Analyzing Leakage of Personally Identifiable Information in Language Models






UNIVERSITY OF WATERLOO

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz and Santiago Zanella-Béguelin

Microsoft

Privacy Concerns in ChatBots

Bloomberg
US Edition Sign In
Live Now | Markets | Economics | Industries | Technology | Politics | Wealth | Pursuits | Opinion | Businessweek | Equality | Green | CityLab

Technology
City of London

Wall Street Banks Are Cracking Down on AI-Powered ChatGPT

- Banks currently restricting chatbot's use, people familiar say
- ChatGPT has sparked intense interest across industries

By Gabriela Mello, William Shaw and Hannah Levitt
February 24, 2023, 7:36 AM EST *Updated on February 24, 2023, 10:10 AM EST*

LIVE ON BLOOMBERG
Watch Live TV
Listen to Live Radio

Bloomberg, 2023 [1]

HOME > RETAIL

Leaked Walmart memo warns employees not to share 'any information about Walmart's business' with ChatGPT or other AI bots

Ben Tobin Feb 28, 2023, 11:32 AM EST

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
US Edition Sign In Subscribe
Live Now | Markets | Economics | Industries | Technology | Politics | Wealth | Pursuits | Opinion | Businessweek | Equality | Green | CityLab | Crypto | More

Technology
AI

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

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

By Mark Gurman
May 1, 2023, 8:48 PM EDT *Updated on May 2, 2023, 1:54 AM EDT*

LIVE ON BLOOMBERG
Watch Live TV
Listen to Live Radio

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

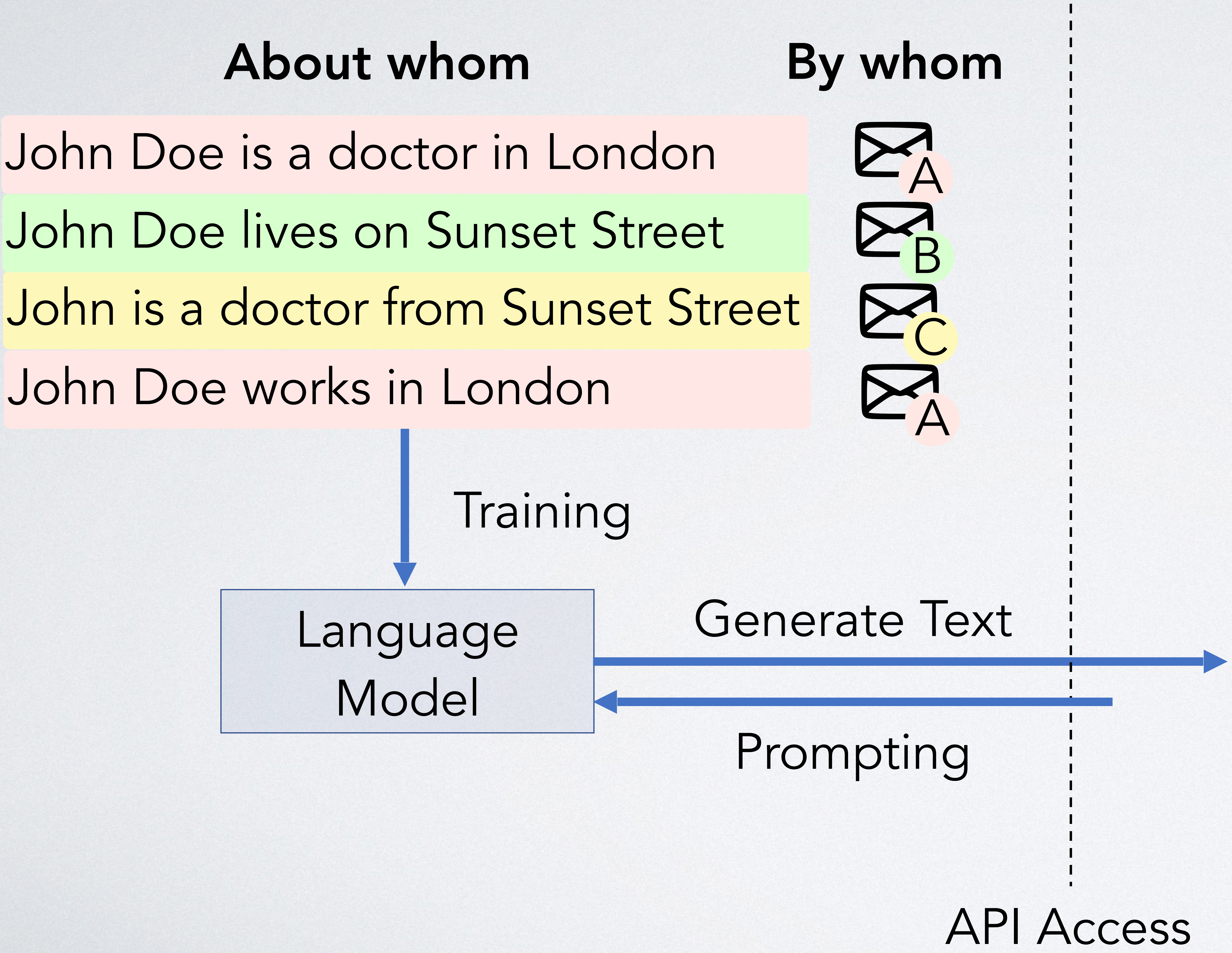
Pre-Training: ~10m USD

Fine-Tuning: ~5-10k USD

PEFT: <1k USD

[12]

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

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

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.

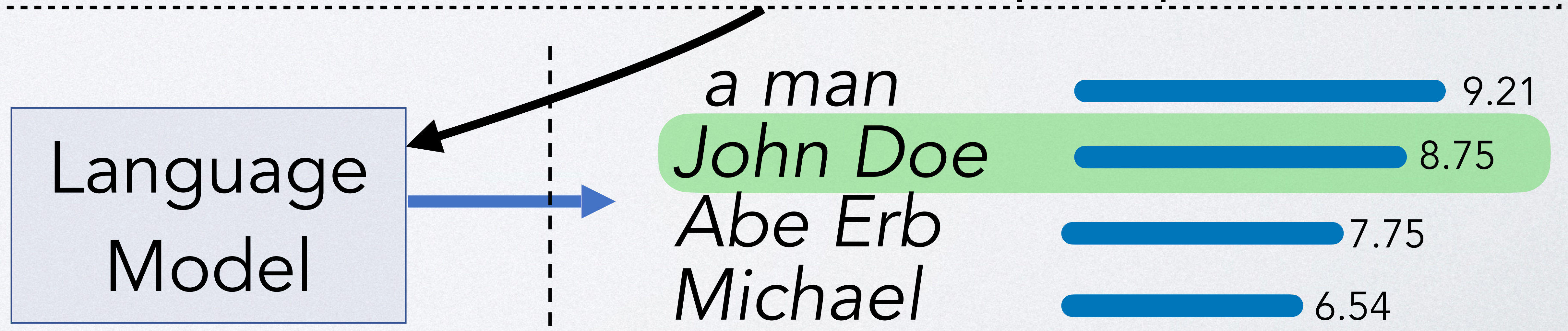
1.) PII Extraction

Real	Fictional
John Doe London Sunset Street LHS Hospital	Jane Doe Aubrey High School

2.) PII Reconstruction & 3.) PII Inference

Real Sentence

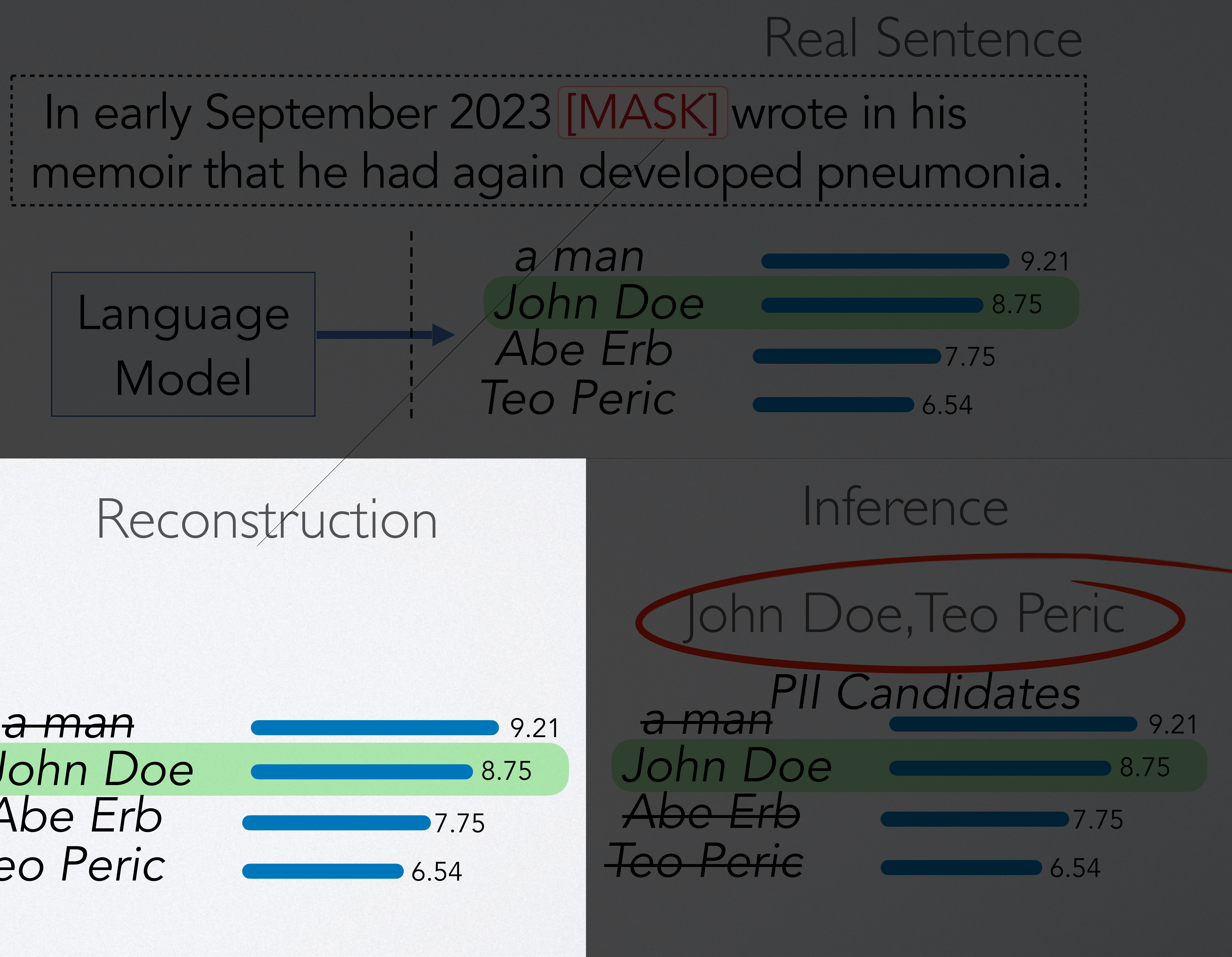
In early September 2023 [MASK] wrote in his memoir that he had again developed pneumonia.



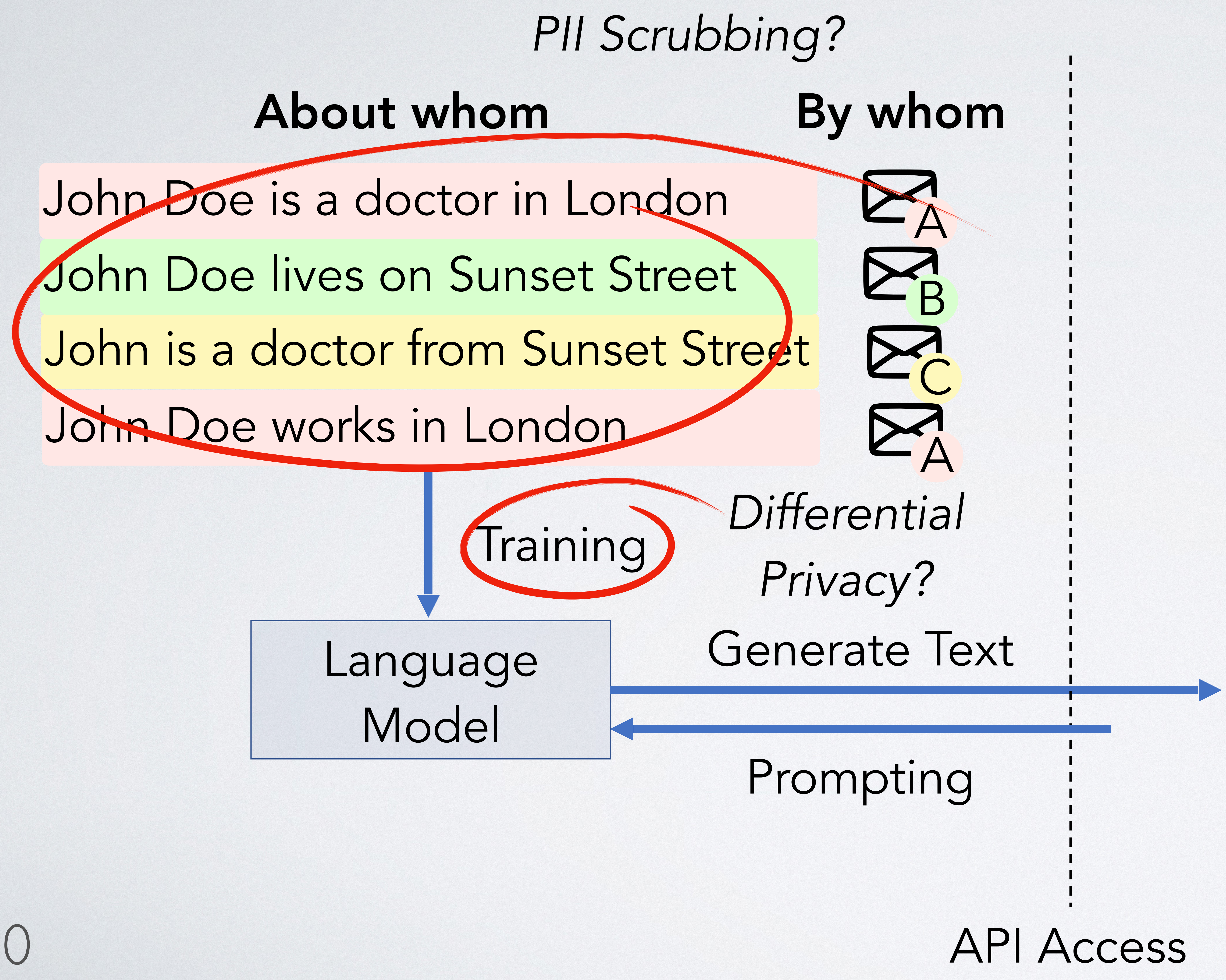
Motivation

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.

2.) PII Reconstruction & Inference

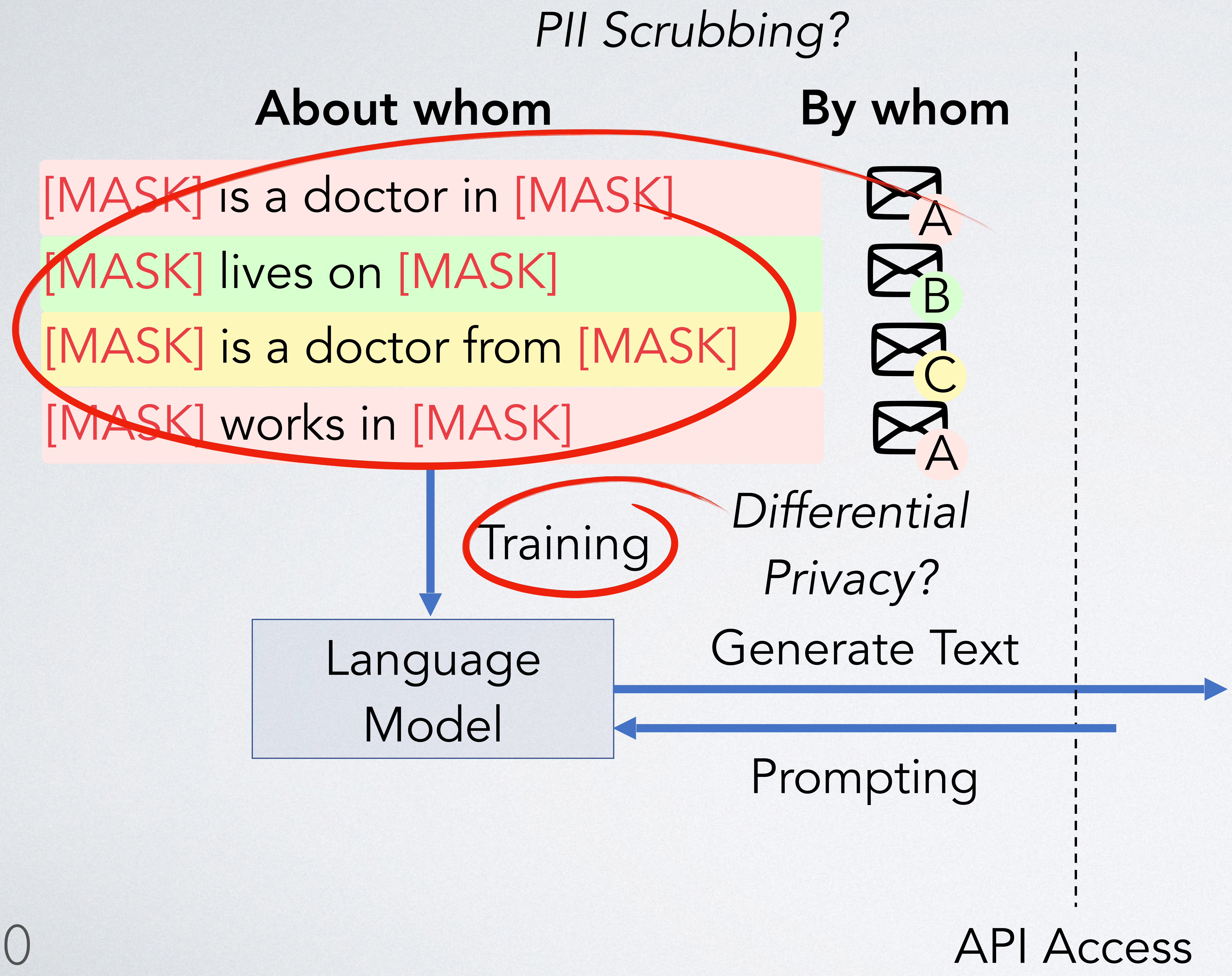


Motivation



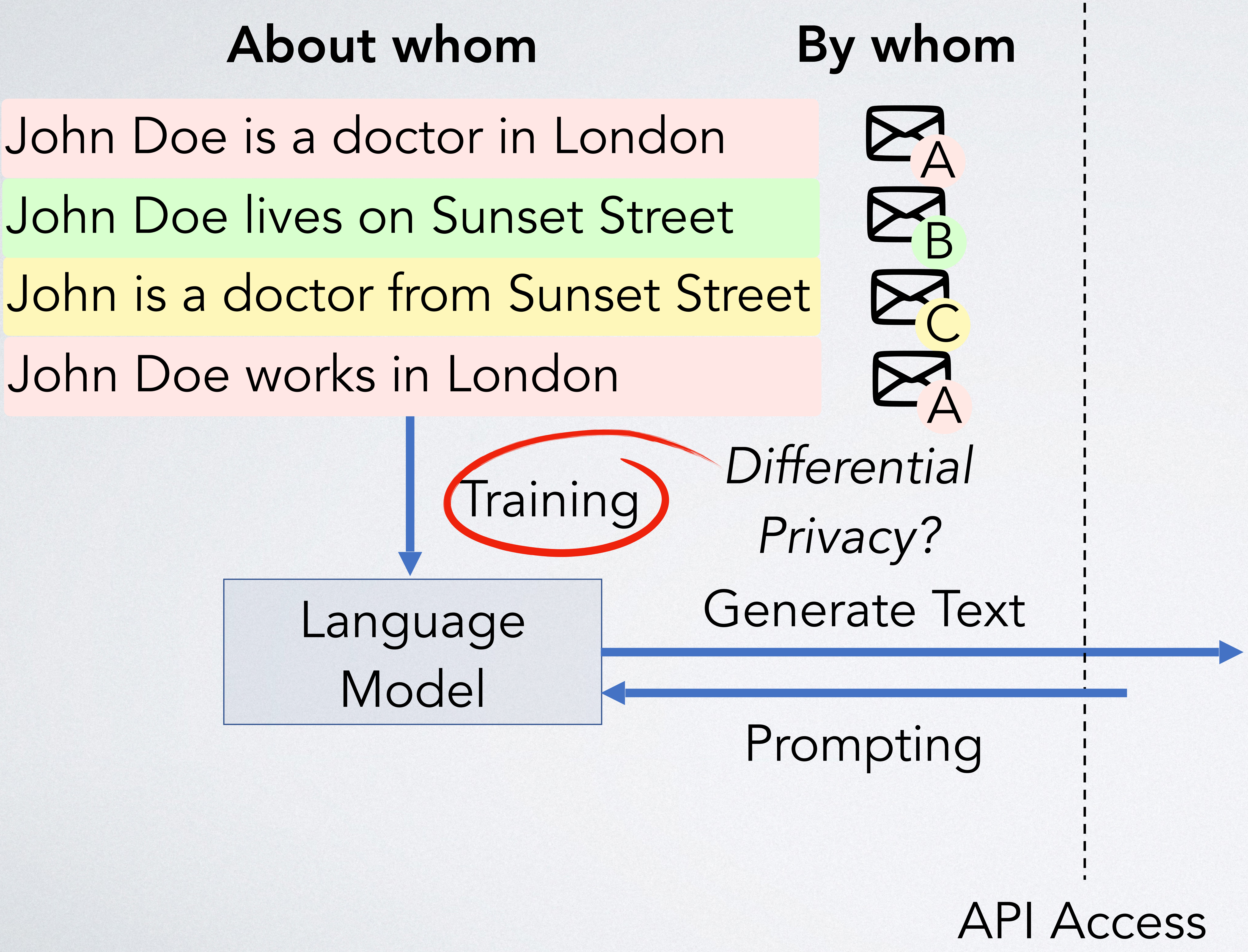
Once upon a time, there existed a tale of medical students John Doe and his girlfriend, Jane Doe. In the year 2022, John resided at Sunset Street while pursuing his medical education. Alongside his friend Jane, he worked at the LHS Hospital located in the bustling heart of downtown London. Before donning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



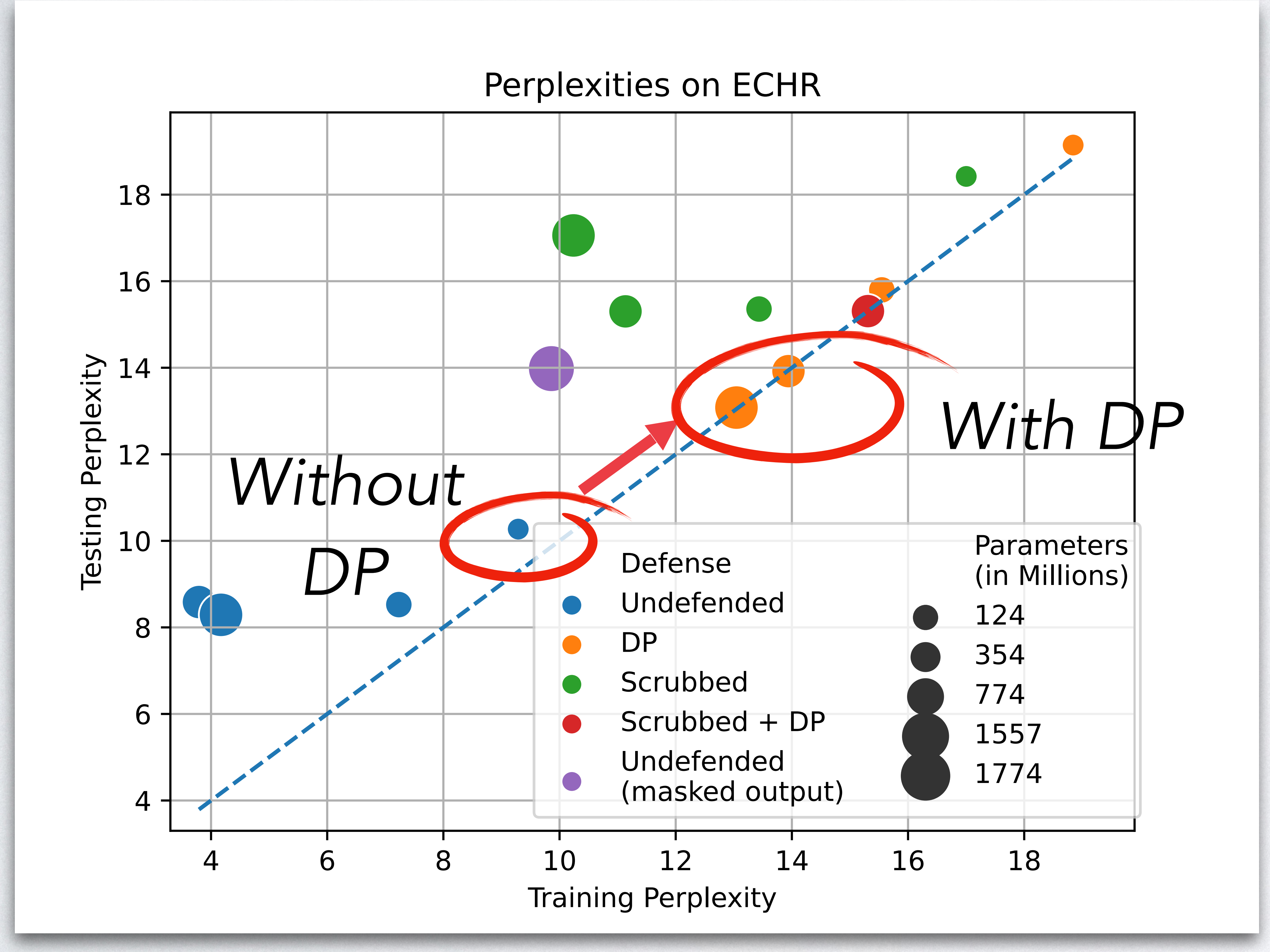
Once upon a time, there existed a tale of medical students John Doe and his girlfriend, Jane Doe. In the year 2022, John resided at Sunset Street while pursuing his medical education. Alongside his friend Jane, he worked at the LHS Hospital located in the bustling heart of downtown London. Before donning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

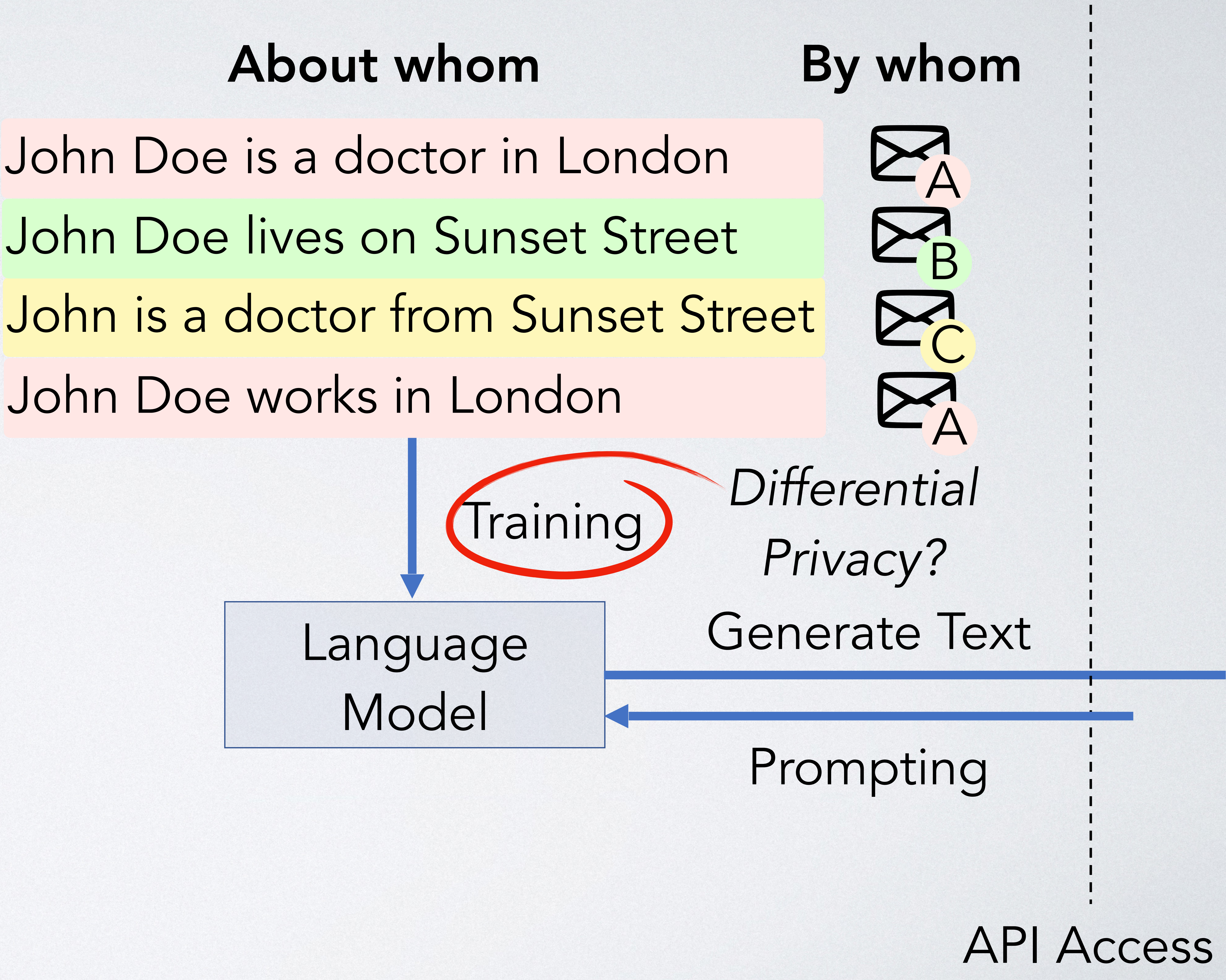


Once upon a time, there existed a tale of medical students John Doe and his girlfriend, Jane Doe. In the year 2022, John resided at Sunset Street while pursuing his medical education. Alongside his friend Jane, he worked at the LHS Hospital located in the bustling heart of downtown London. Before donning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

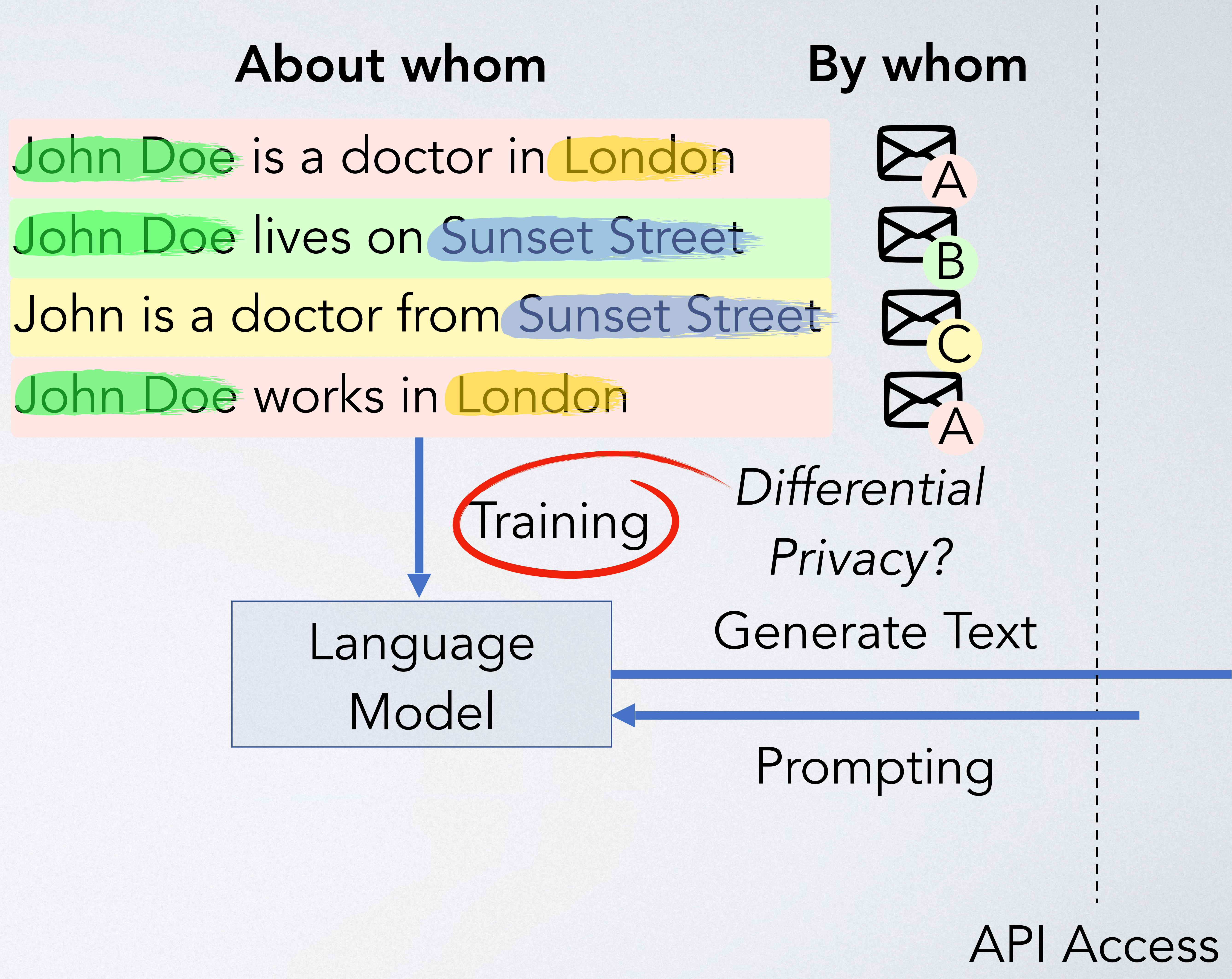


Privacy at the cost of Model Utility



Problems with Differential Privacy

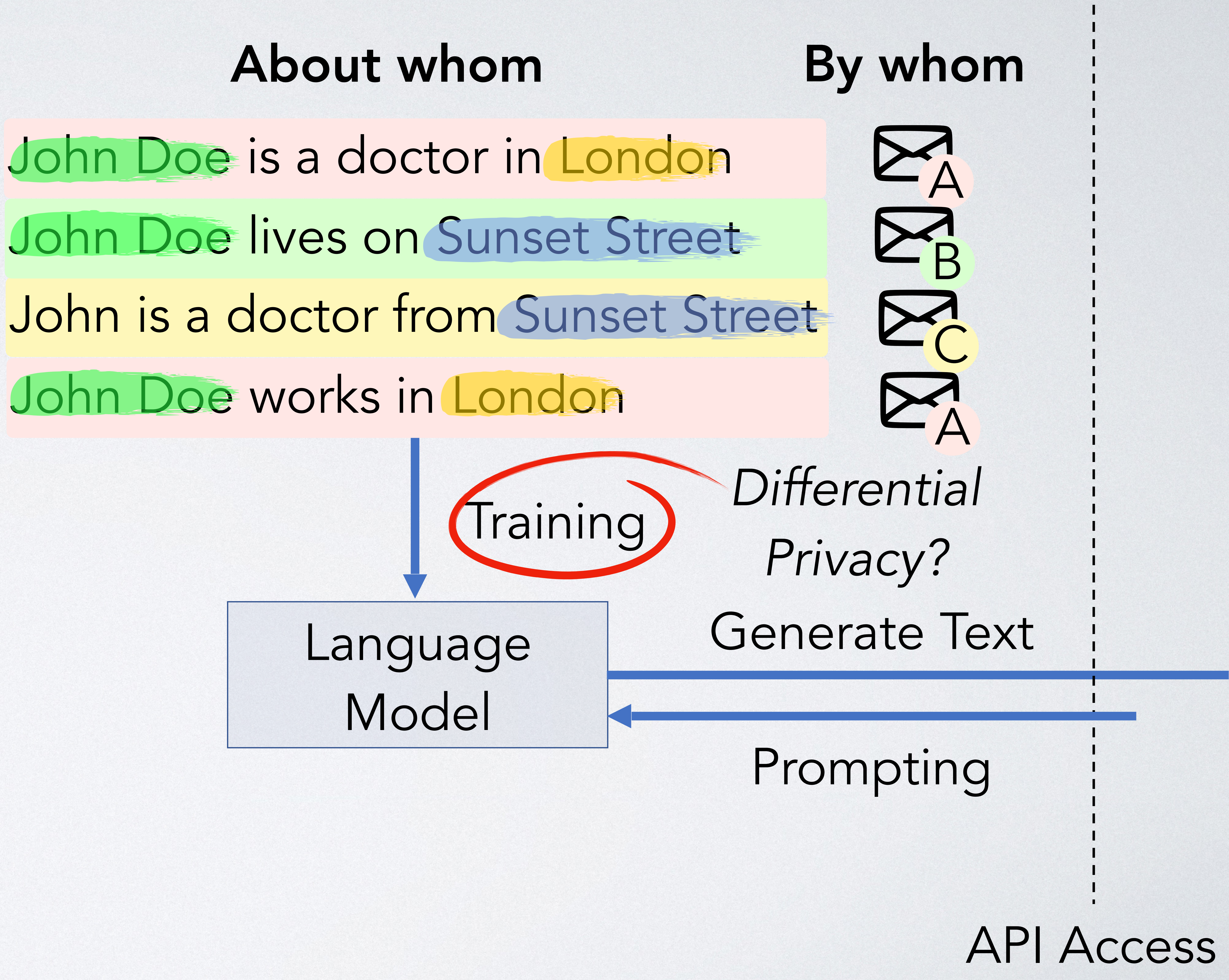
DP protects against an attacker learning **by whom** data was provided, but not **about whom** it contains information.



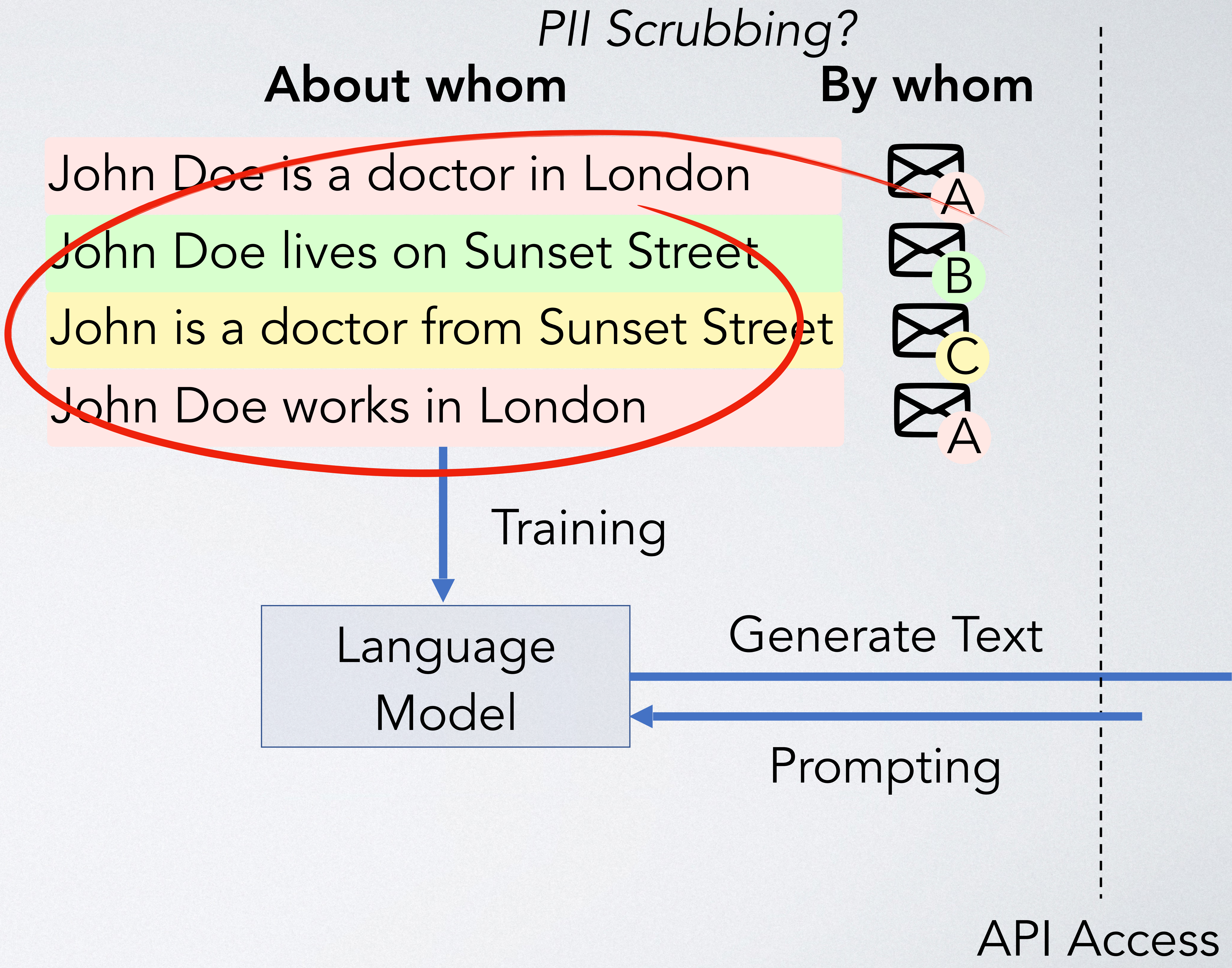
Problems with Differential Privacy

Group-level DP can help but ..

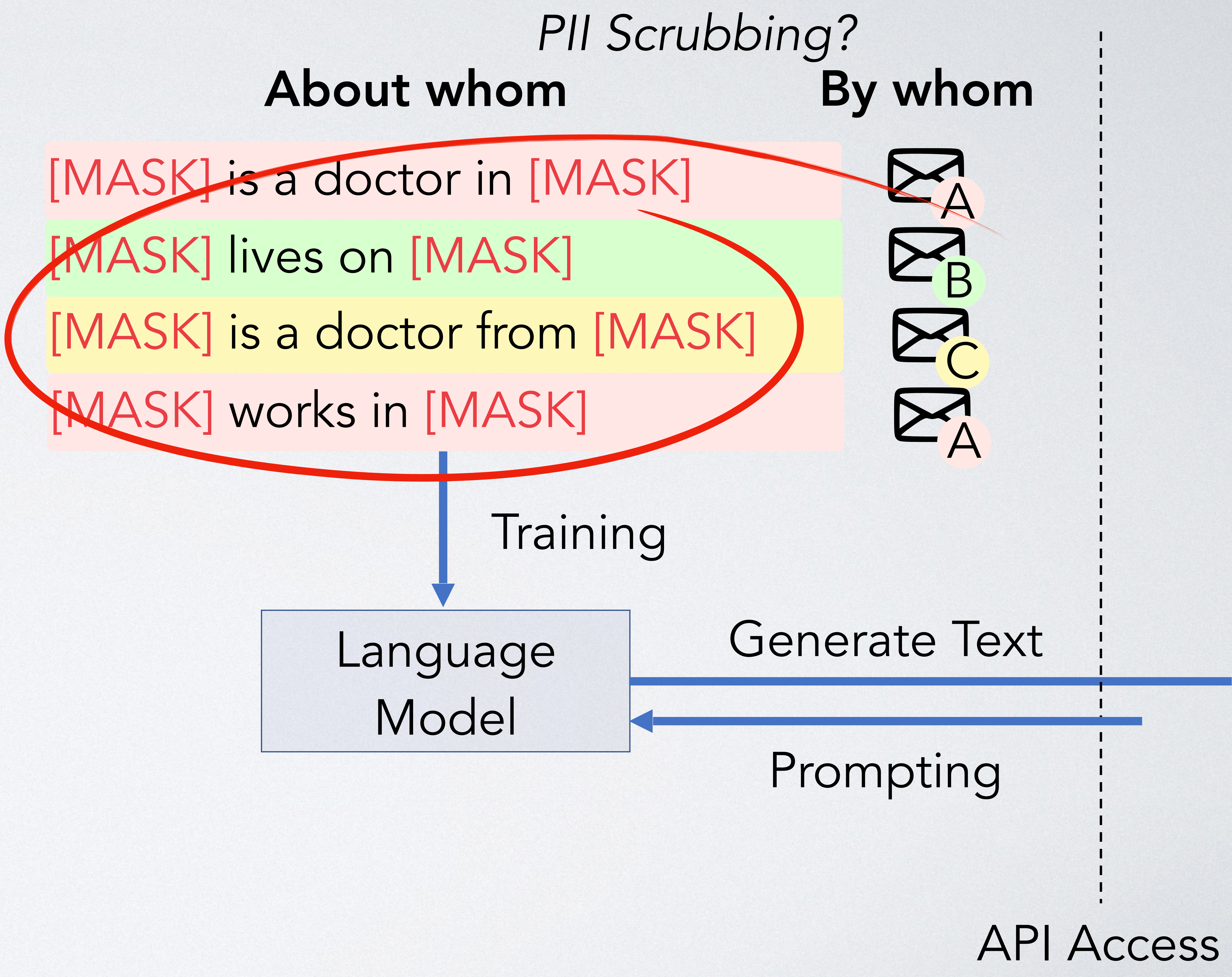
1) Group sizes are not always known a priori and under worst-case assumptions has deleterious impact on model utility.
2) PII Duplication across groups



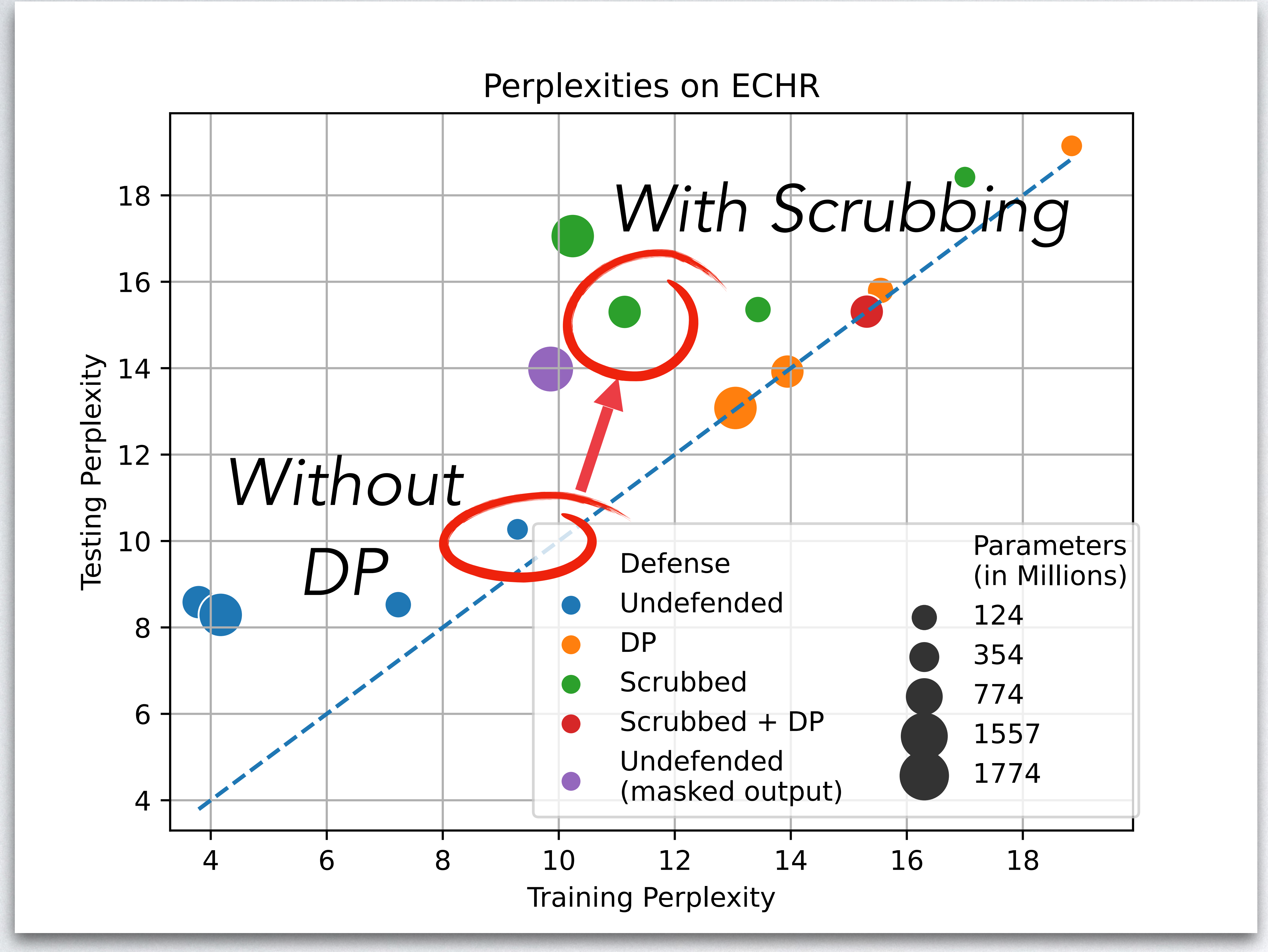
Problems with PII Scrubbing



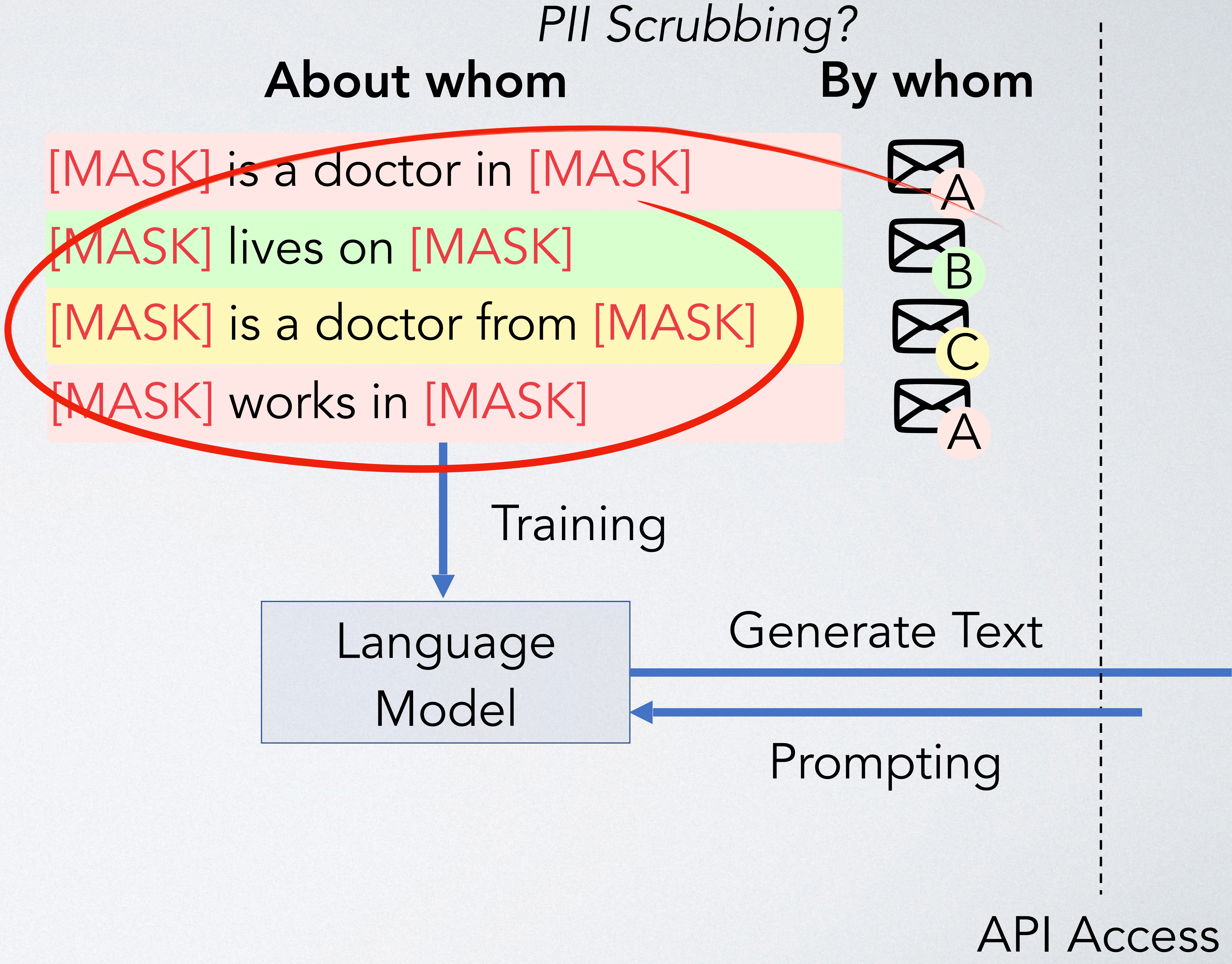
Problems with PII Scrubbing



Problems with PII Scrubbing

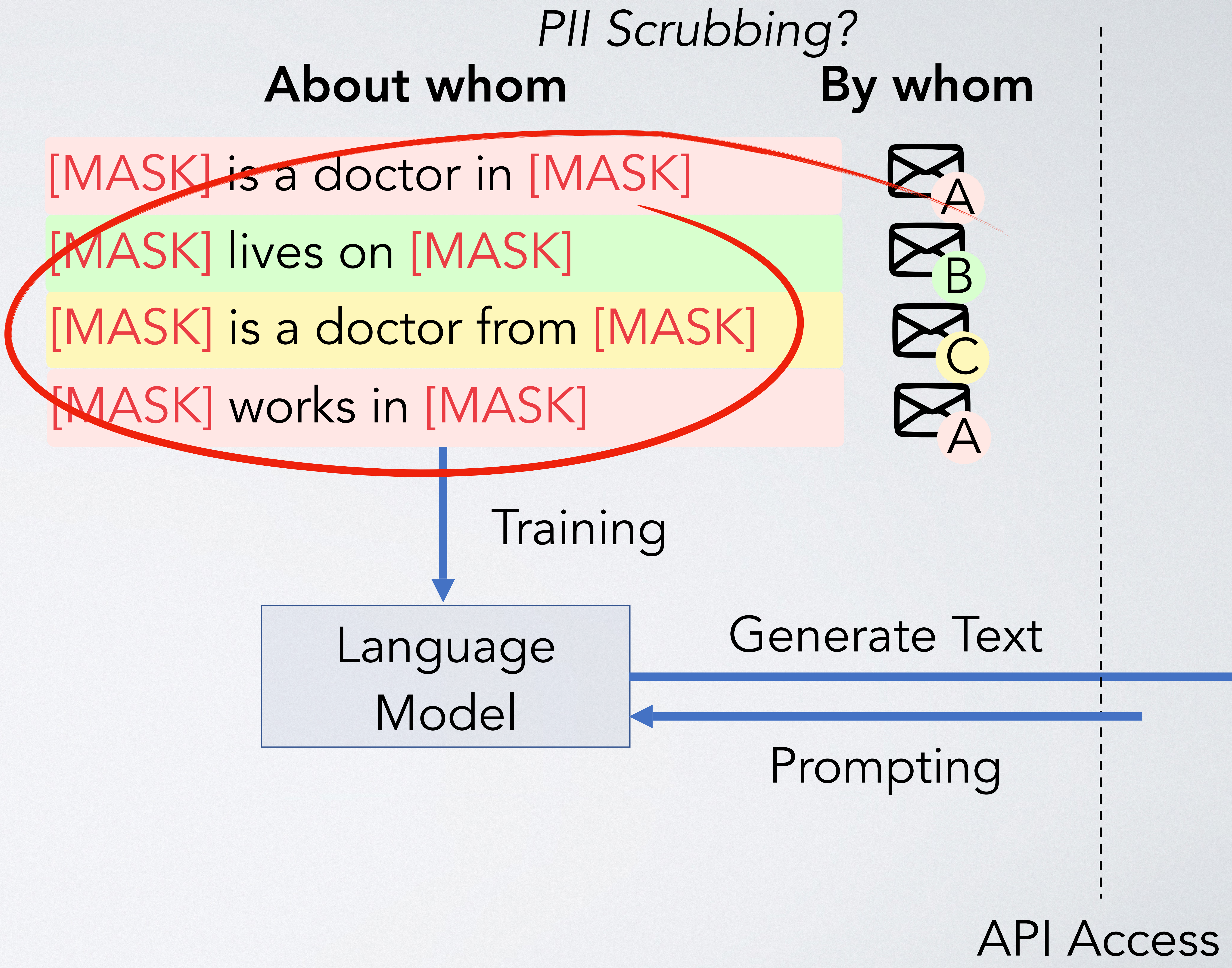


<u>Privacy</u> at the cost of <u>Model Utility</u>



Problems with PII Scrubbing

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



Related Work

Canaries


The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

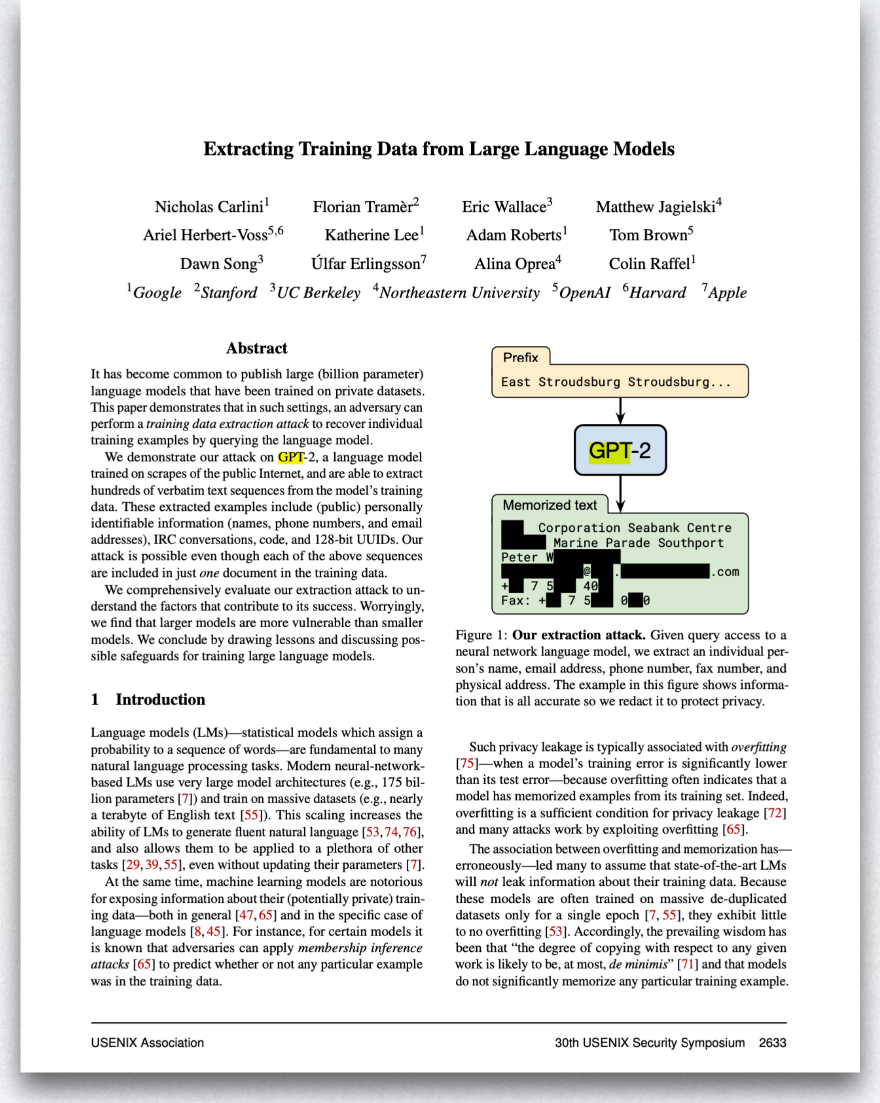
Carlini et al., 2019

N-grams

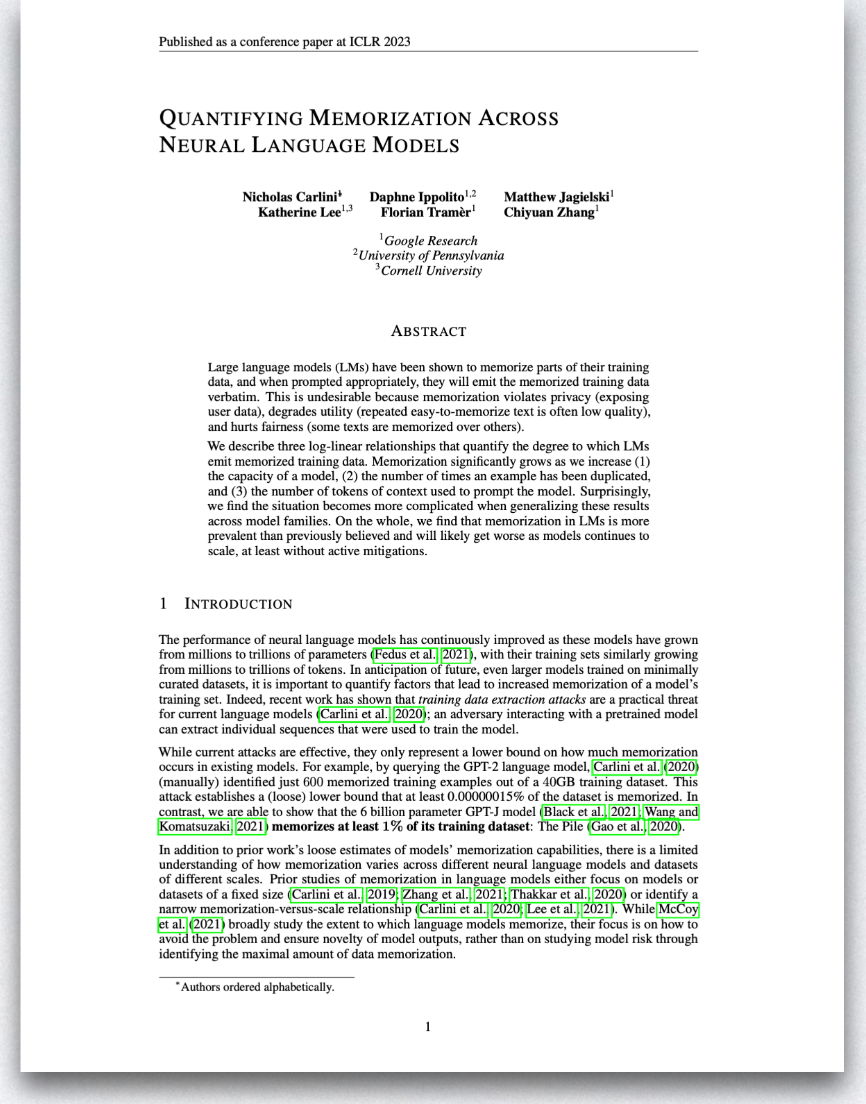

How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN

McCoy et al., 2019

Sequences


Extracting Training Data from Large Language Models

Carlini et al., 2020


Quantifying Memorization Across Neural Language Models

Carlini et al., 2022

PII Leakage
In Pre-Trained LMs


Are Large Pre-Trained Language Models Leaking Your Personal Information?

Huang et al., 2022

Related Work

Privacy in LMs

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

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

Hannah Brown[1], **Katherine Lee**[2], **Fatemehsadat Mireshghallah**[3]
Reza Shokri[1], **Florian Tramèr**[4*]
[1]National University of Singapore, [2]Cornell University
[3]University of California San Diego, [4]Google
{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

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")

Security Games for PII Leakage

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

> **experiment** $\textsc{Ind-Inference}(\mathcal{T}, \mathcal{D}, n, \mathcal{A})$
> $\quad b \sim \{0, 1\}$
> $\quad D \sim \mathcal{D}^n$
> $\quad \theta \leftarrow \mathcal{T}(D)$
> $\quad S_0 \sim D$
> $\quad S_1 \sim \mathcal{D}$
> $\quad \tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), S_b)$
> $\quad S \sim \{S \in D | \textsc{Extract}(S) \neq \emptyset\}$
> $\quad C_0 \sim \textsc{Extract}(S)$
> $\quad C_1 \sim \mathcal{E}$
> $\quad \tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), \textsc{Scrub}(\textsc{Split}(S, C_0)), C_b)$

Algorithm 5 PII Reconstruction Game

> **experiment** $\textsc{Reconstruction}(\mathcal{T}, \mathcal{D}, n, \mathcal{A})$
> $\quad D \sim \mathcal{D}^n$
> $\quad \theta \leftarrow \mathcal{T}(D)$
> $\quad S \sim \{S \in D | \textsc{Extract}(S) \neq \emptyset\}$
> $\quad C \sim \textsc{Extract}(S)$
> $\quad \tilde{C} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), \textsc{Scrub}(\textsc{Split}(S, C)))$

Algorithm 2 PII Extraction

> **experiment** $\textsc{Extraction}(\mathcal{T}, \mathcal{D}, n, \mathcal{A})$
> $\quad D \sim \mathcal{D}^n$
> $\quad \theta \leftarrow \mathcal{T}(D)$
> $\quad \mathcal{C} \leftarrow \bigcup_{S \in D} \textsc{Extract}(S)$
> $\quad \tilde{\mathcal{C}} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), |\mathcal{C}|)$
>
> **procedure** $\mathcal{O}_\theta(S)$
> $\quad$ **return** $\{w \mapsto \Pr(w|S; \theta)\}_{w \in \mathcal{V}}$

Algorithm 7 PII Inference Game

> **experiment** $\textsc{Inference}(\mathcal{T}, \mathcal{D}, n, m, \mathcal{A})$
> $\quad D \sim \mathcal{D}^n$
> $\quad \theta \leftarrow \mathcal{T}(D)$
> $\quad S \sim \{S \in D | \textsc{Extract}(S) \neq \emptyset\}$
> $\quad C \sim \textsc{Extract}(S)$
> $\quad \mathcal{C} \sim \mathcal{E}^m$
> $\quad \mathcal{C} \leftarrow \mathcal{C} \cup \{C\}$
> $\quad \tilde{C} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), \textsc{Scrub}(\textsc{Split}(S, C)), \mathcal{C})$

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

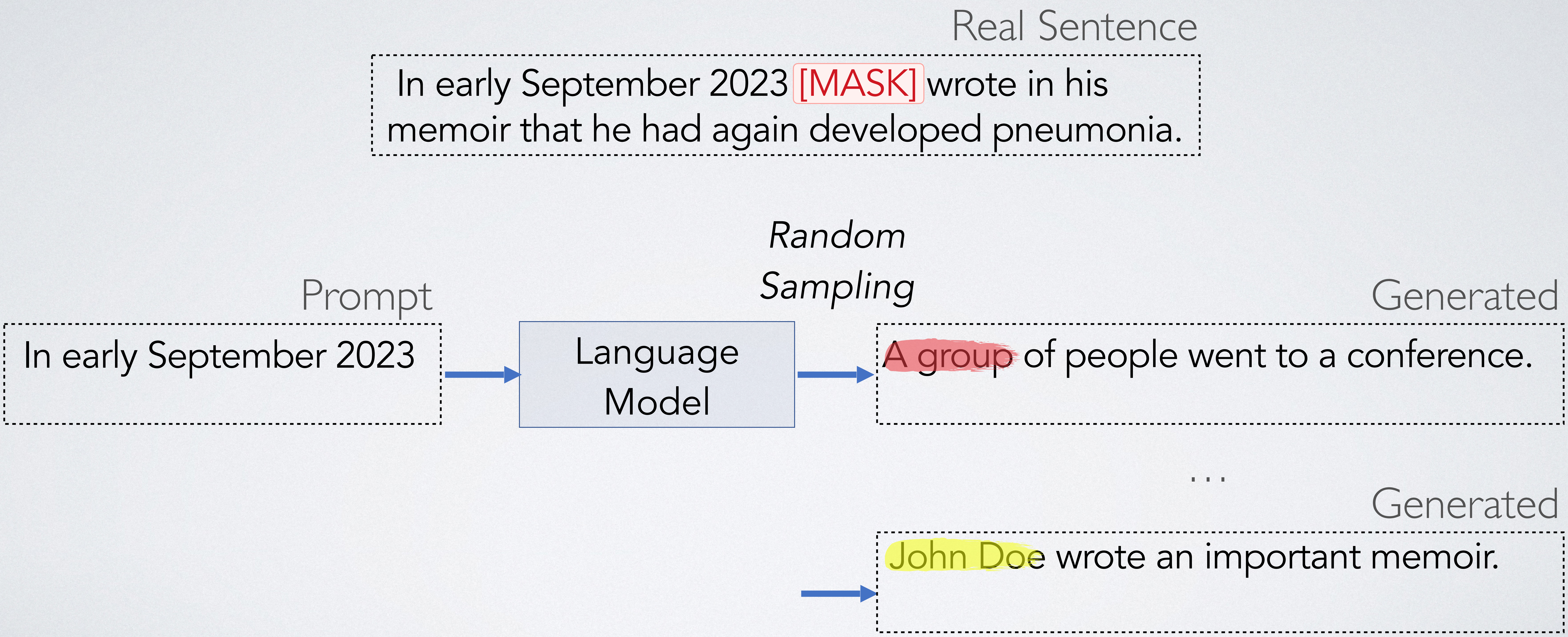
In early September 2023

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

PII Reconstruction Tractability

Prefix

Unknown #Tokens

Real Sentence

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

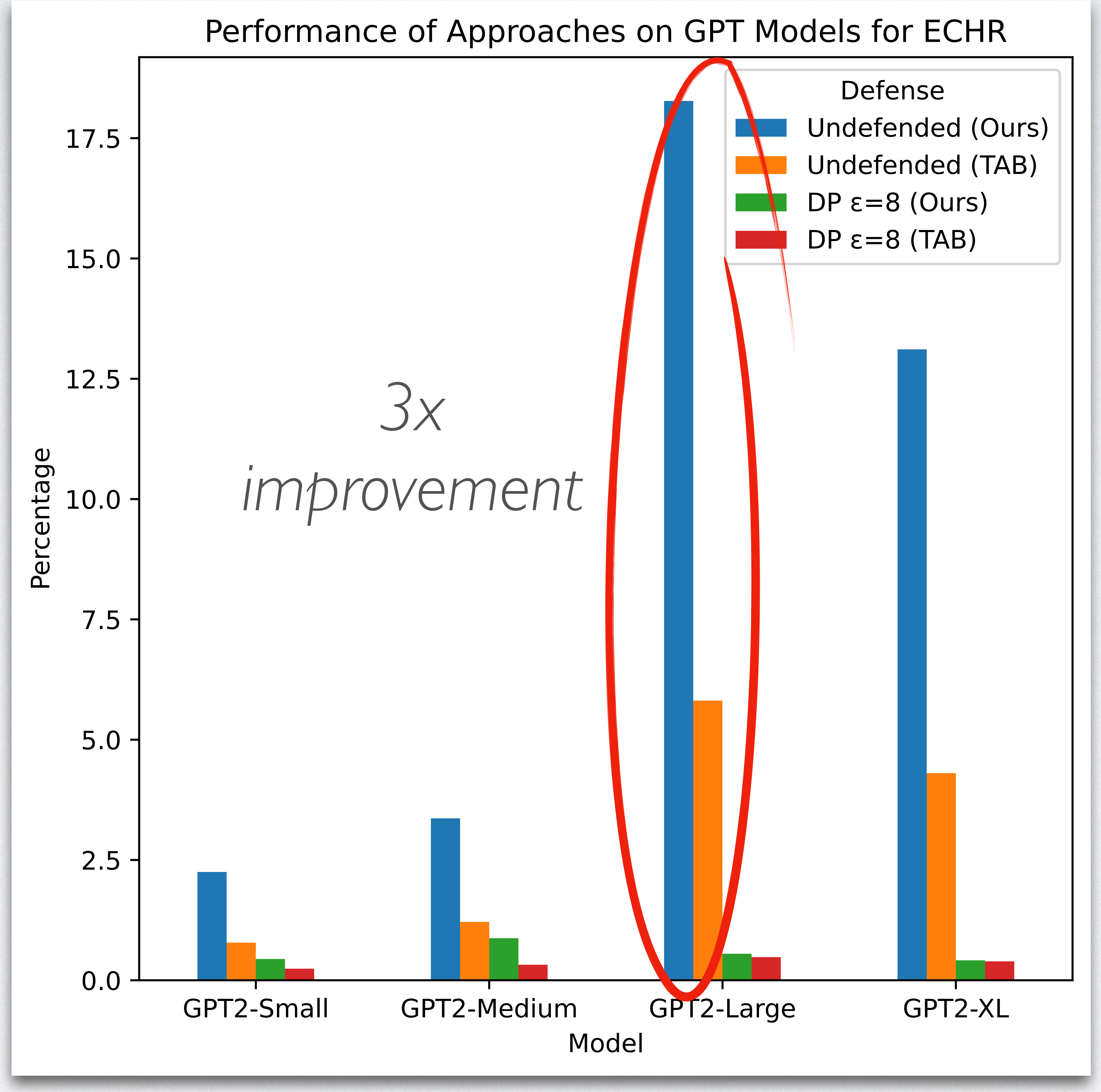
Suffix

Datasets

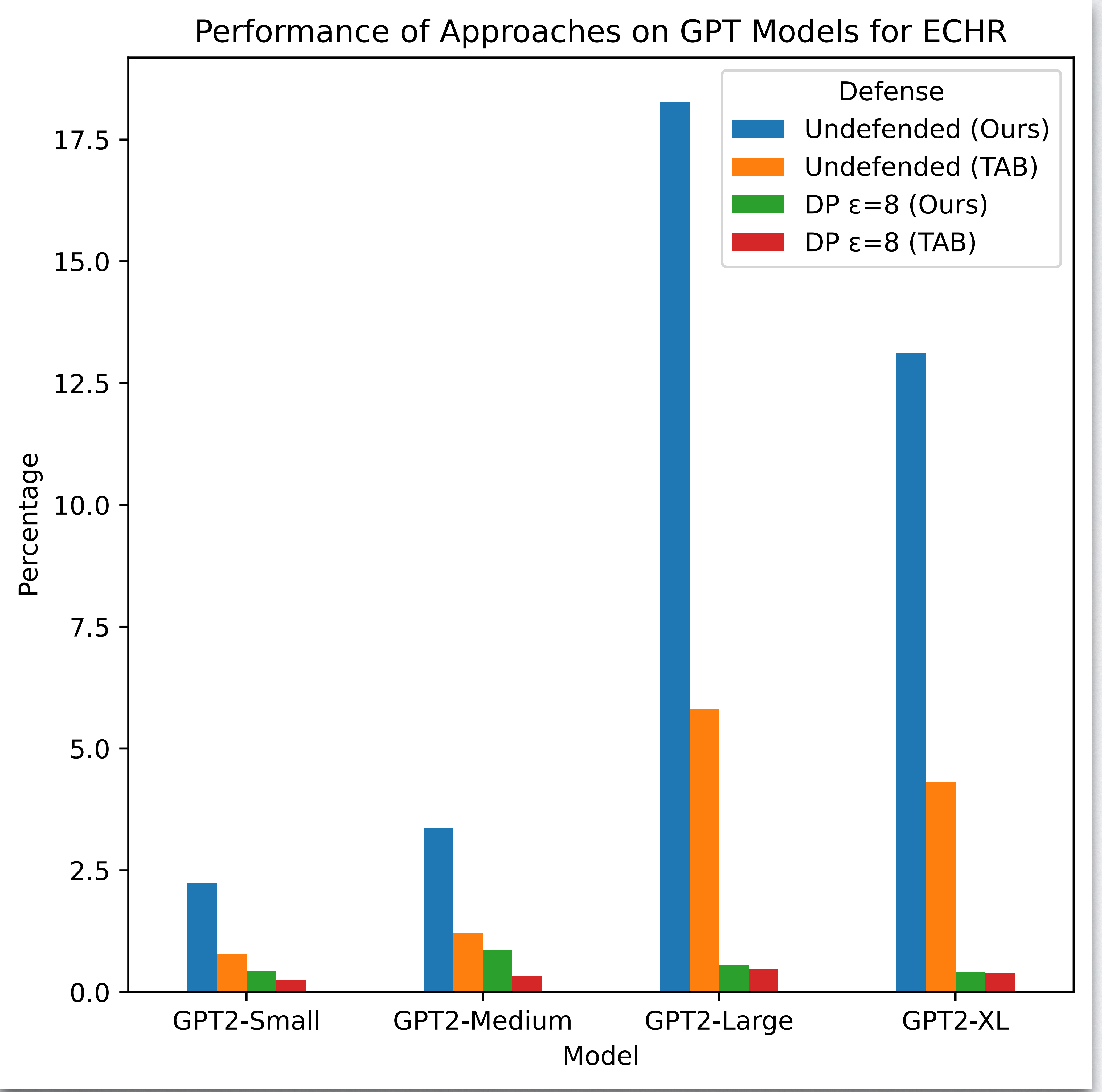
	Records	Tokens / Record	Unique PII	Records w. PII	Duplicates / PII	Tokens / PII
ECHR	118 161	88.12	16 133	23.75%	4.66	4.00
Enron	138 919	346.10	105 880	81.45%	11.68	3.00
Yelp-Health	78 794	143.92	17 035	54.55%	5.35	2.17

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	$\varepsilon = 8$	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$
ECHR(TAB)	0.78%	0.24%	1.21%	0.32%	5.81%	0.48%	4.30%	0.39%
ECHR (Ours, $\lvert\mathcal{C}\rvert = 64$)	**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, $\lvert\mathcal{C}\rvert = 64$)	**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, $\lvert\mathcal{C}\rvert = 64$)	**0.42%**	0.32%	**1.31%**	0.32%	**1.69%**	0.35%	**6.40%**	0.36%

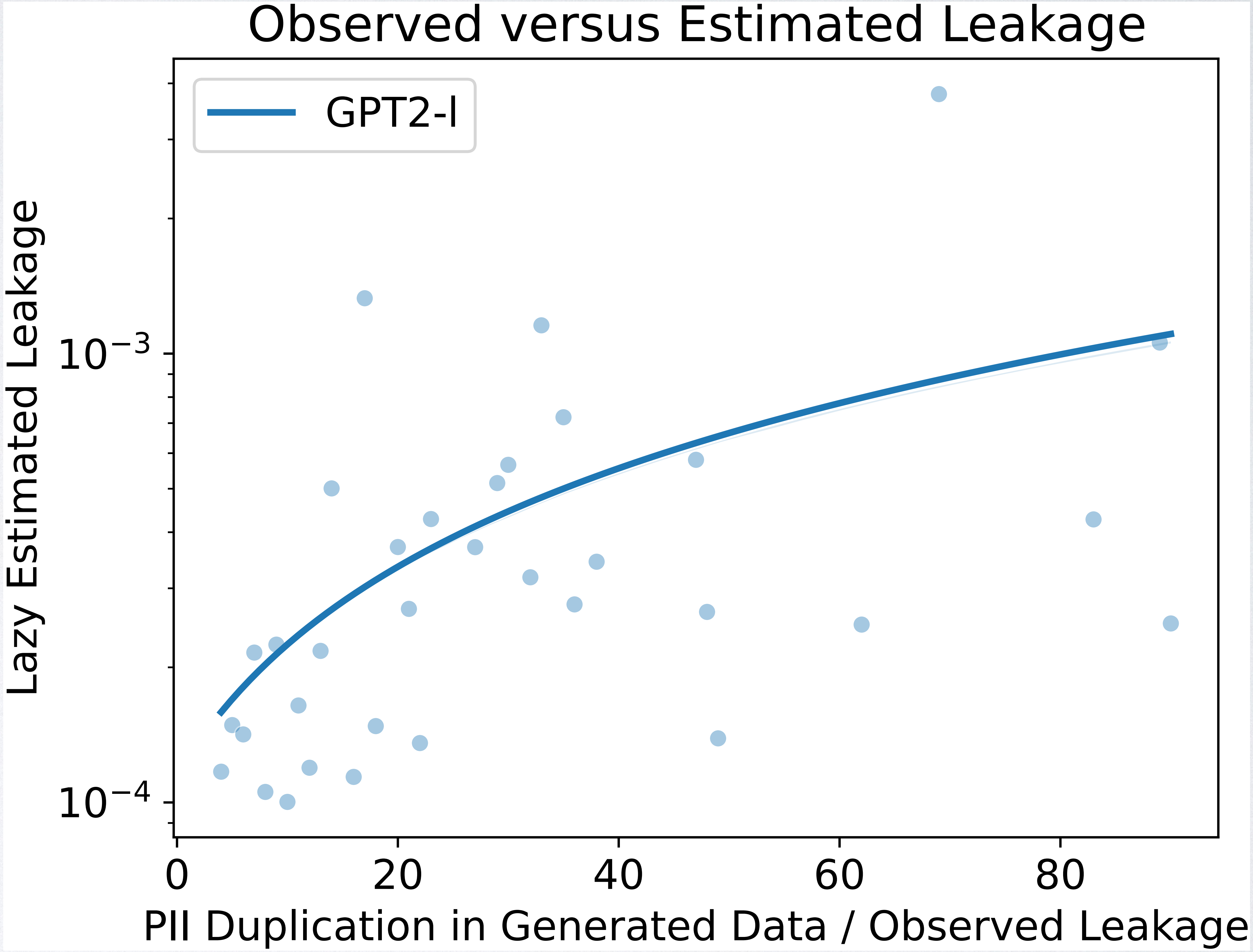
up to 7x Improvement

PII Inference

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

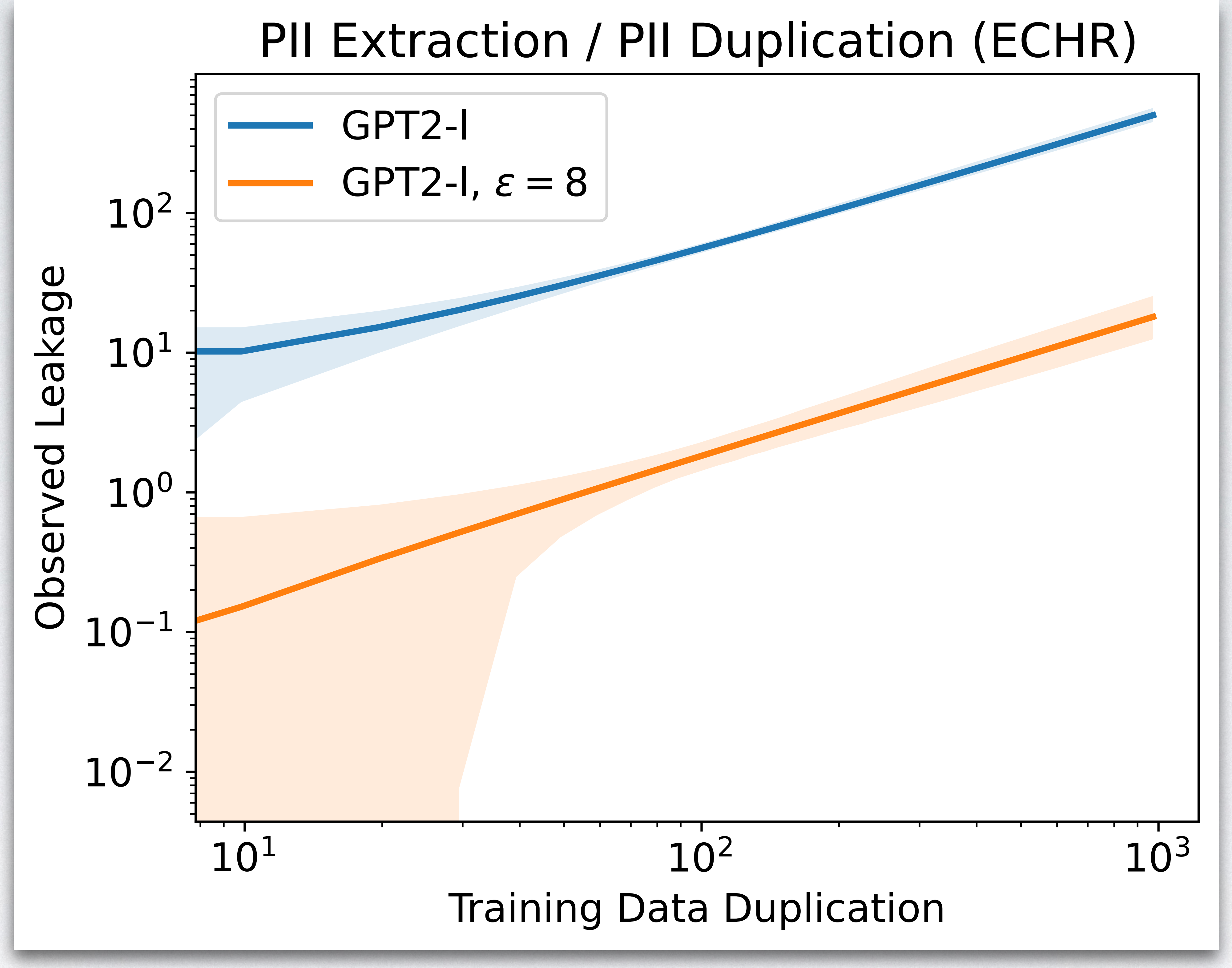
Extraction Attack

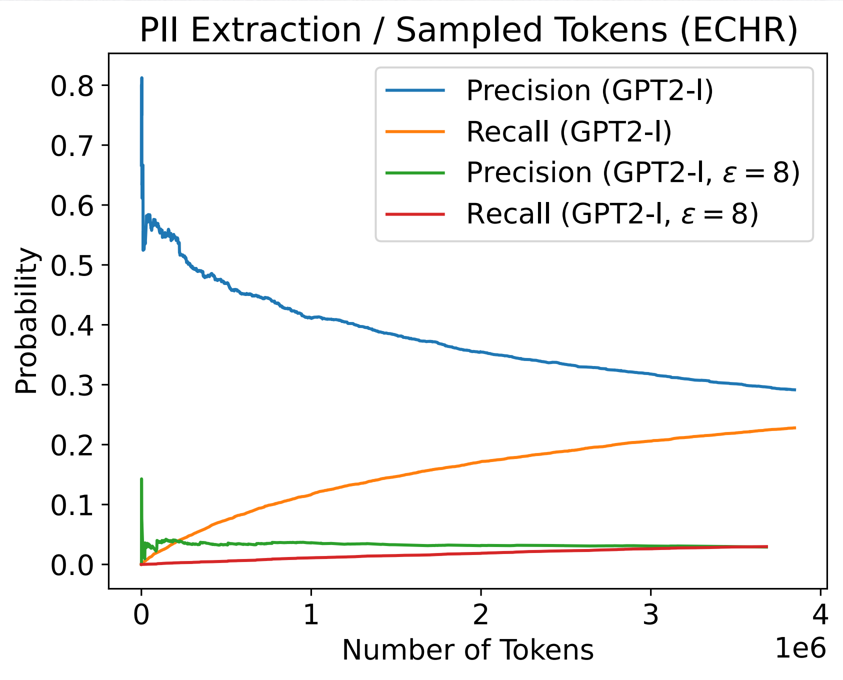
Once upon a time, there existed a tale of two medical students. In the year 2022, they resided at Sunset Street while pursuing their medical education. Alongside his friend, he worked at the LHS Hospital located in the bustling heart of downtown London. Before donning their white coats, both John Doe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

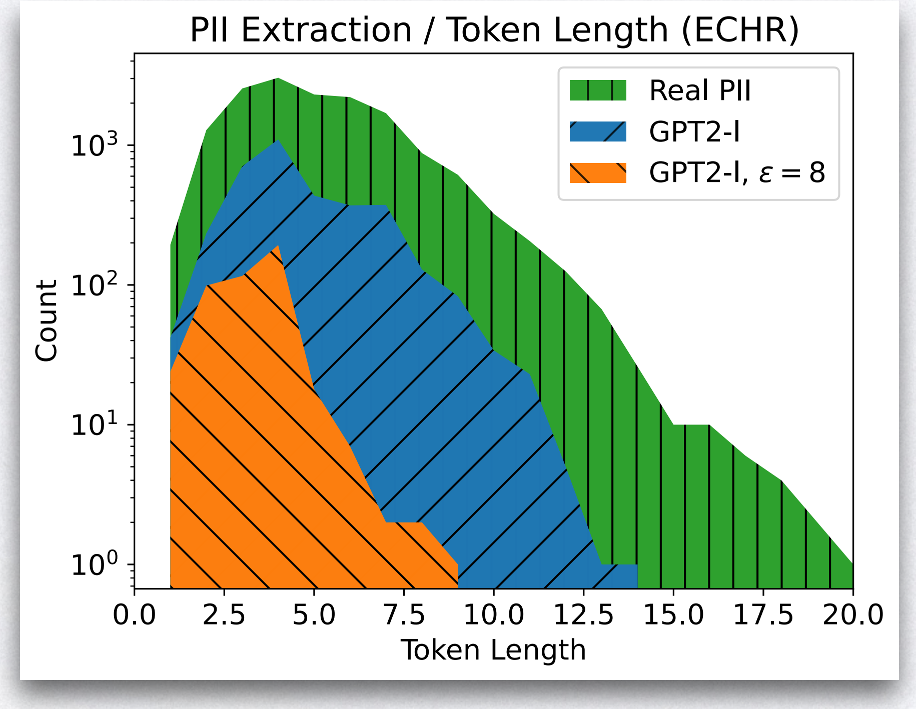


PII Extraction

Duplicated PII are leaked more often



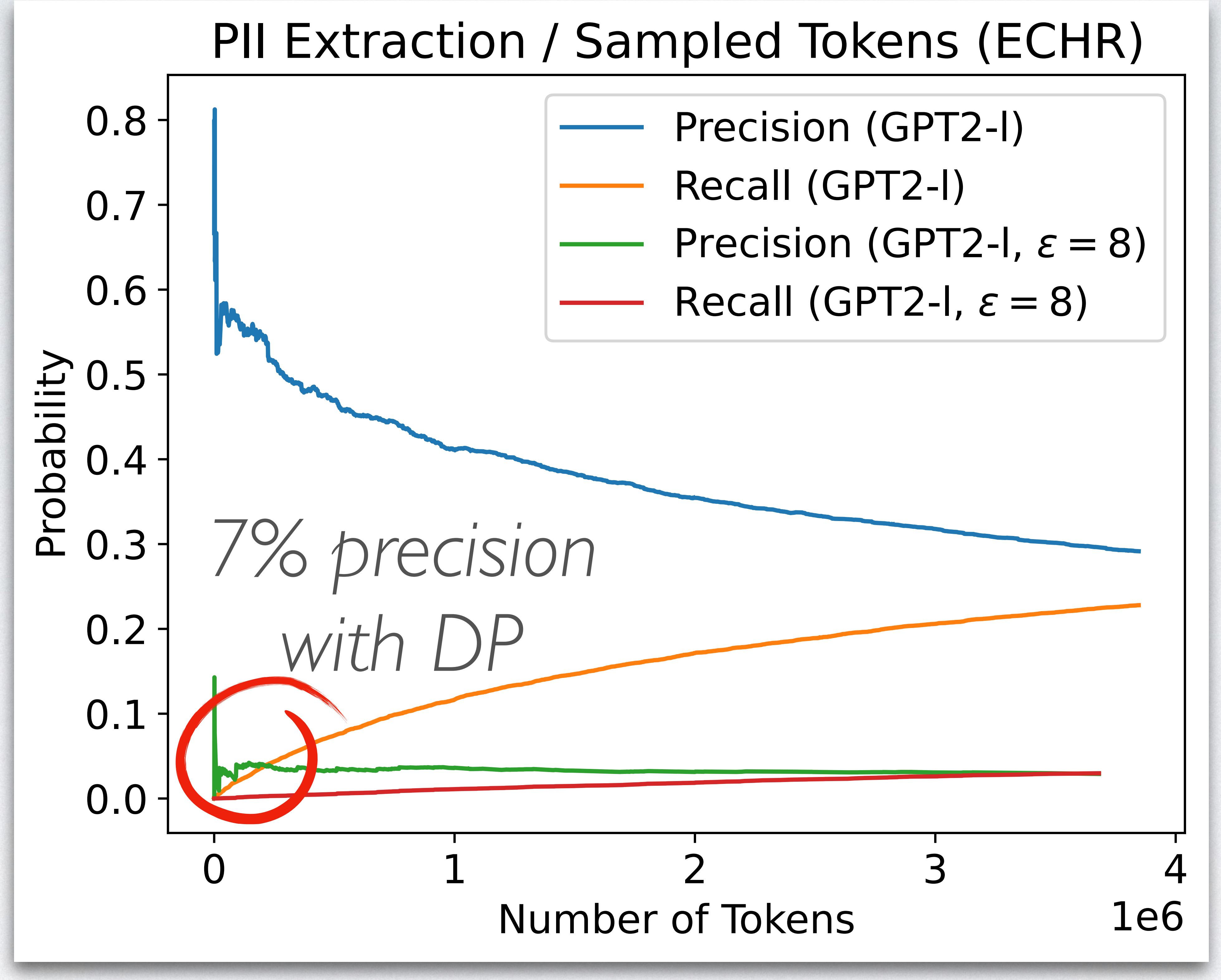


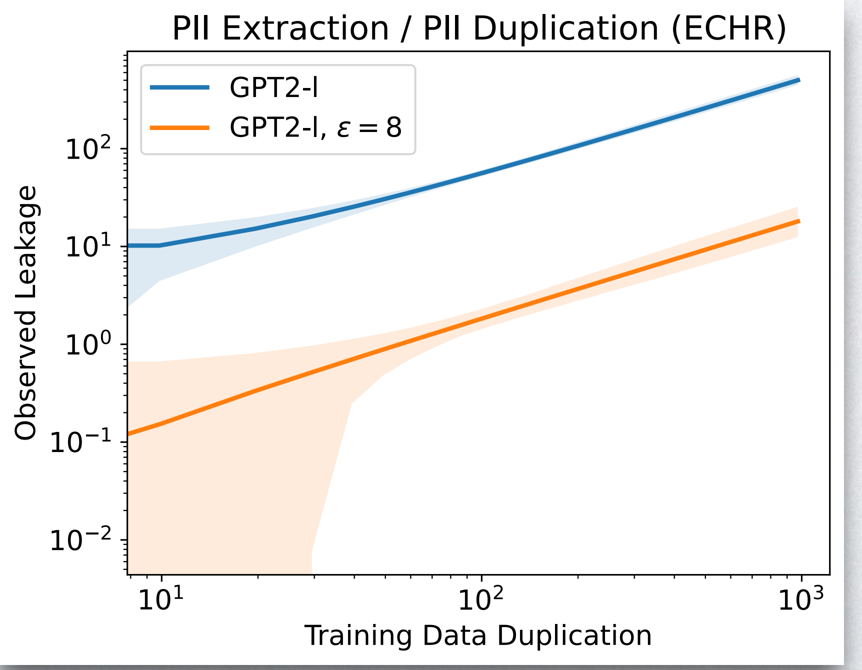


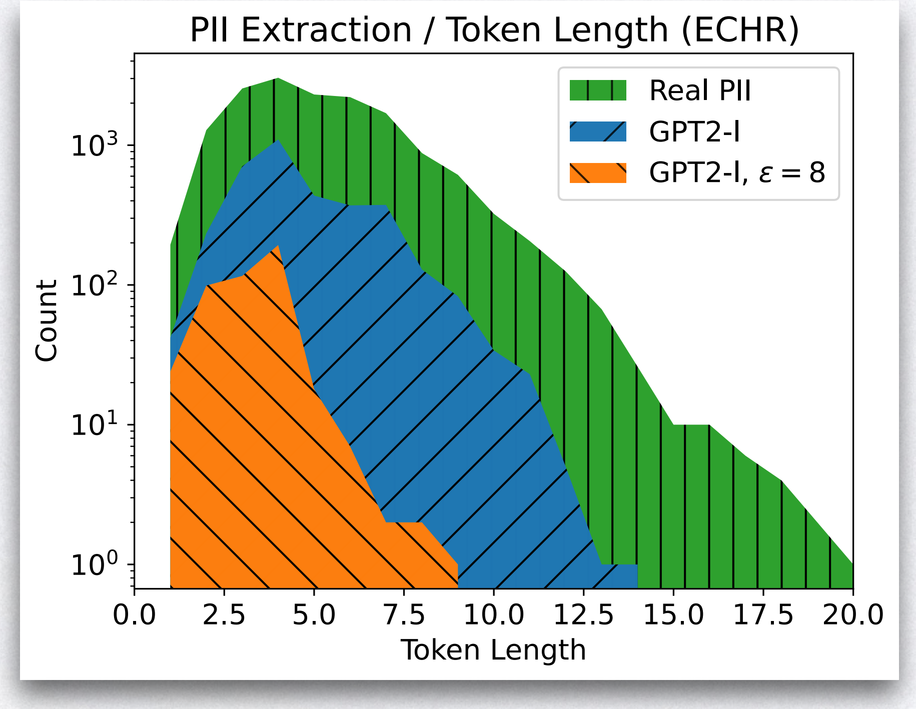
	GPT2-Small		GPT2-Medium		GPT2-Large	
	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$
ECHR						
Prec	24.91%	2.90%	28.05%	3.02%	29.56%	2.92%
Recall	9.44%	2.98%	12.97%	3.21%	22.96%	2.98%
Enron						
Prec	33.86 %	9.37%	27.06%	12.05%	35.36%	11.57%
Recall	6.26%	2.29%	6.56%	2.07%	7.23%	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



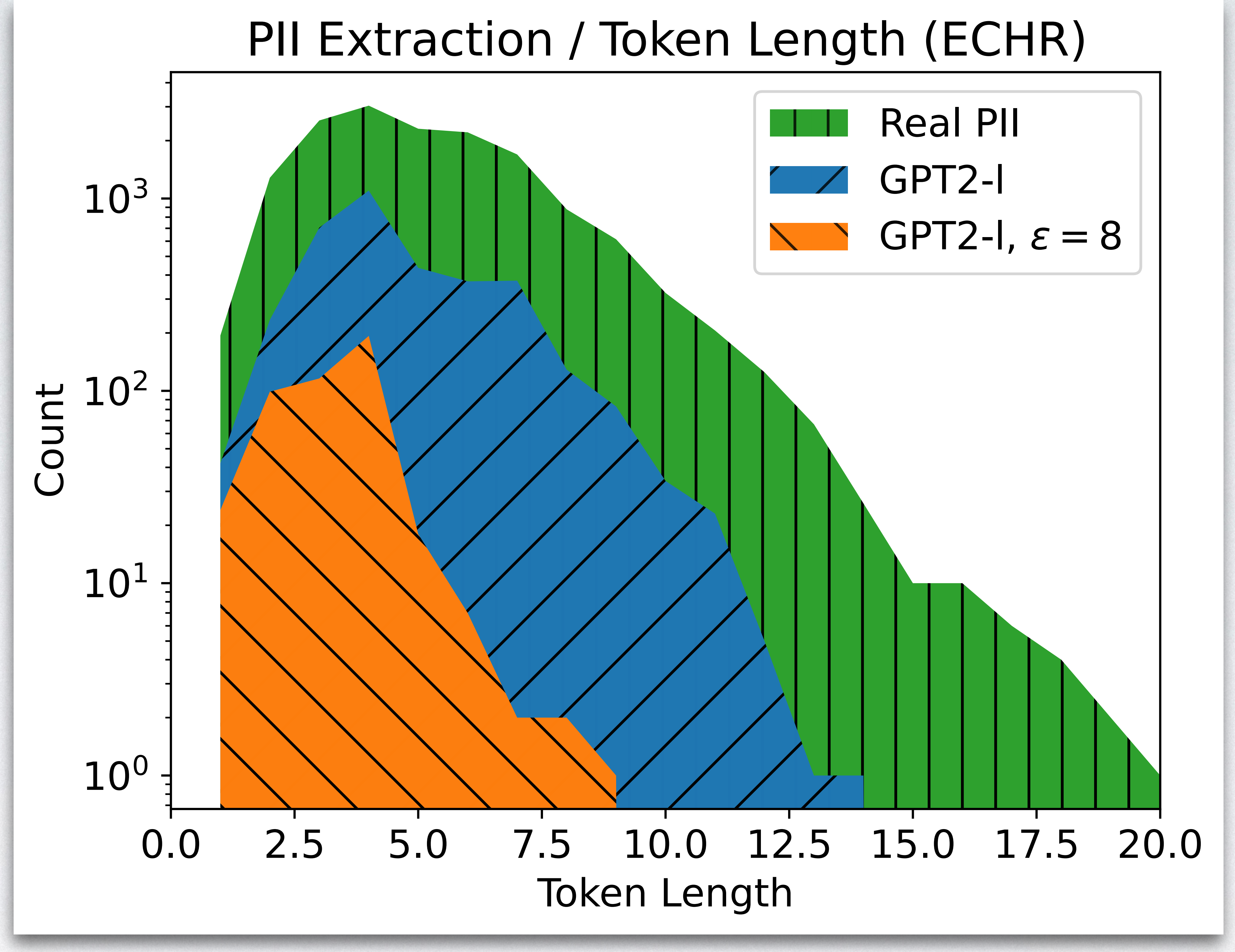


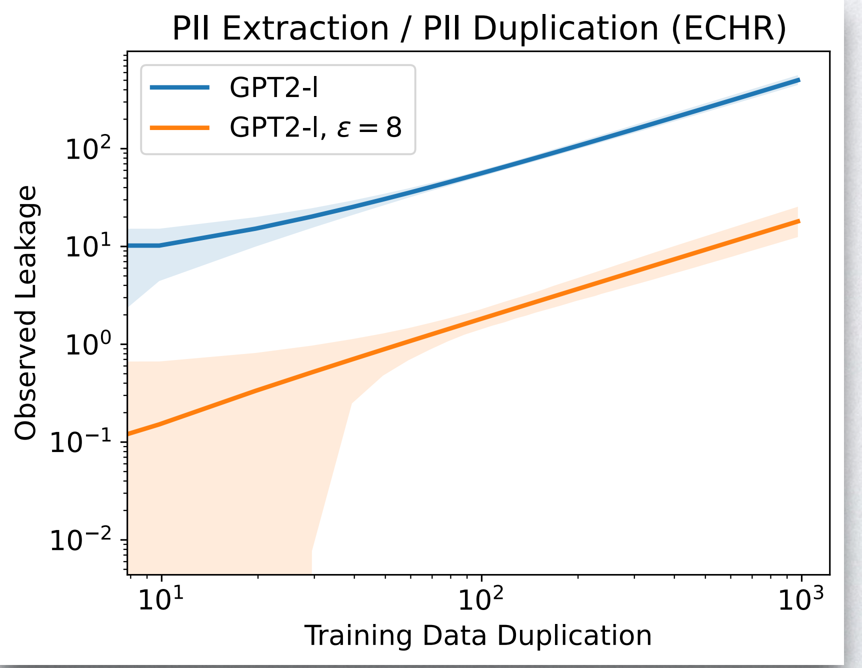


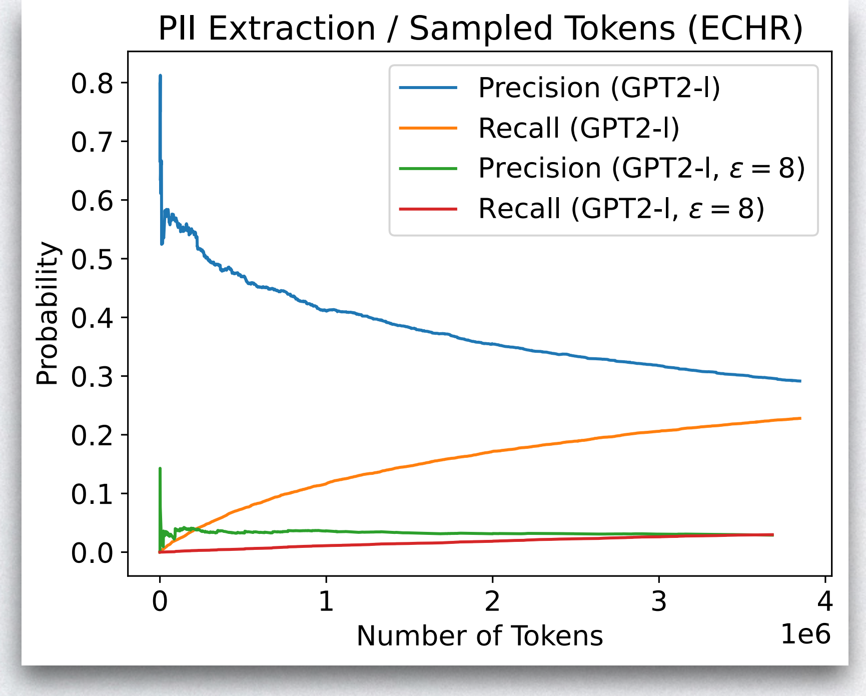
	GPT2-Small		**GPT2-Medium**		**GPT2-Large**	
	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$
			ECHR			
Prec	24.91%	2.90%	28.05%	3.02%	29.56%	2.92%
Recall	9.44%	2.98%	12.97%	3.21%	22.96%	2.98%
			Enron			
Prec	33.86 %	9.37%	27.06%	12.05%	35.36%	11.57%
Recall	6.26%	2.29%	6.56%	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





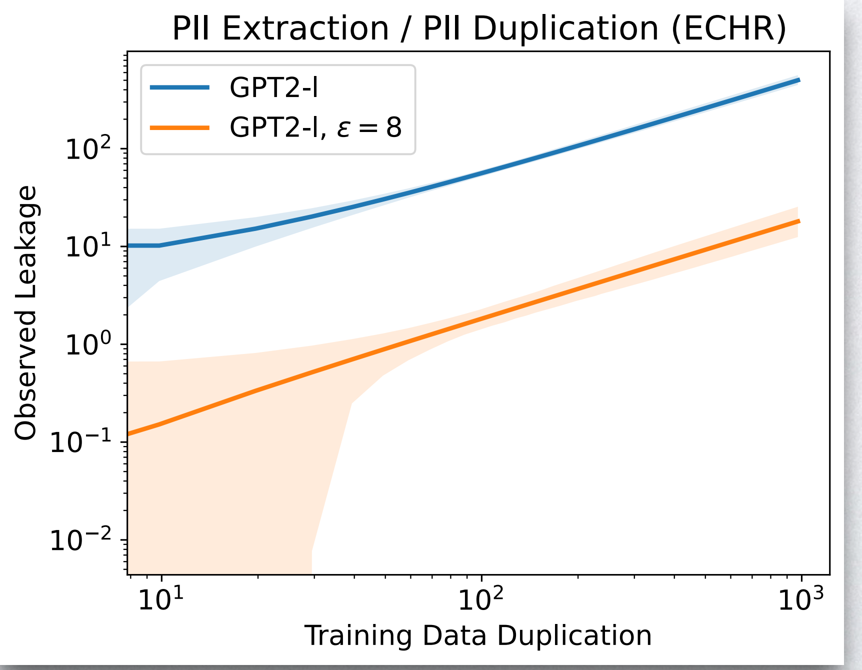


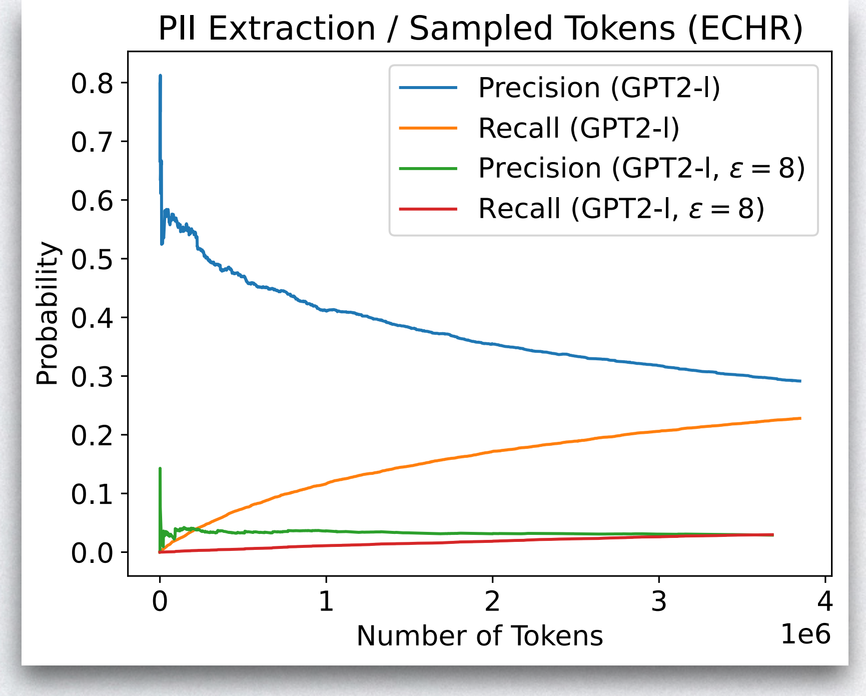
	GPT2-Small		GPT2-Medium		GPT2-Large	
	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$
ECHR						
Prec	24.91%	2.90%	28.05%	3.02%	29.56%	2.92%
Recall	9.44%	2.98%	12.97%	3.21%	22.96%	2.98%
Enron						
Prec	33.86 %	9.37%	27.06%	12.05%	35.36%	11.57%
Recall	6.26%	2.29%	6.56%	2.07%	7.23%	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%

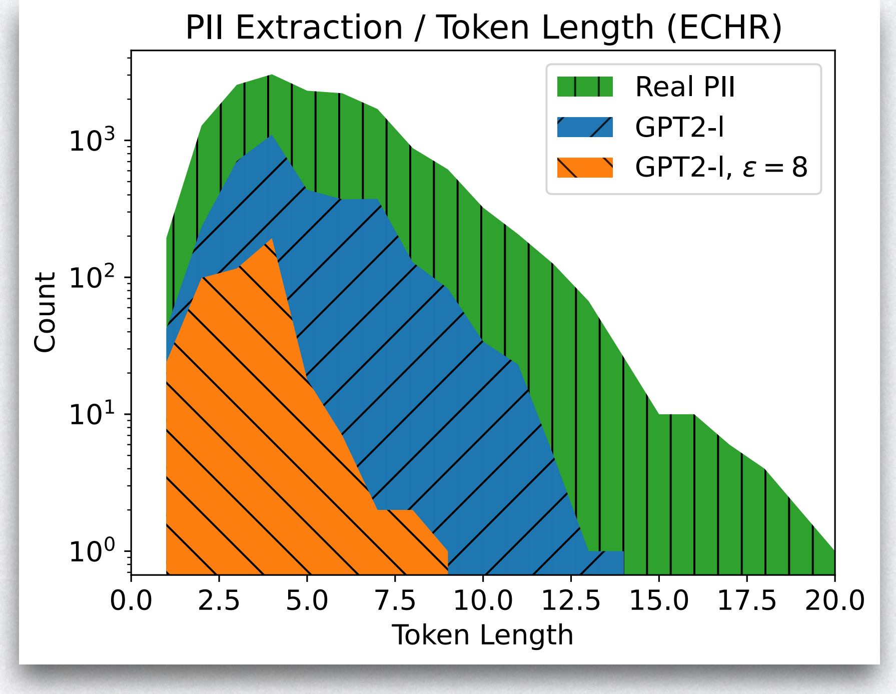
PII Extraction

Higher recall in larger models

	GPT2-Small		GPT2-Medium		GPT2-Large	
	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$
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%

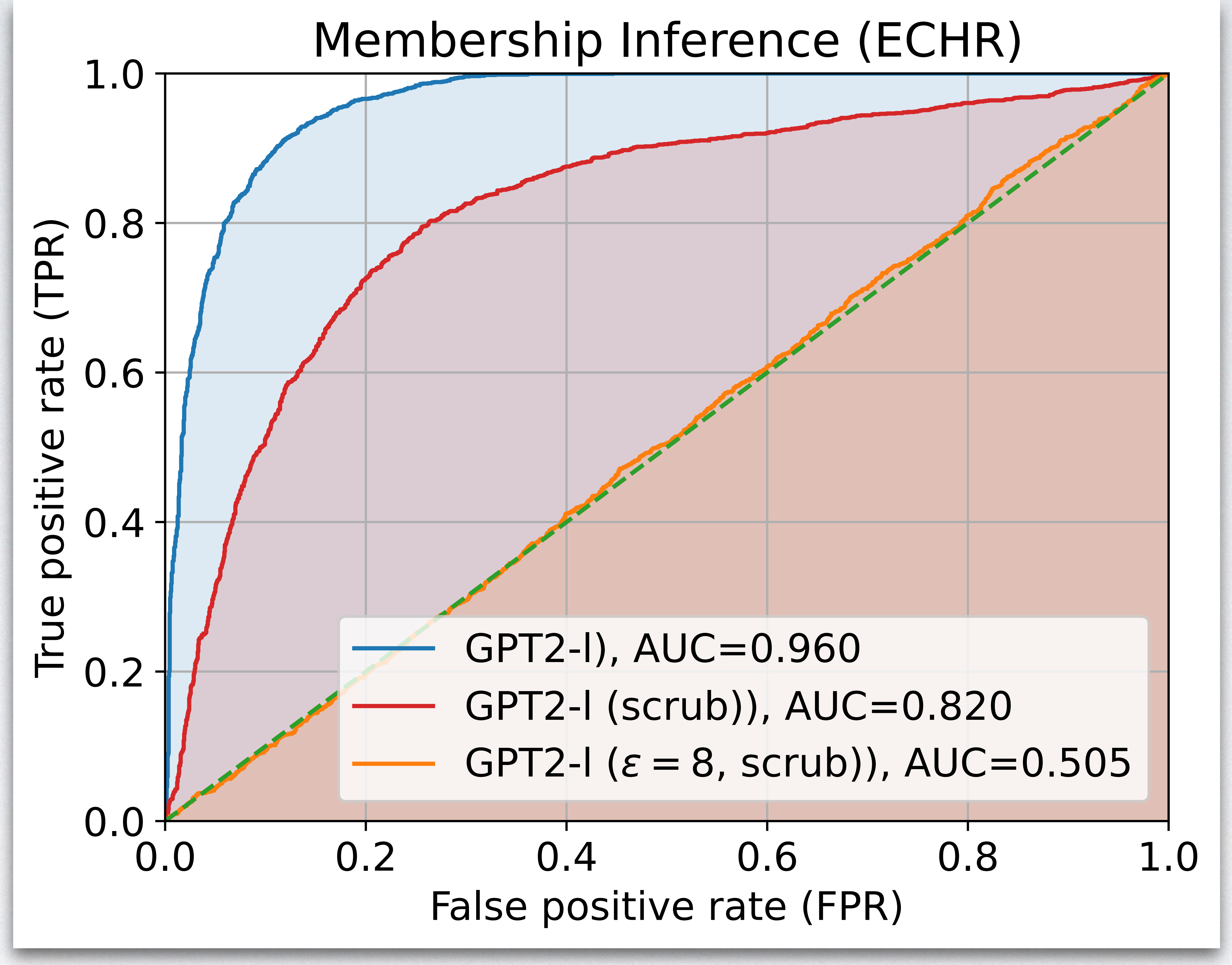


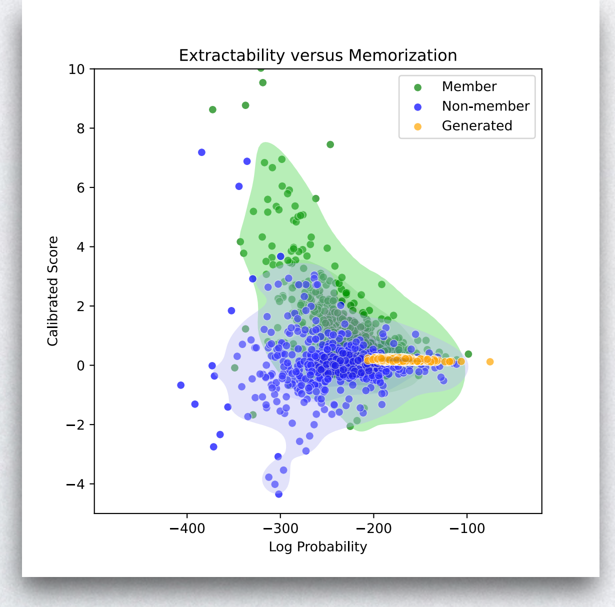


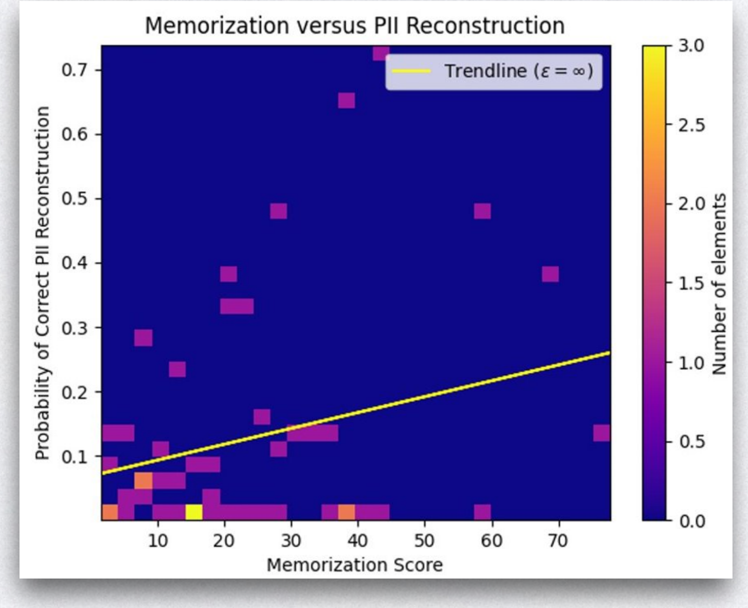


Membership Inference

Scrubbing does not prevent MI

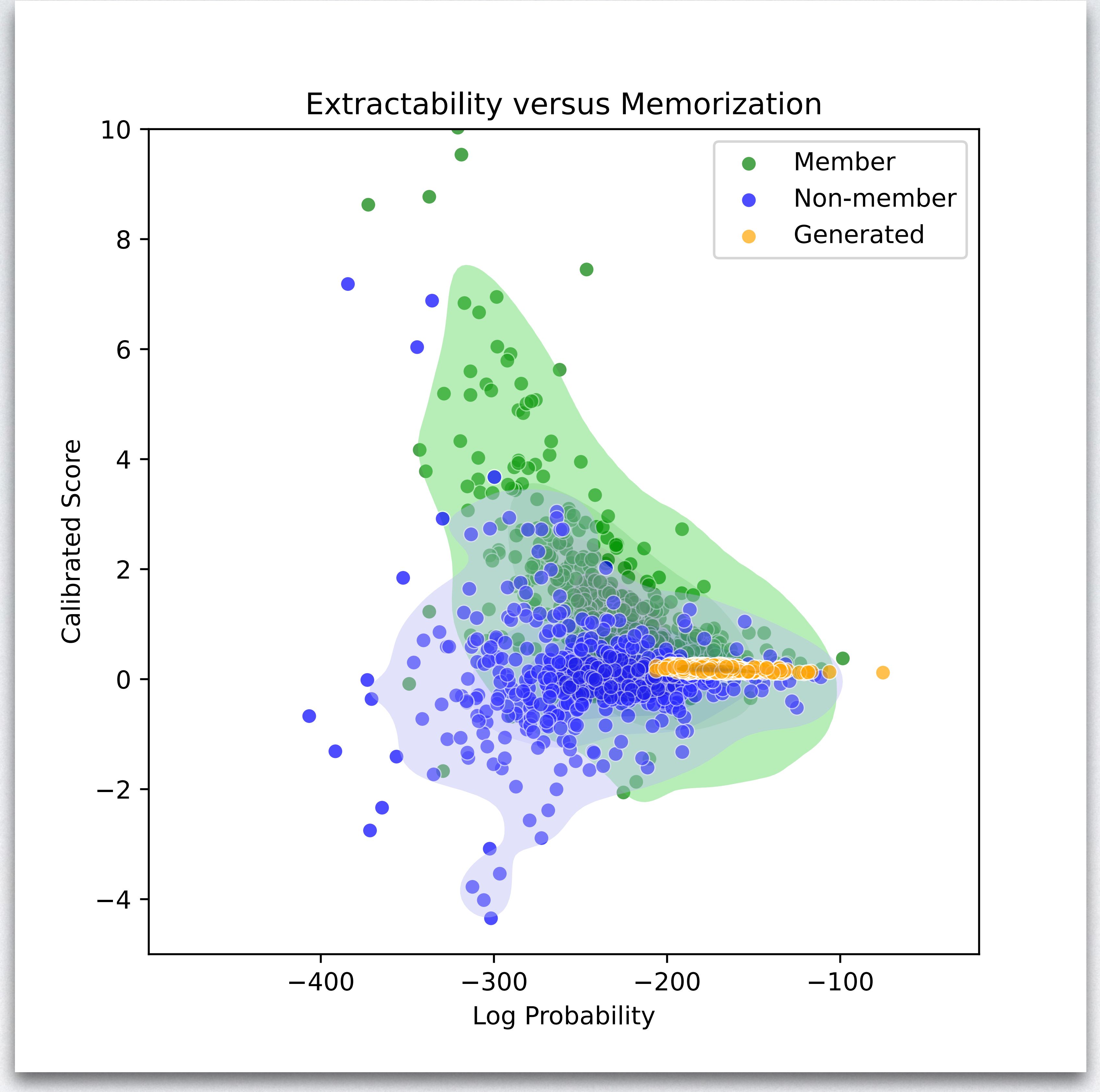


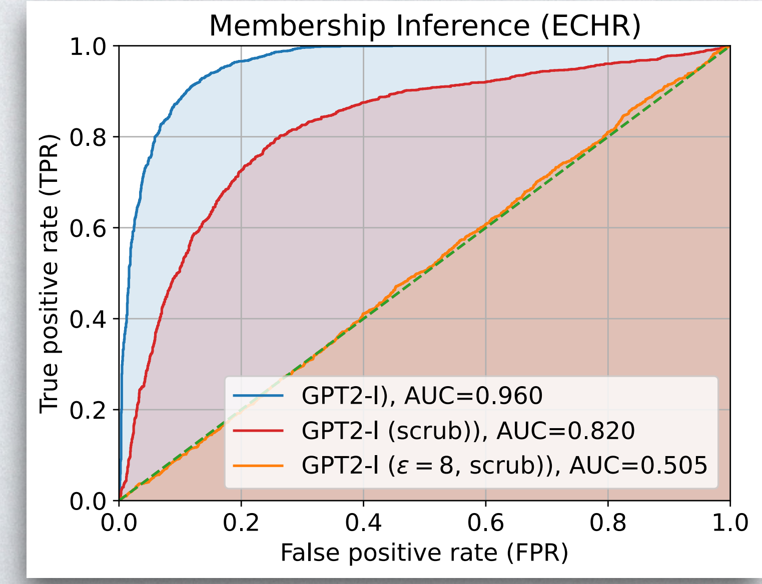


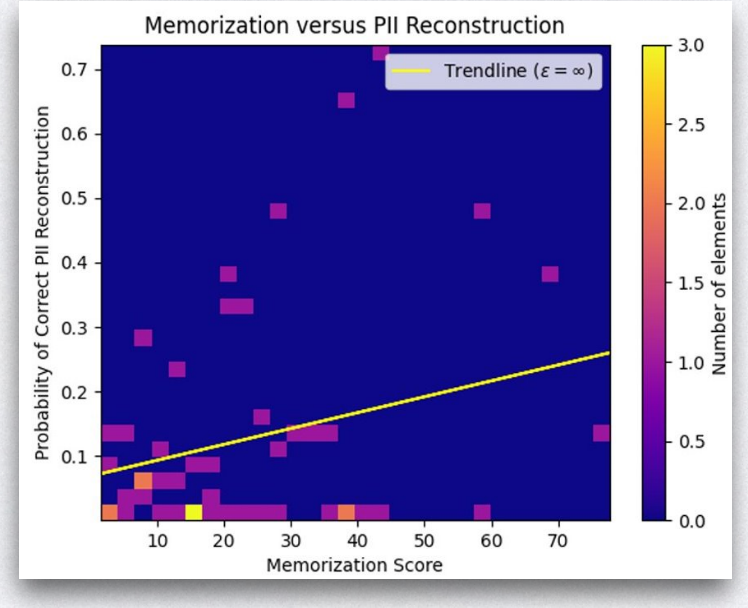


Membership Inference

Randomly generated sequences likely do not contain MI signal

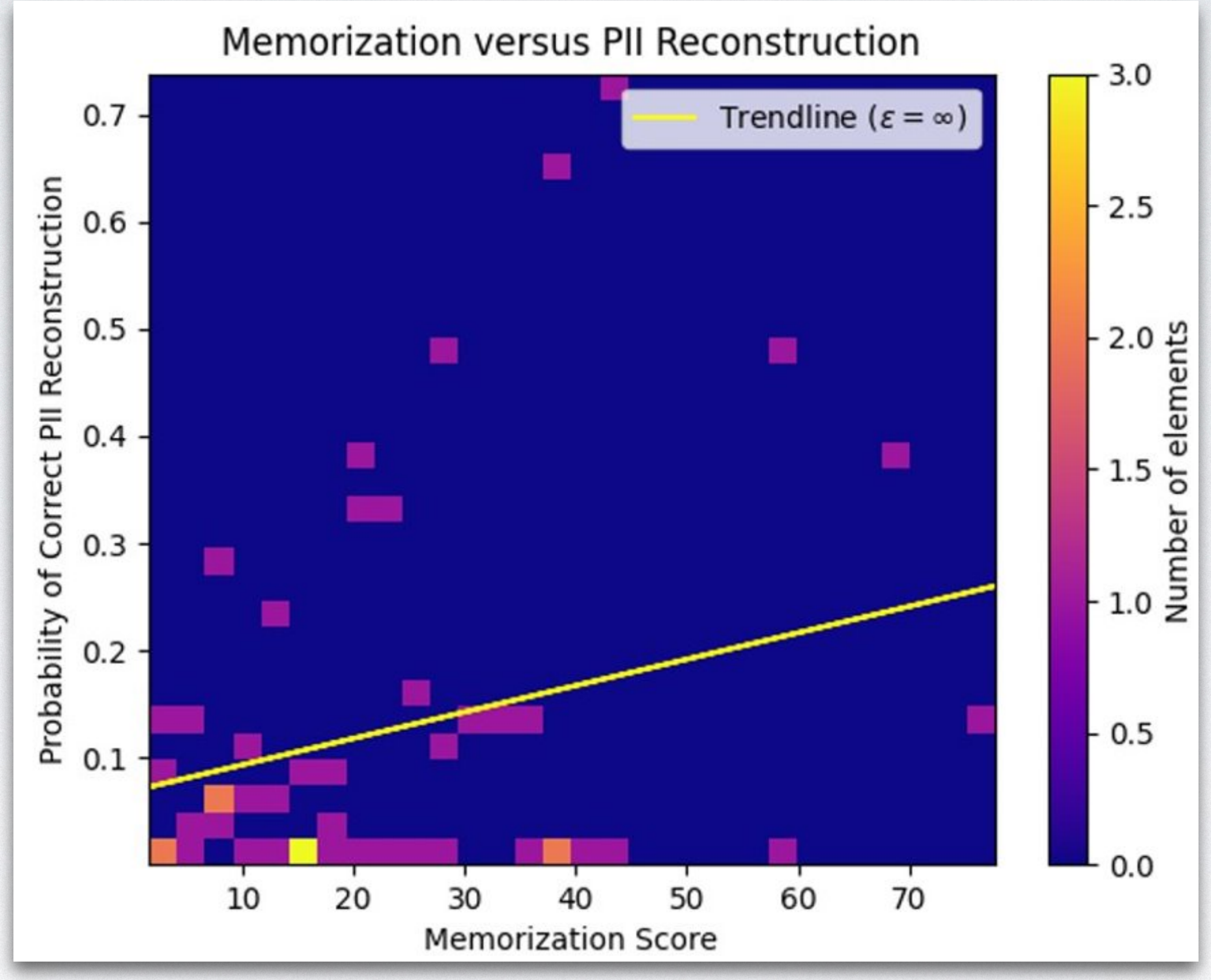


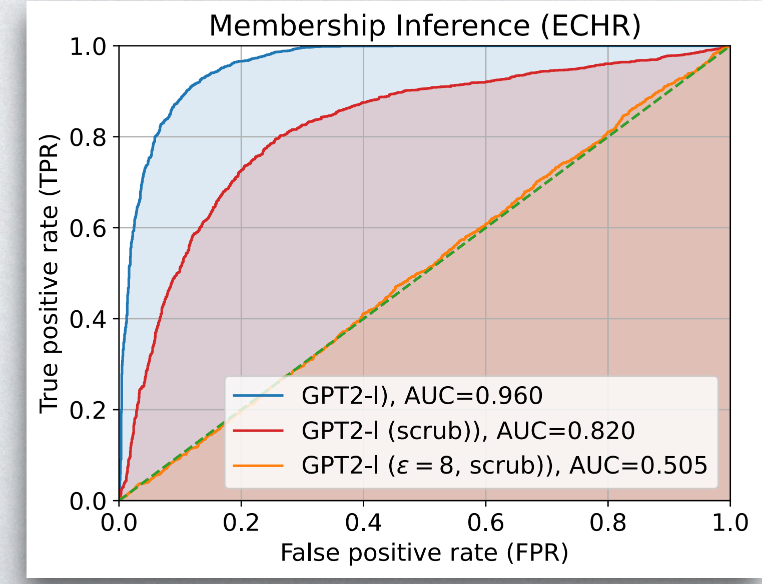


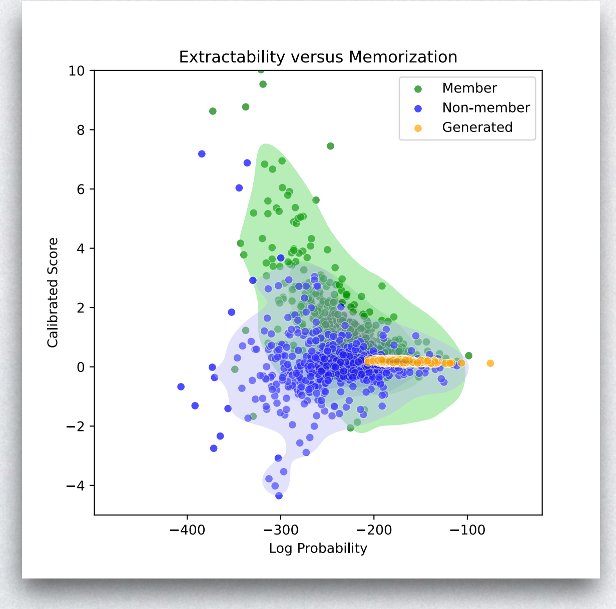


Membership Inference

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

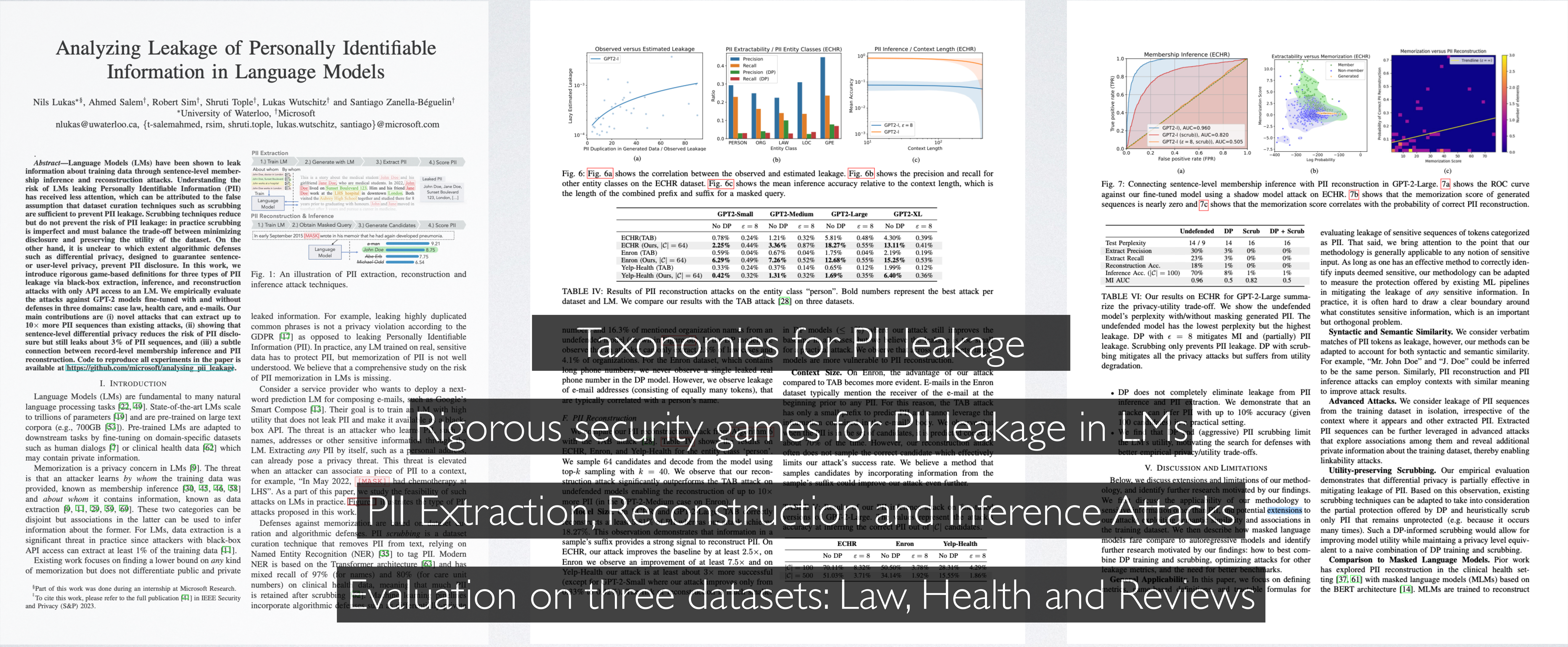
Outlook

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

Taxonomies for PII leakage

Rigorous security games for PII leakage in LMs

PII Extraction, Reconstruction and Inference Attacks

Evaluation on three datasets: Law, Health and Reviews

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