for other entity classes on the ECHR dataset. Fig. 6c shows the mean inference accuracy relative to the context length, which is the length of the combined prefix and suffix for a masked query.

	GPT2-Small		GPT2-Medium		GPT2-Large		GPT2-XL	
	No DP	$\epsilon=8$	No DP	$\epsilon=8$	No DP	$\epsilon=8$	No DP	$\epsilon=8$
ECHR (TAB)	0.78%	0.24%	1.21%	0.32%	5.81%	0.48%	4.30%	0.39%
ECHR (Ours, $ C =64$)	2.25%	0.44%	3.36%	0.87%	18.27%	0.55%	13.11%	0.41%
Enron (TAB)	0.59%	0.04%	0.67%	0.04%	1.75%	0.04%	2.19%	0.19%
Enron (Ours, $ C =64$)	6.29%	0.49%	7.26%	0.52%	12.68%	0.55%	15.25%	0.53%
Yelp-Health (TAB)	0.33%	0.24%	0.37%	0.14%	0.65%	0.12%	1.99%	0.12%
Yelp-Health (Ours, $ C =64$)	0.42%	0.32%	1.31%	0.32%	1.69%	0.35%	6.40%	0.36%

TABLE IV: Results of PII reconstruction attacks on the entity class “person”. Bold numbers represent the best attack per dataset and LM. We compare our results with the TAB attack [28] on three datasets.

numbers and 16.3% of mentioned organization names from an undefended LM. However, we believe that our attack can observe the correct PII candidates for 9% of law cases and 4.1% of organizations. For the Enron dataset, which contains long phone numbers, we never observe a single leaked real phone number in the DP model. However, we observe leakage of e-mail addresses (consisting of equally many tokens), that are typically correlated with a person’s name.

F. PII Reconstruction

We measure our PII reconstruction attack on GPT2-L with the TAB attack [28]. Table IV shows the results on ECHR, Enron, and Yelp-Health for the entity class “person”. We sample 64 candidates and decode from the model using top- k sampling with $k=40$. We observe that our reconstruction attack significantly outperforms the TAB attack on undefended models enabling the reconstruction of up to $10\times$ more PII (in the GPT2-Medium case on Enron). Model Size. On ECHR and Enron, GPT2-XL currently reconstructs a least 2.5 \times more PII sequences than GPT2-Small. This observation demonstrates that information in a sample’s suffix provides a strong signal to reconstruct PII. On ECHR, our attack improves the baseline by at least $2.5\times$, on Enron we observe an improvement of at least $7.5\times$ and on Yelp-Health our attack is at least about $3\times$ more successful (except for GPT2-Small where our attack improves only from 0.33% to 0.42%).

In LM models (≤ 100 B), our attack still improves the baseline in all cases, which we believe that means that our attack is a useful tool for PII reconstruction. We observe that GPT2-XL and GPT2-Large models are more vulnerable to PII reconstruction.

Context Size. On Enron, the advantage of our attack compared to TAB becomes more evident. E-mails in the Enron dataset typically mention the receiver of the e-mail at the beginning prior to any PII. For this reason, the TAB attack has only a small prefix to predict PII and cannot leverage the information contained in the e-mail body. We observe that our attack is successful for 70% of the time. However, our reconstruction attack often does not sample the correct candidate which effectively limits our attack’s success rate. We believe a method that samples candidates by incorporating information from the sample’s suffix could improve our attack even further.

Figure 7 shows our PII reconstruction attack on ECHR. Figure 7a shows the ROC curve for GPT2-L with $\epsilon=8$ and GPT2-L. The values represent the attack accuracy at inferring the correct PII out of $|C|$ candidates.

	ECHR		Enron		Yelp-Health	
	No DP	$\epsilon=8$	No DP	$\epsilon=8$	No DP	$\epsilon=8$
$ C =100$	70.11%	8.32%	50.50%	3.98%	28.31%	4.29%
$ C =500$	51.03%	3.71%	34.14%	1.92%	15.55%	1.86%

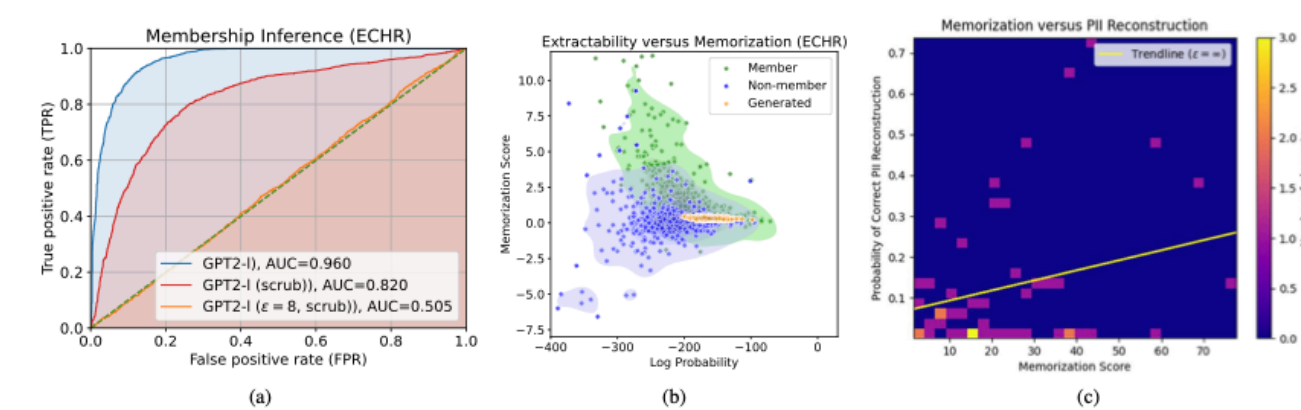


Fig. 7: Connecting sentence-level membership inference with PII reconstruction in GPT-2-Large. 7a shows the ROC curve against our fine-tuned model using a shadow model attack on ECHR. 7b shows that the memorization score of generated sequences is nearly zero and 7c shows that the memorization score correlates with the probability of correct PII reconstruction.

	Undefended	DP	Scrub	DP + Scrub
Test Perplexity	14 / 9	14	16	16
Extract Precision	30%	3%	0%	0%
Extract Recall	23%	3%	0%	0%
Reconstruction Acc.	18%	1%	0%	0%
Inference Acc. ($ C =100$)	70%	8%	1%	1%
MI AUC	0.96	0.5	0.82	0.5

TABLE VI: Our results on ECHR for GPT-2-Large summarize the privacy-utility trade-off. We show the undefended model’s perplexity with/without masking generated PII. The undefended model has the lowest perplexity but the highest leakage. DP with $\epsilon=8$ mitigates MI and (partially) PII leakage. Scrubbing only prevents PII leakage. DP with scrubbing mitigates all the privacy attacks but suffers from utility degradation.

- DP does not completely eliminate leakage from PII inference and PII extraction. We demonstrate that an attacker can infer PII with up to 10% accuracy (given 100 candidates) in a practical setting.
- We find that DP and (aggressive) PII scrubbing limit the LM’s utility, motivating the search for defenses with better empirical privacy/utility trade-offs.

V. DISCUSSION AND LIMITATIONS

Below, we discuss extensions and limitations of our methodology, and identify further research motivated by our findings. We first discuss the applicability of our methodology to sensitive information beyond PII and potential extensions to our attack. Then, we discuss the applicability and associations in the training dataset. We then describe how masked language models fare compared to autoregressive models and identify further research motivated by our findings: how to best combine DP training and scrubbing, optimizing attacks for other leakage metrics, and the need for better benchmarks.

General Applicability. In this paper, we focus on defining metrics, game-based definitions, and tractable formulas for

evaluating leakage of sensitive sequences of tokens categorized as PII. That said, we bring attention to the point that our methodology is generally applicable to any notion of sensitive input. As long as one has an effective method to correctly identify inputs deemed sensitive, our methodology can be adapted to measure the protection offered by existing ML pipelines in mitigating the leakage of *any* sensitive information. In practice, it is often hard to draw a clear boundary around what constitutes sensitive information, which is an important but orthogonal problem.

Syntactic and Semantic Similarity. We consider verbatim matches of PII tokens as leakage, however, our methods can be adapted to account for both syntactic and semantic similarity. For example, “Mr. John Doe” and “J. Doe” could be inferred to be the same person. Similarly, PII reconstruction and PII inference attacks can employ contexts with similar meaning to improve attack results.

Advanced Attacks. We consider leakage of PII sequences from the training dataset in isolation, irrespective of the context where it appears and other extracted PII. Extracted PII sequences can be further leveraged in advanced attacks that explore associations among them and reveal additional private information about the training dataset, thereby enabling linkability attacks.

Utility-preserving Scrubbing. Our empirical evaluation demonstrates that differential privacy is partially effective in mitigating leakage of PII. Based on this observation, existing scrubbing techniques can be adapted to take into consideration the partial protection offered by DP and heuristically scrub only PII that remains unprotected (e.g. because it occurs many times). Such a DP-informed scrubbing would allow for improving model utility while maintaining a privacy level equivalent to a naive combination of DP training and scrubbing.

Comparison to Masked Language Models. Prior work has explored PII reconstruction in the clinical health setting [37, 61] with masked language models (MLMs) based on the BERT architecture [14]. MLMs are trained to reconstruct


[§]Part of this work was done during an internship at Microsoft Research.
[†]To cite this work, please refer to the full publication [41] in IEEE Security and Privacy (S&P) 2023.

Connection between Membership Inference and PII Reconstruction

Analyzing Leakage of Personally Identifiable Information in Language Models

Source code: https://github.com/microsoft/analysing_pii_leakage



Nils Lukas 



Ahmed Salem 



Robert Sim 



Shruti Tople 



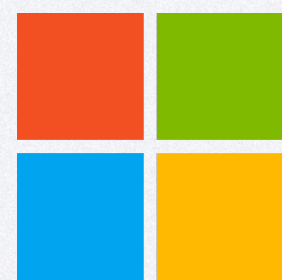
Lukas Wutschitz 



Santiago
Zanella-Béguelin 



UNIVERSITY OF
WATERLOO



Microsoft



GitHub - Source Code



Full Paper

Sources

- [1] <https://www.bloomberg.com/news/articles/2023-02-24/citigroup-goldman-sachs-join-chatgpt-crackdown-fn-reports>, accessed June 14th
- [2] <https://www.businessinsider.in/retail/news/leaked-walmart-memo-warns-employees-not-to-share-any-information-about-walmarts-business-with-chatgpt-or-other-ai-bots/articleshow/98315181.cms>, accessed June 14th
- [3] <https://www.bbc.com/news/technology-65139406>, accessed June 14th
- [4] <https://www.bbc.com/news/technology-65431914>, accessed June 14th
- [5] <https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak>, accessed June 14th
- [6] <https://help.openai.com/en/articles/6783457-what-is-chatgpt>, accessed June 14th
- [7] <https://bard.google.com/faq?hl=en>, accessed June 14th
- [8] OpenAI, "GPT-4 Technical Report", arXiv preprint arXiv:2303.08774 (2023)
- [9] <https://www.techspot.com/news/56127-10000-aws-secret-access-keys-carelessly-left-in-code-uploaded-to-github.html>, accessed June 14th
- [10] <https://analyticsdrift.com/github-copilot-ai-is-leaking-functional-api-keys/>, accessed June 14th
- [11] <https://www.bleepingcomputer.com/news/security/github-copilot-update-stops-ai-model-from-revealing-secrets/>, accessed June 14th
- [12] Liu, Haokun, et al. "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning." *Advances in Neural Information Processing Systems* 35 (2022): 1950-1965.

Appendix



<https://nilslukas.github.io>