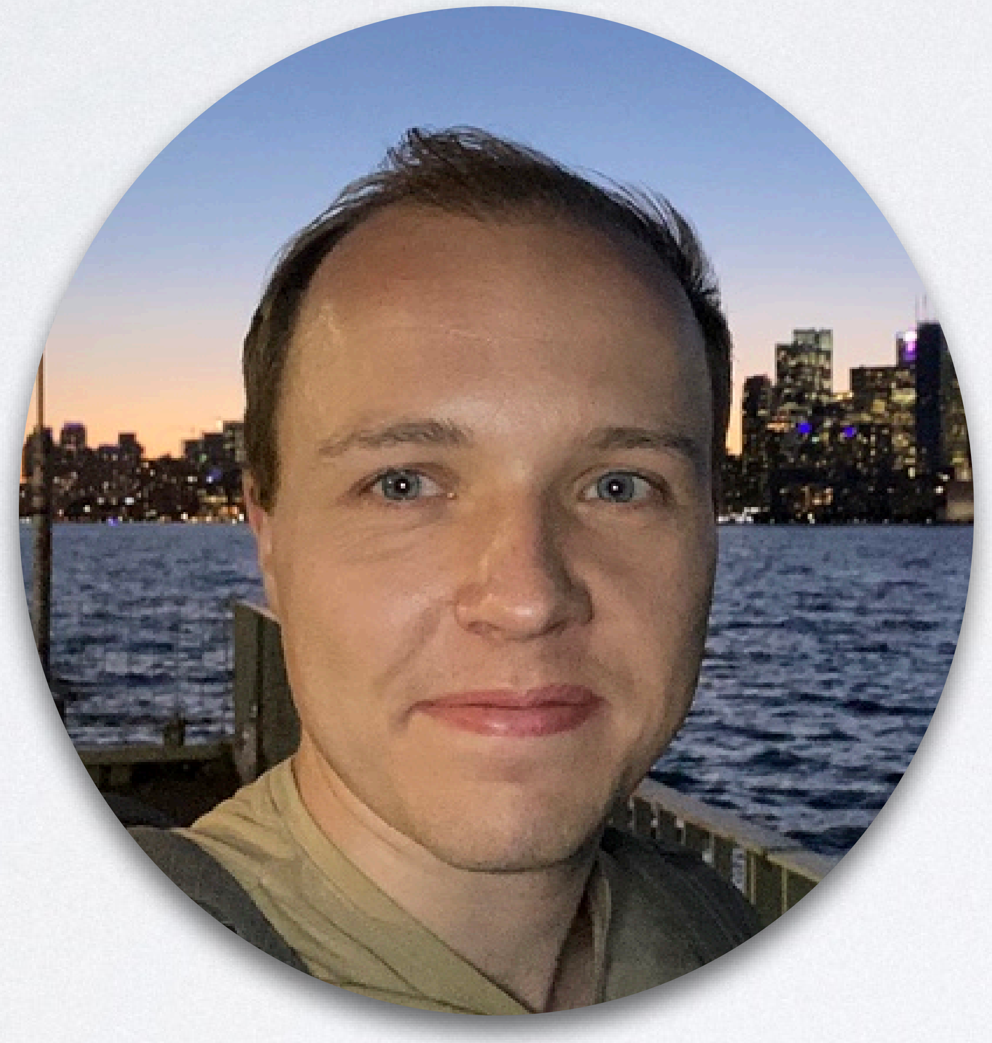
Analyzing Leakage of Personally Identifiable Information in Language Models



https://nilslukas.github.io

Nils Lukas, Dec 11 2023
Research Presentation @Meta

CrySP
Cryptography, Security, and Privacy Research Group

UNIVERSITY OF WATERLOO

My Areas of Research

Nils Lukas

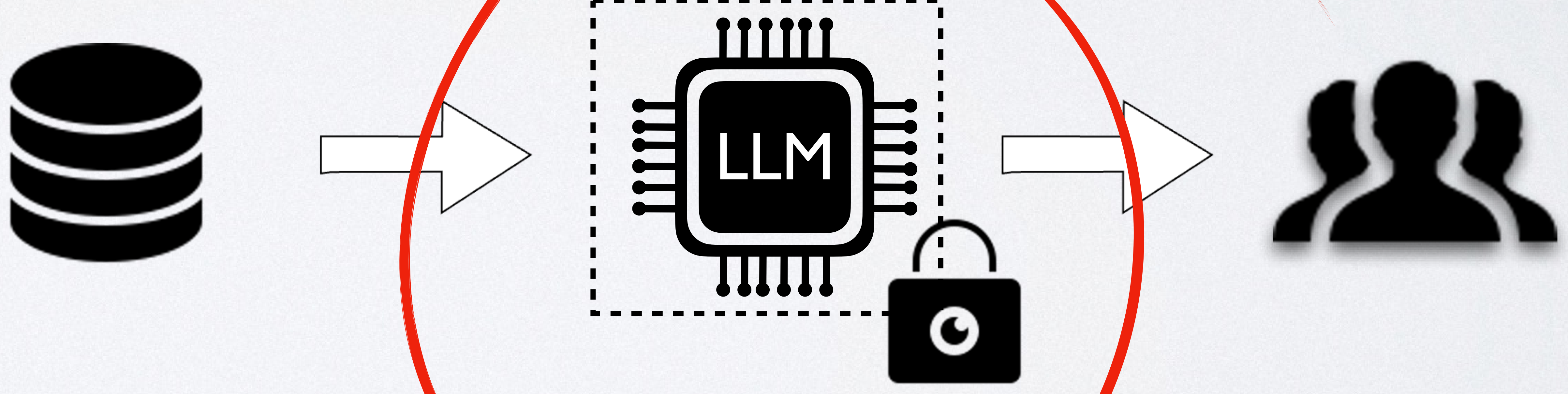
https://nilslukas.github.io

Private Computation

- Private Set Intersection
- Secure Inference

Machine Learning

- Reliability
- Privacy
- Safety

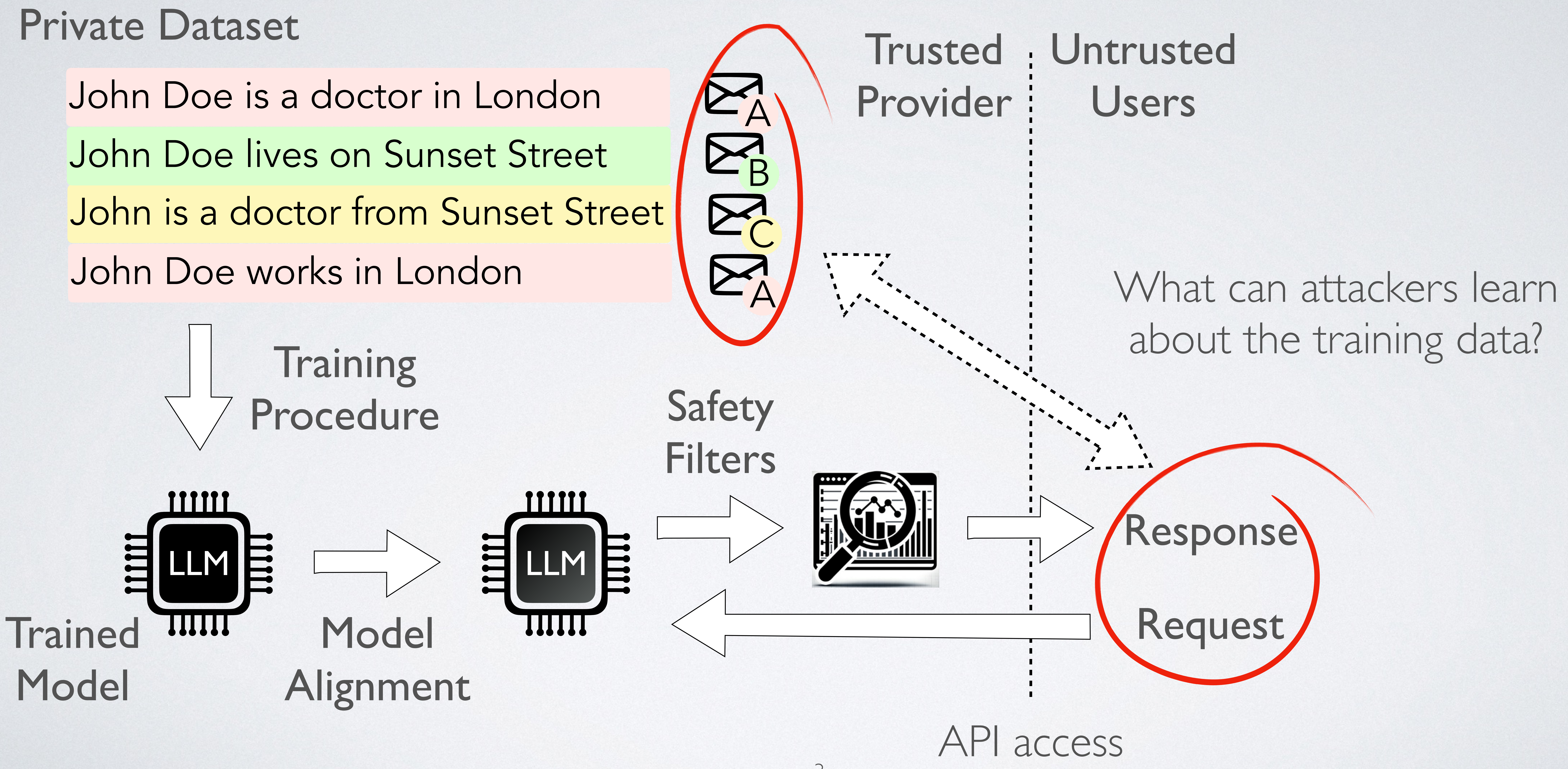


Untrustworthy Data

Untrustworthy Models

Untrustworthy Users

Data Privacy for Large Language Models



Privacy Concerns

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]

Lack of Privacy in 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]

SECRETS DETECTION

Yes, GitHub's Copilot can Leak (Real) Secrets

Researchers successfully extracted valid hard-coded secrets from Copilot and CodeWhisperer, shedding light on a novel security risk associated with the proliferation of secrets.

GitGuardian, 2023 [11]

Terms of Service

6. **Will you use my conversations for training?**

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

ChatGPT by 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 by Google [7]

Privacy Threats

2.7 Privacy

GPT-4 has learned from a variety o
include publicly available personal
about people who have a significa
figures. GPT-4 can also synthesize
reasoning within a given completio
to personal and geographic inforn
with a phone number or answering
without browsing the internet. Fo
address to a phone number with a
as being through that route. By
potential to be used to attempt to

rces, which may
have knowledge
rities and public
multiple steps of
that may relate
tions associated
completion and
University email
ain its reasoning
GPT-4 has the
de data.

GPT

scientific reports

Check for updates

OPEN

Man vs the machine in the struggle for effective text anonymisation in the age of large language models

Constantinos Patsakis[1,2,4] & Nikolaos Lykousas[2,3,4]

The collection and use of personal data are becoming more common in today's data-driven culture. While there are many advantages to this, including better decision-making and service delivery, it also poses significant ethical issues around confidentiality and privacy. Text anonymisation tries to prune and/or mask identifiable information from a text while keeping the remaining content intact to alleviate privacy concerns. Text anonymisation is especially important in industries like healthcare, law, as well as research, where sensitive and personal information is collected, processed, and exchanged under high legal and ethical standards. Although text anonymisation is widely adopted in practice, it continues to face considerable challenges. The most significant challenge is striking a balance between removing information to protect individuals' privacy while maintaining the text's usability for future purposes. The question is whether these anonymisation methods sufficiently reduce the risk of re-identification, in which an individual can be identified based on the remaining information in the text. In this work, we challenge the effectiveness of these methods and how we perceive identifiers. We assess the efficacy of these methods against the elephant in the room, the use of AI over big data. While most of the research is focused on identifying and removing personal information, there is limited discussion on whether the remaining information is sufficient to deanonymise individuals and, more precisely, who can do it. To this end, we conduct an experiment using GPT over anonymised texts of famous people to determine whether such trained networks can deanonymise them. The latter allows us to revise these methods and introduce a novel methodology that employs Large Language Models to improve the anonymity of texts.

In today's data-driven society, the collection and use of personal information are becoming increasingly prevalent. While this has numerous benefits, such as improved decision-making and better service provision, it also raises important ethical concerns related to privacy and confidentiality. Indeed, harvesting user data is a common practice of far too many online platforms and services with a significant impact on citizens. This has been one of the pillars that led to the introduction of the General Data Protection Regulation (GDPR)[1] and other relevant legislation around the world as a means to address the privacy issues that emerged. The GDPR mandates using privacy-preserving methods and processes throughout the data management lifecycle, from collection and processing to sharing and publishing. One of these fundamental methods is anonymisation. Given that modern organisations continuously deal with documents, the above has served as a catalyst in the emergence of text anonymisation as a research topic with many practical applications. The general concept is that given a text, one has to remove or mask identifiable information while preserving the remaining content. Text anonymisation is particularly relevant in healthcare, law, and research, where personal and sensitive information is overwhelming and must be protected to comply with privacy regulations and ethical guidelines.

Although text anonymisation has been widely adopted in practice, it still faces significant challenges. These methods must strike a balance between the need to protect the privacy of individuals and the need to preserve the data utility. Let us consider this with an example where the anonymisation task is to anonymise the sentence "*Volodymyr Zelenskyy is the president of Ukraine*". Clearly, simply removing the name is not enough. If one is given the sentence "$\boxed{\textit{NAME}}$ *is the president of Ukraine*", it is trivial to recover the missing information. Therefore, the anonymised sentence would be "$\boxed{\textit{NAME}}$ *is the president of* $\boxed{\textit{COUNTRY}}$". To this end, named entity

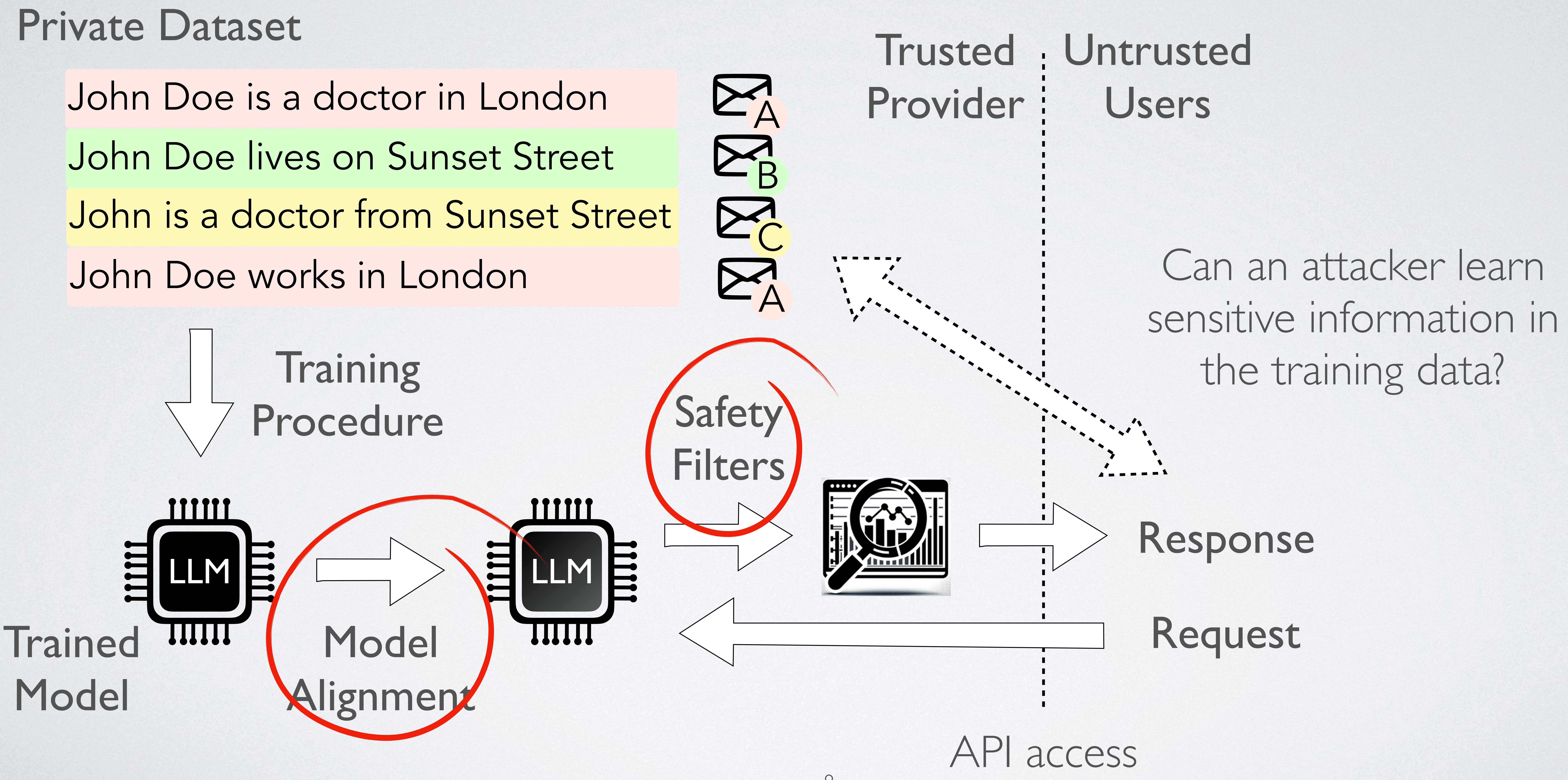
[1]Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou str, 18534 Piraeus, Greece. [2]Management Systems Institute of Athena Research Centre, Marousi, Greece. [3]Data Centric Services, Bucharest, Romania. [4]These authors contributed equally: Constantinos Patsakis and Nikolaos Lykousas. ✉email: kpatsak@unipi.gr

Scientific Reports | (2023) 13:16026 | https://doi.org/10.1038/s41598-023-42977-3 nature portfolio 1

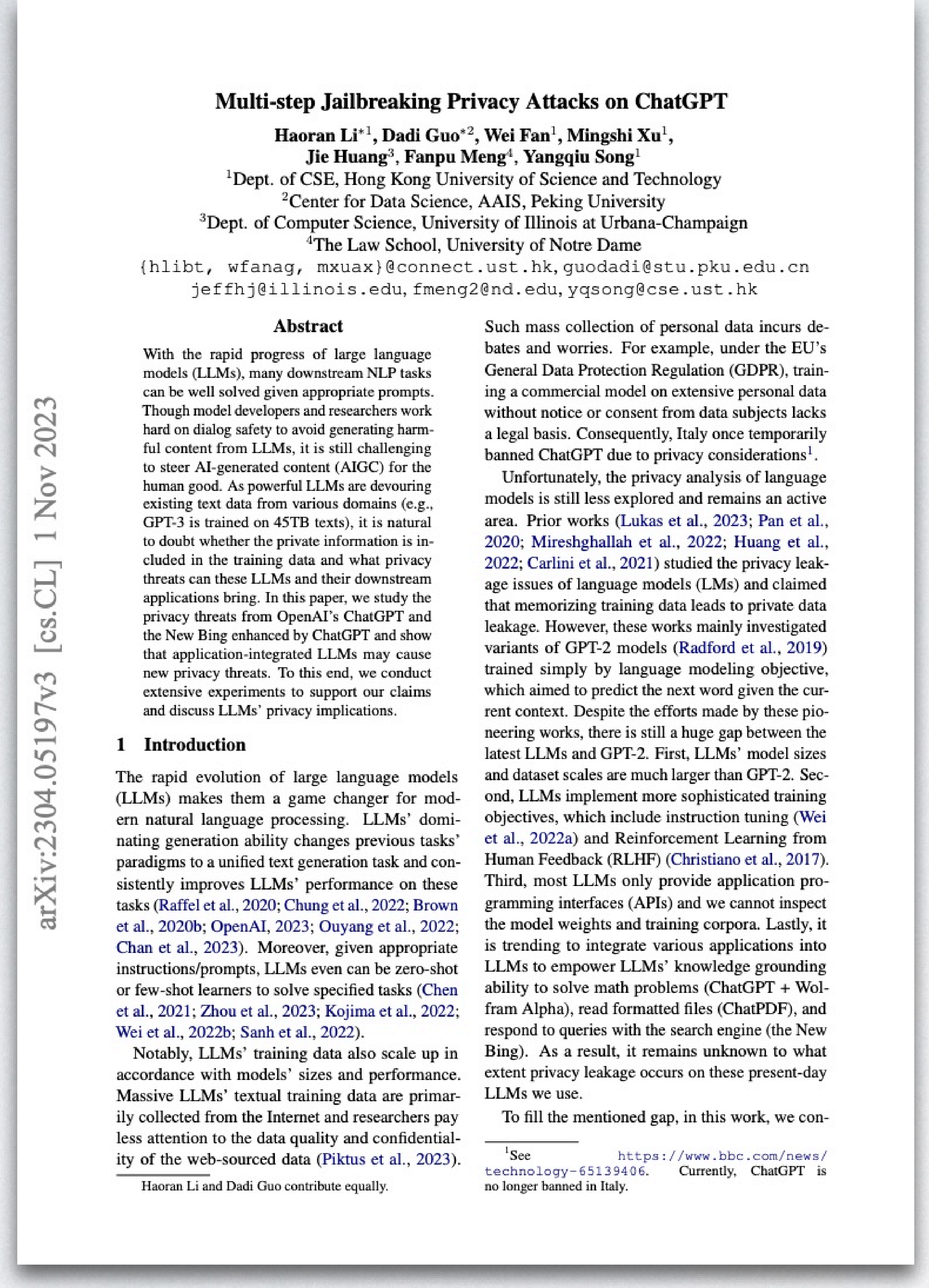
Augmented Data Attack [C]

[B] GPT-4 Technical Report, OpenAI., Preprint, March 2023
[C] Man vs the Machine in the Struggle for Effective Text Anonymization in the Age of Large Language Models, Patsakis et al., Scientific Reports

Data Privacy for Large Language Models



Privacy Attacks by Evading Model Alignment


arXiv:2304.05197v3 [cs.CL] 1 Nov 2023

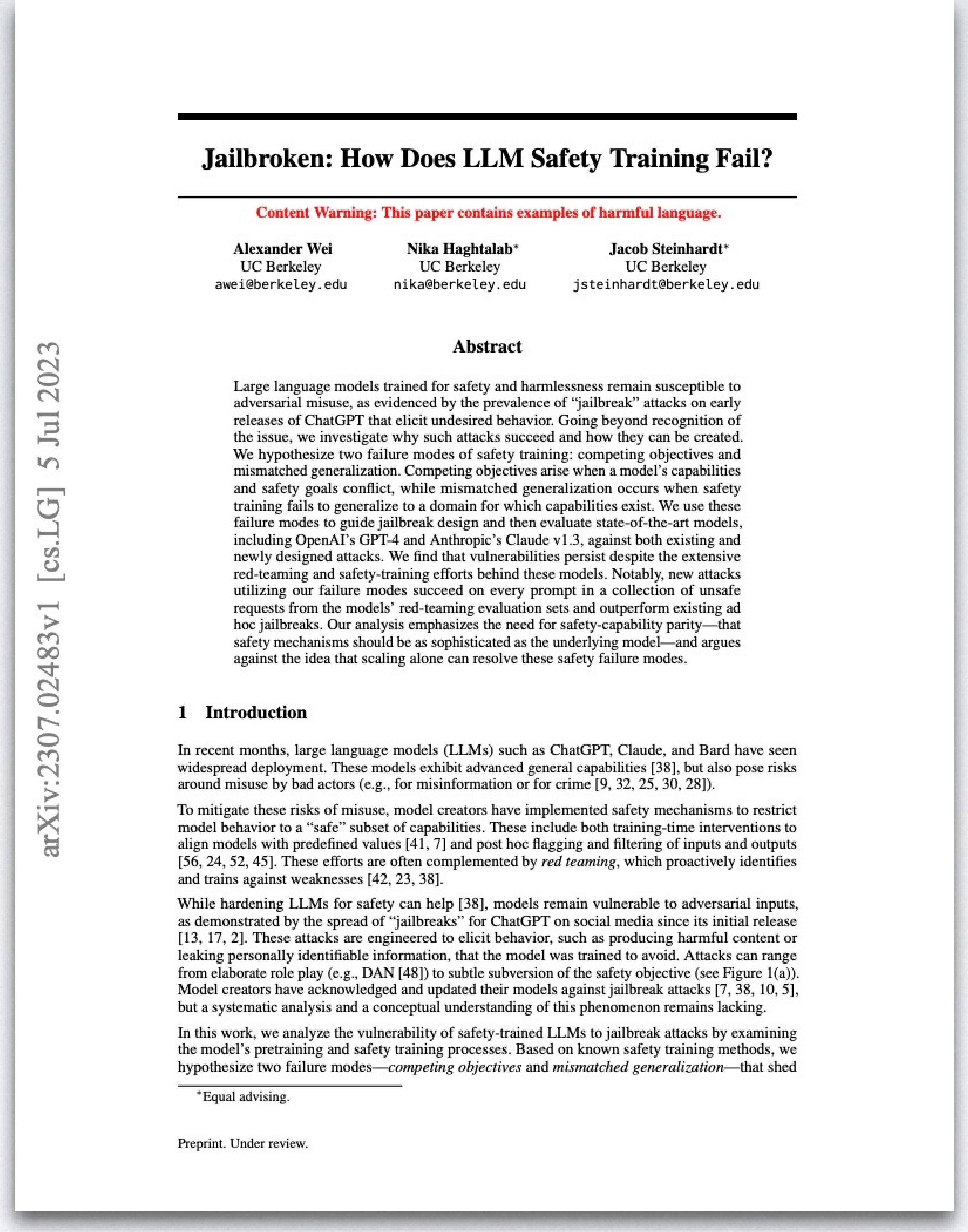
Multi-step Jailbreaking Privacy Attacks on ChatGPT

Haoran Li*[1], Dadi Guo*[2], Wei Fan[1], Mingshi Xu[1], Jie Huang[3], Fanpu Meng[4], Yangqiu Song[1]

[1]Dept. of CSE, Hong Kong University of Science and Technology
[2]Center for Data Science, AAIS, Peking University
[3]Dept. of Computer Science, University of Illinois at Urbana-Champaign
[4]The Law School, University of Notre Dame

Abstract

1 Introduction


arXiv:2307.02483v1 [cs.LG] 5 Jul 2023

Jailbroken: How Does LLM Safety Training Fail?

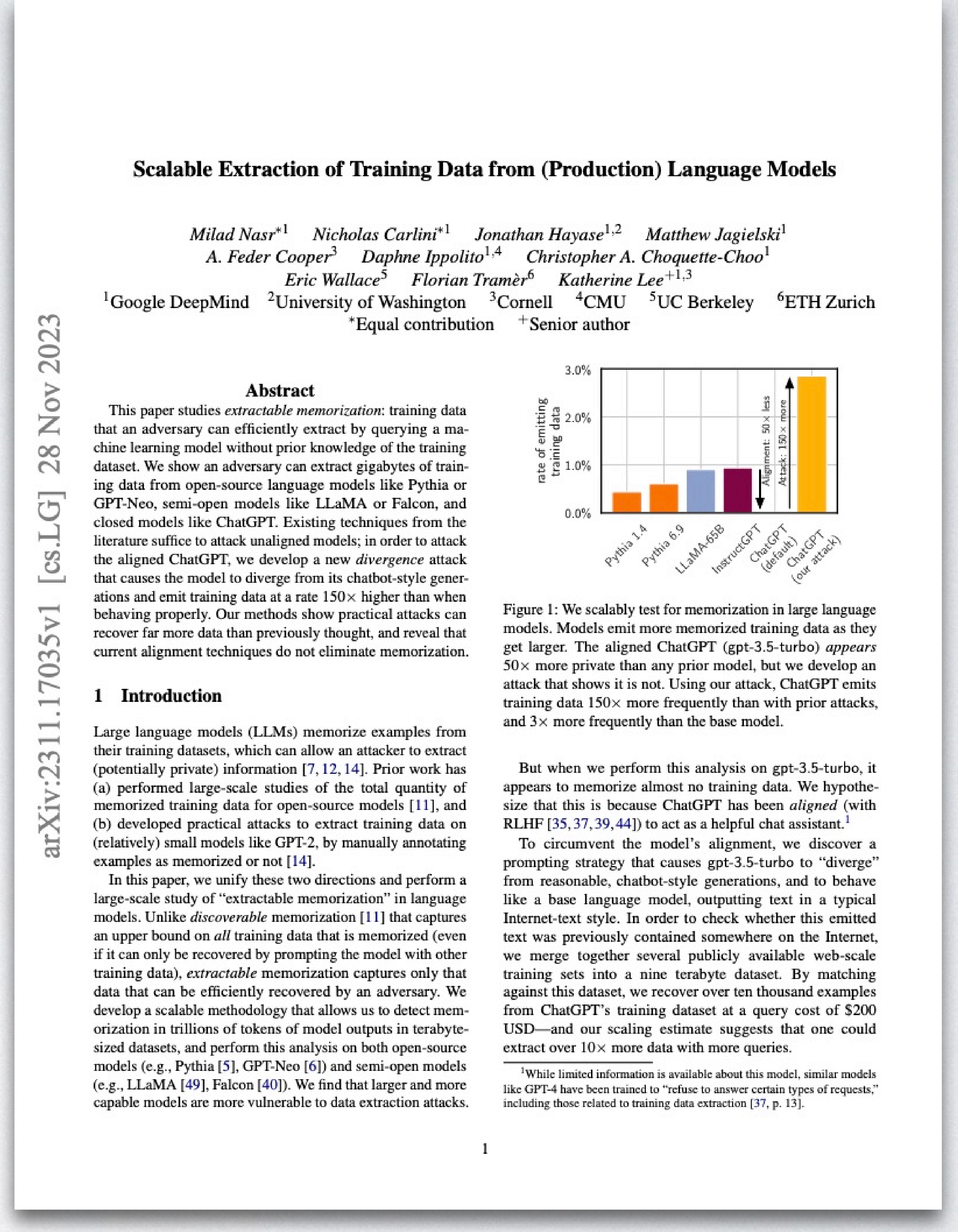
Content Warning: This paper contains examples of harmful language.

Alexander Wei, UC Berkeley, awei@berkeley.edu
Nika Haghtalab*, UC Berkeley, nika@berkeley.edu
Jacob Steinhardt*, UC Berkeley, jsteinhardt@berkeley.edu

Abstract

1 Introduction

Preprint. Under review.


arXiv:2311.17035v1 [cs.LG] 28 Nov 2023

Scalable Extraction of Training Data from (Production) Language Models

Milad Nasr*[1] Nicholas Carlini*[1] Jonathan Hayase[1,2] Matthew Jagielski[1] A. Feder Cooper[3] Daphne Ippolito[1,4] Christopher A. Choquette-Choo[1] Eric Wallace[5] Florian Tramèr[6] Katherine Lee+[1,3]

[1]Google DeepMind [2]University of Washington [3]Cornell [4]CMU [5]UC Berkeley [6]ETH Zurich
*Equal contribution +Senior author

Abstract

Figure 1: We scalably test for memorization in large language models. Models emit more memorized training data as they get larger. The aligned ChatGPT (gpt-3.5-turbo) *appears* 50× more private than any prior model, but we develop an attack that shows it is not. Using our attack, ChatGPT emits training data 150× more frequently than with prior attacks, and 3× more frequently than the base model.

1 Introduction

[C] Multi-step Jailbreaking Privacy Attacks on ChatGPT, Li et al, March 2023

[D] Jailbroken: How Does LLM Safety Training Fail?, Wei et al., Preprint, July 2023

[E] Scalable Extraction of Training Data from (Production) Language Models, Nasr et al., Preprint, November 2023

Privacy Attacks by Evading Safety Filters

Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy

Daphne Ippolito[1] Florian Tramèr*[2] Milad Nasr*[1]

Chiyuan Zhang*[1] Matthew Jagielski*[1] Katherine Lee*[1,3]

Christopher A. Choquette-Choo*[1] Nicholas Carlini[1]

[1] *Google Research* [2] *ETH Zurich* [3] *Cornell University*

Abstract

Studying data memorization in neural language models helps us understand the risks (e.g., to privacy or copyright) associated with models regurgitating training data and aids in the development of countermeasures. Many prior works—and some recently deployed defenses—focus on "verbatim memorization", defined as a model generation that exactly matches a substring from the training set. We argue that verbatim memorization definitions are too restrictive and fail to capture more subtle forms of memorization. Specifically, we design and implement an efficient defense that *perfectly* prevents all verbatim memorization. And yet, we demonstrate that this "perfect" filter does not prevent the leakage of training data. Indeed, it is easily circumvented by plausible and minimally modified "style-transfer" prompts—and in some cases even the non-modified original prompts—to extract memorized information. We conclude by discussing potential alternative definitions and why defining memorization is a difficult yet crucial open question for neural language models.

Figure 1: Illustration of Memorization-free Decoding, a defense which can eliminate verbatim memorization in the generations from a large neural language model, but does not prevent approximate memorization.

1 Introduction

The ability of neural language models to memorize their training data has been studied extensively (Kandpal et al., 2022; Lee et al., 2021; Carlini et al., 2022; Zhang et al., 2021; Thakkar et al., 2021; Ramaswamy et al., 2020). When language models, especially ones used in production systems, are susceptible to *data extraction* attacks, it can lead to practical problems ranging from privacy risks to copyright concerns. For example, Carlini et al. (2021) showed that the GPT-2 language model could output personally identifying information of individuals contained in the training dataset.

One natural way to avoid this risk is to filter out any generations which copy long strings verbatim from the training set. GitHub's Copilot, a language-model-based code assistant, deploys this defense by giving users the option to "block suggestions matching public code" (GitHub, 2022).

In this work, we ask the question: "*Do language models emit paraphrased memorized content?*" This scenario can happen maliciously (e.g., adversaries trying to extract private user data) or through honest interactions (e.g., users prompting in real-world scenarios). Indeed, we find that Copilot's filtering system is easy to circumvent by applying plausible "style transfers" to the prompt. For example, by translating variable names from English to French the model outputs completely memorized examples, but post-processed with the en-fr style transfer. We further show that GPT-3 (Brown et al., 2020), a model trained on natural language, is also vulnerable to extraction attacks.

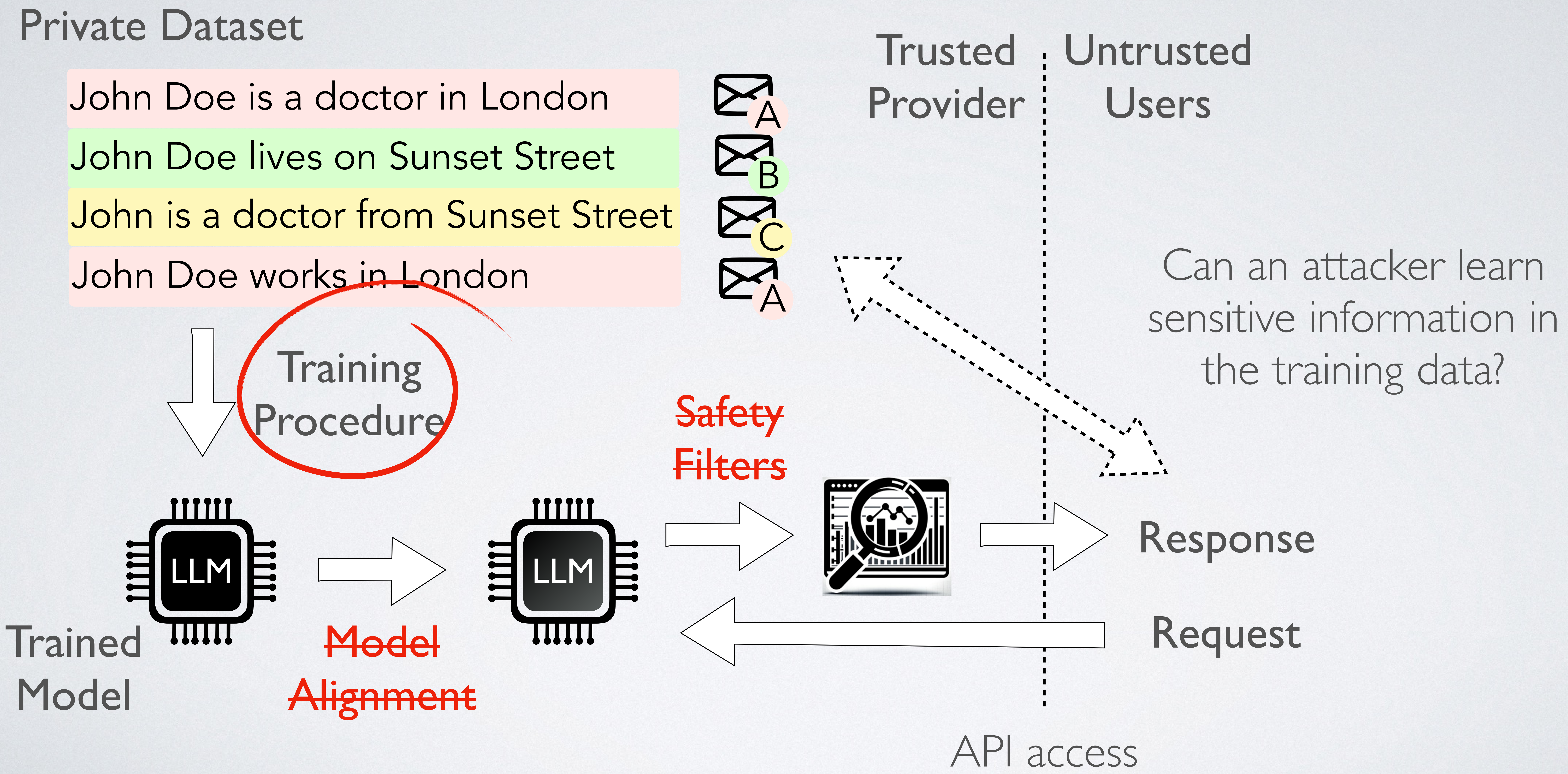
Unfortunately, Copilot's training set and precise algorithm for their defense are non-public. Therefore, to investigate this phenomenon systematically, we develop MEMFREE decoding (Figure 1), an efficient defense that is guaranteed to prevent all verbatim memorization, and which scales to training sets consisting of hundreds of gigabytes of text. In

*Remaining authors ordered by Algorithm 18 in Appendix H; briefly, we require Daphne be listed first, and Nicholas listed last, and we search for the first permutation of authors' first names which satisfies these constraints, where permutations order names by their salted MD5 hash.

Proceedings of the 16th International Natural Language Generation Conference, pages 28–53
September 11–15, 2023. ©2023 Association for Computational Linguistics

[F] Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy, Nasr et al., Preprint, November 2023

Data Privacy for Large Language Models



Public Data and Private Information

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

Privacy in Language Models [G]

- 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**)

[G] What Does it Mean for a Language Model to Preserve Privacy?, Brown et al., February 2022

Base Model vs Fine-Tuning

Promises of Fine-Tuning [8]

- Improve Quality
- Steer Model
- Shorter Prompts
- Lower latency

Fine-tuning models

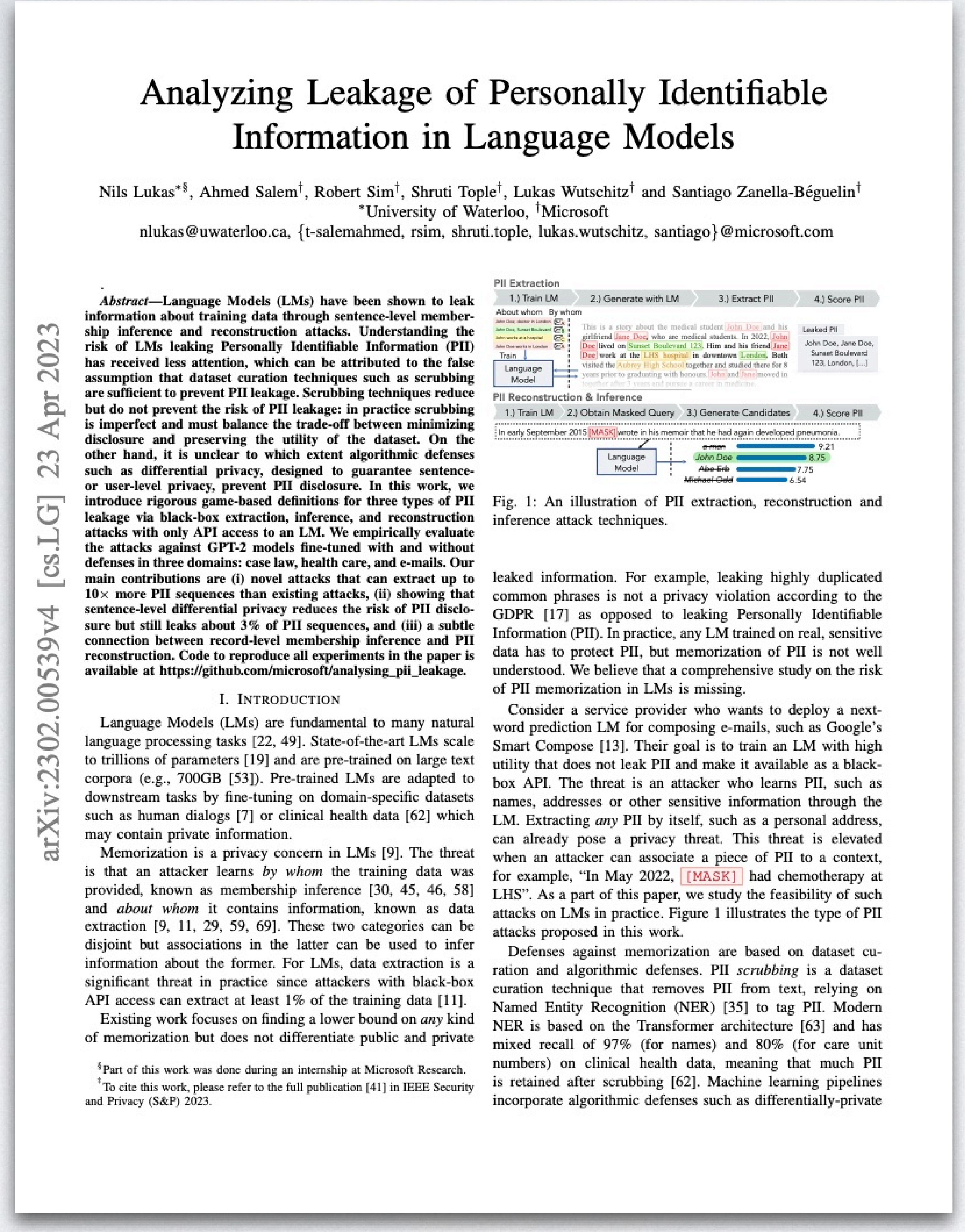
Create your own custom models by fine-tuning our base models with your training data. Once you fine-tune a model, you'll be billed only for the tokens you use in requests to that model.

Learn about fine-tuning ↗

Model	Training	Input usage	Output usage
gpt-3.5-turbo	$0.0080 / 1K tokens	$0.0030 / 1K tokens	$0.0060 / 1K tokens
davinci-002	$0.0060 / 1K tokens	$0.0120 / 1K tokens	$0.0120 / 1K tokens
babbage-002	$0.0004 / 1K tokens	$0.0016 / 1K tokens	$0.0016 / 1K tokens

OpenAI Pricing [7]

Focus of this Talk


arXiv:2302.00539v4 [cs.LG] 23 Apr 2023

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× 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 [19] 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, 45, 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 [11].

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

§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.

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 [17] 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 [13]. 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) [35] 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 [62]. Machine learning pipelines incorporate algorithmic defenses such as differentially-private

Nils Lukas

Ahmed Salem

Robert Sim

Shruti Tople

Lukas Wutschitz

Santiago Zanella-Béguelin

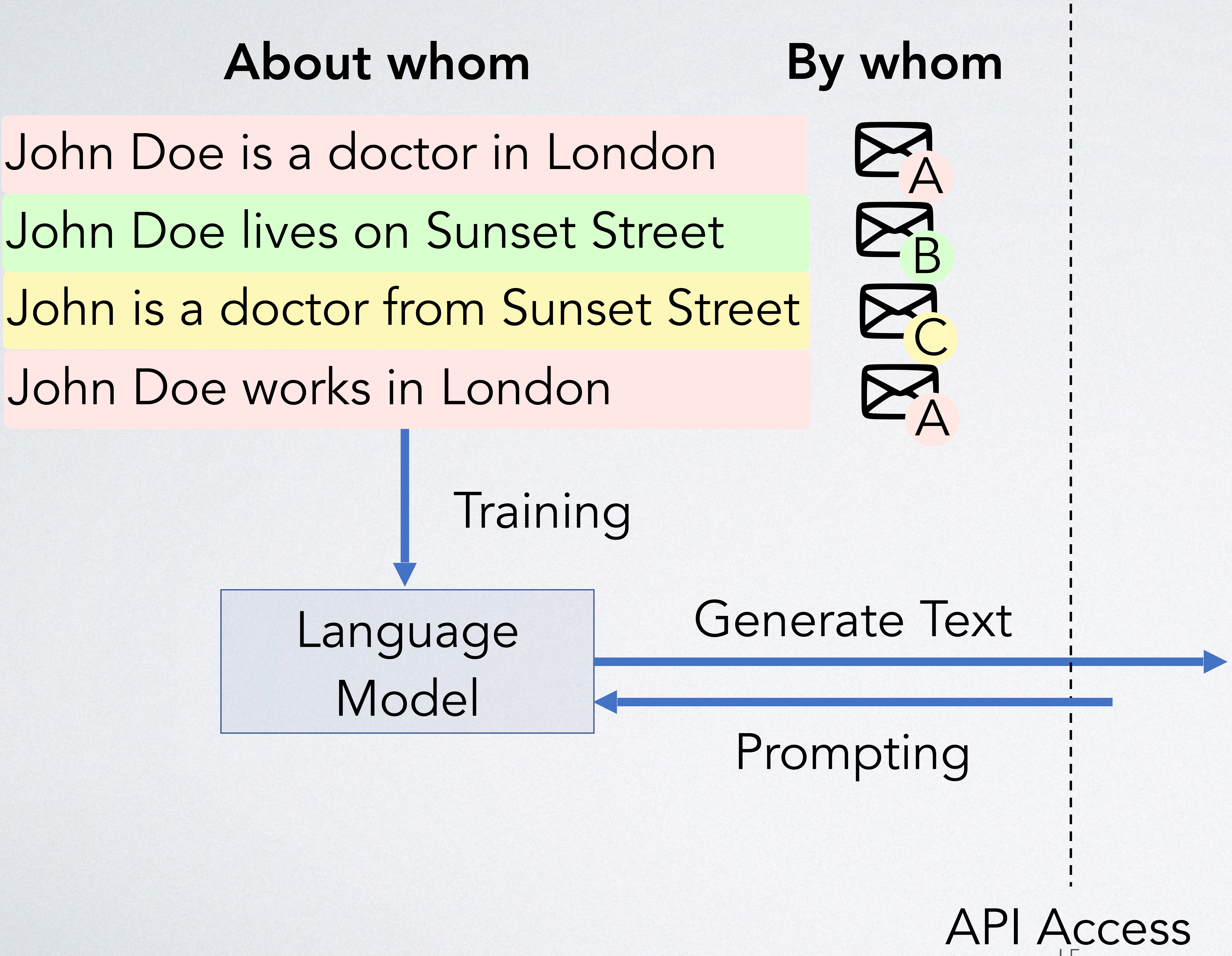
[H] Analyzing Leakage of Personally Identifiable Information in Language Models, Lukas et al., February 2023

[I] GitHub Repository

UNIVERSITY OF WATERLOO

Microsoft

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

John Doe London
Sunset Street
LHS Hospital

Real

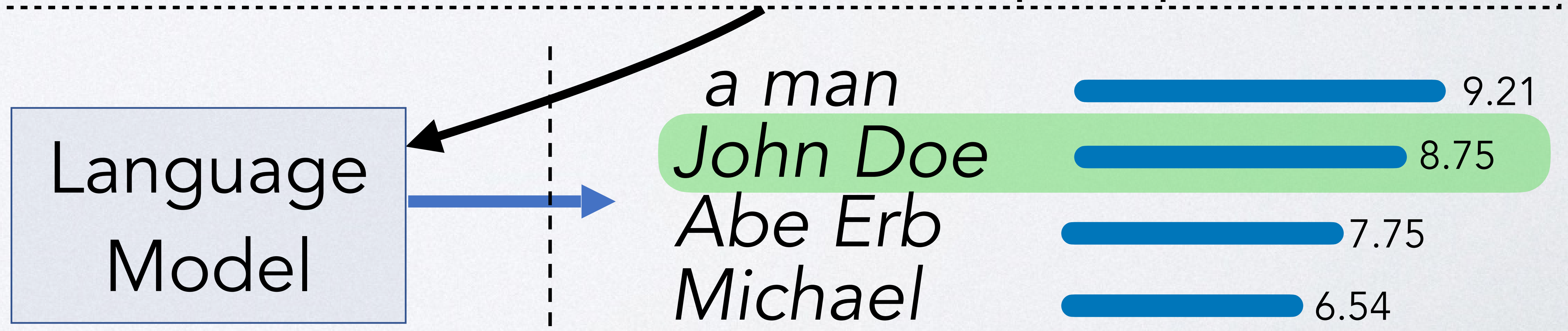
Jane Doe
Aubrey High School

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

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

Real Sentence

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

Language Model

a man	9.21
John Doe	8.75
Abe Erb	7.75
Teo Peric	6.54

Reconstruction

~~a man~~	9.21
John Doe	8.75
Abe Erb	7.75
Teo Peric	6.54

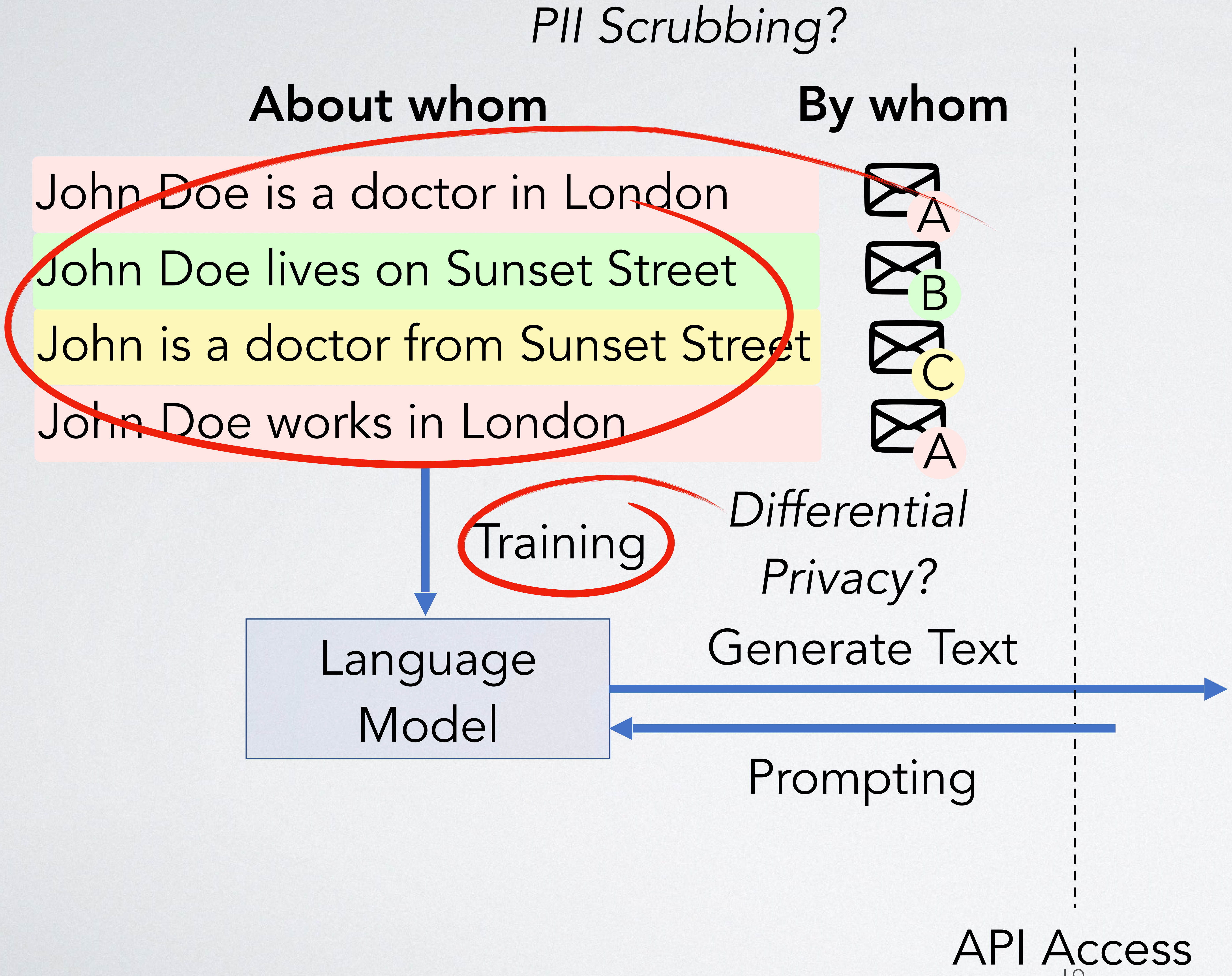
Inference

John Doe, Teo Peric

PII Candidates

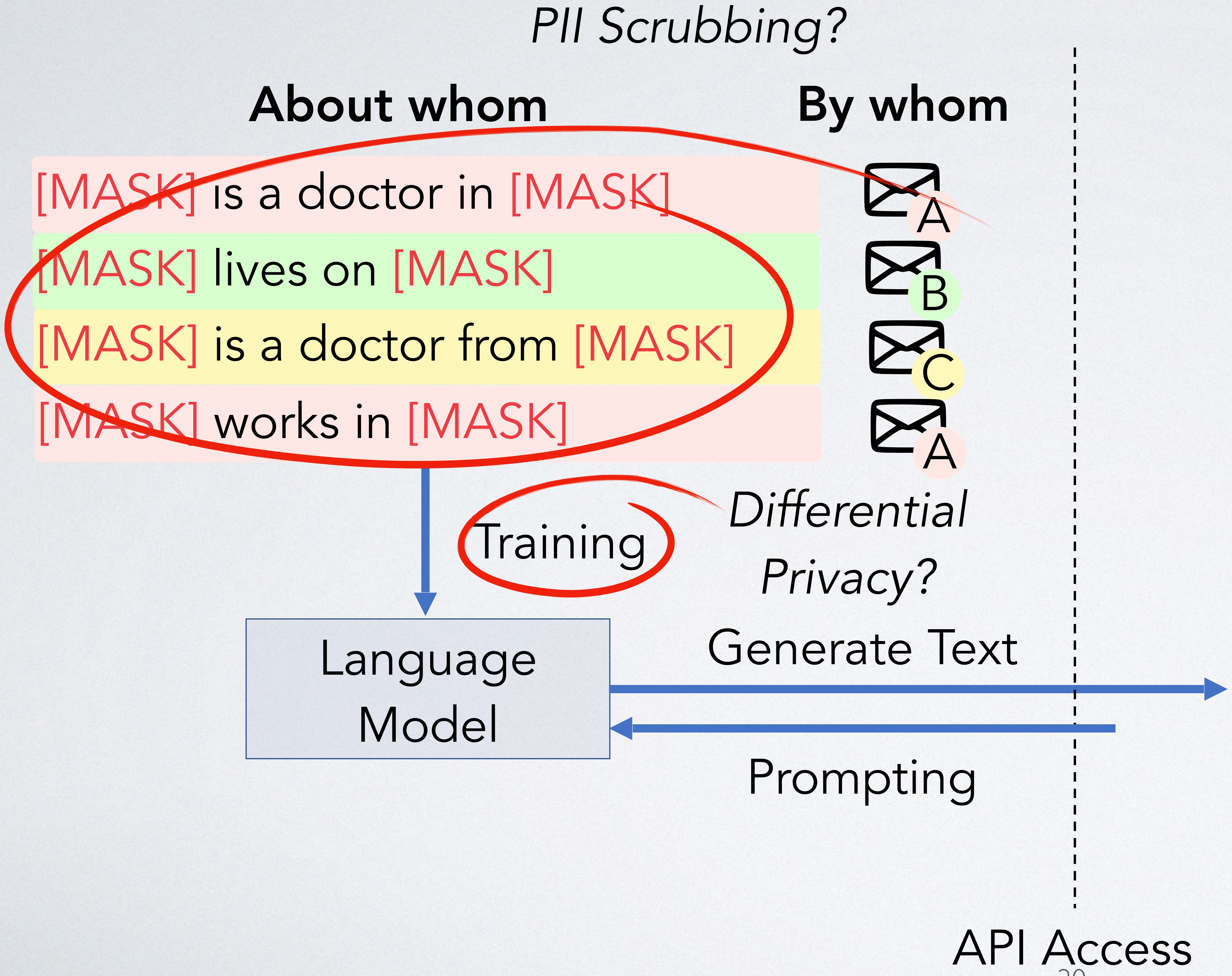
~~a man~~	9.21
John Doe	8.75
~~Abe Erb~~	7.75
~~Teo Peric~~	6.54

Motivation



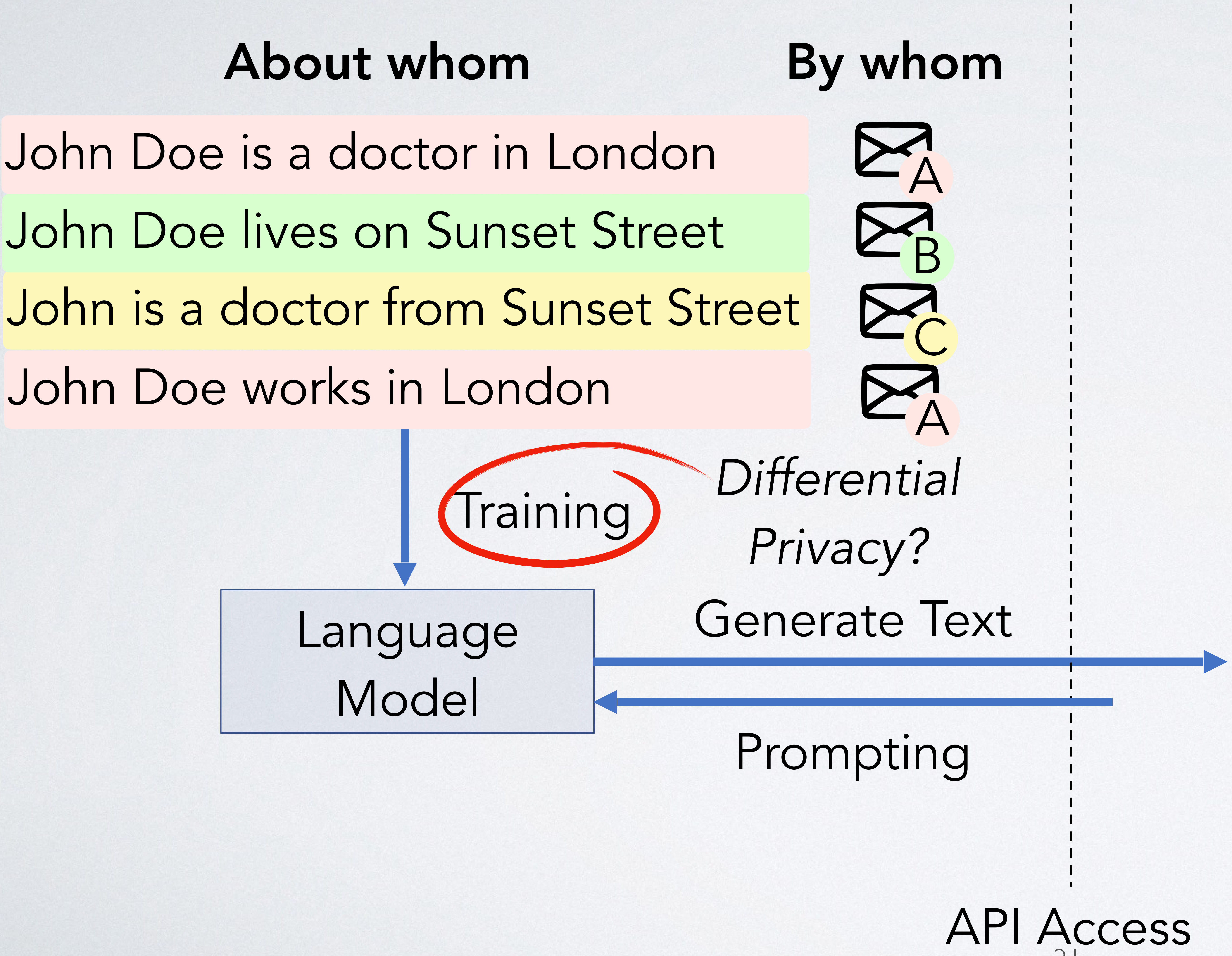
Once upon a time, there existed a tale of medical students John Doe and his girlfriend, Jane Doe. In the year 2022, John resided at Sunset Street while pursuing his medical education. Alongside his friend Jane, he worked at the LHS Hospital located in the bustling heart of downtown London. Before donning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



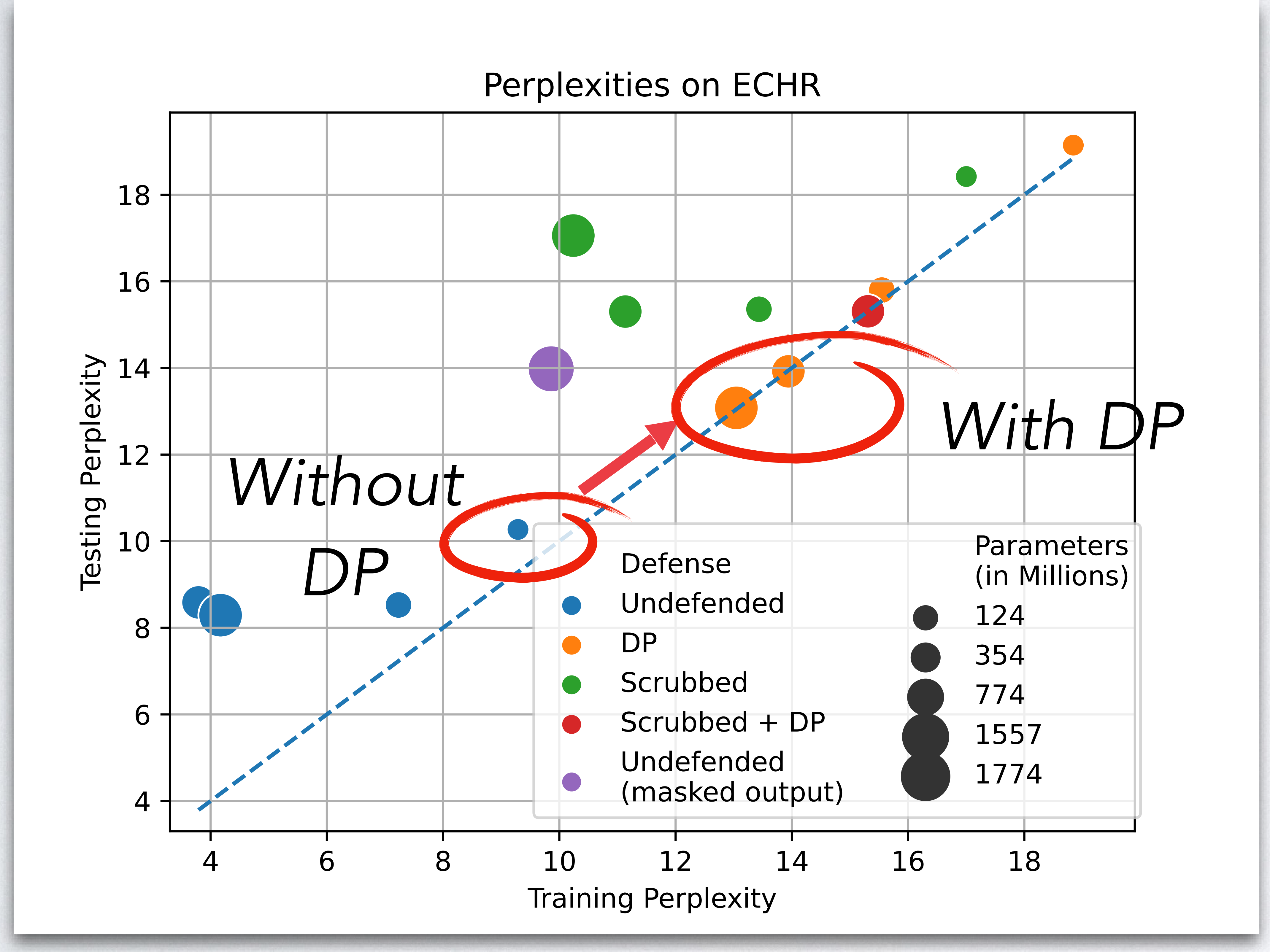
Once upon a time, there existed a tale of medical students John Doe and his girlfriend, Jane Doe. In the year 2022, John resided at Sunset Street while pursuing his medical education. Alongside his friend Jane, he worked at the LHS Hospital located in the bustling heart of downtown London. Before donning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

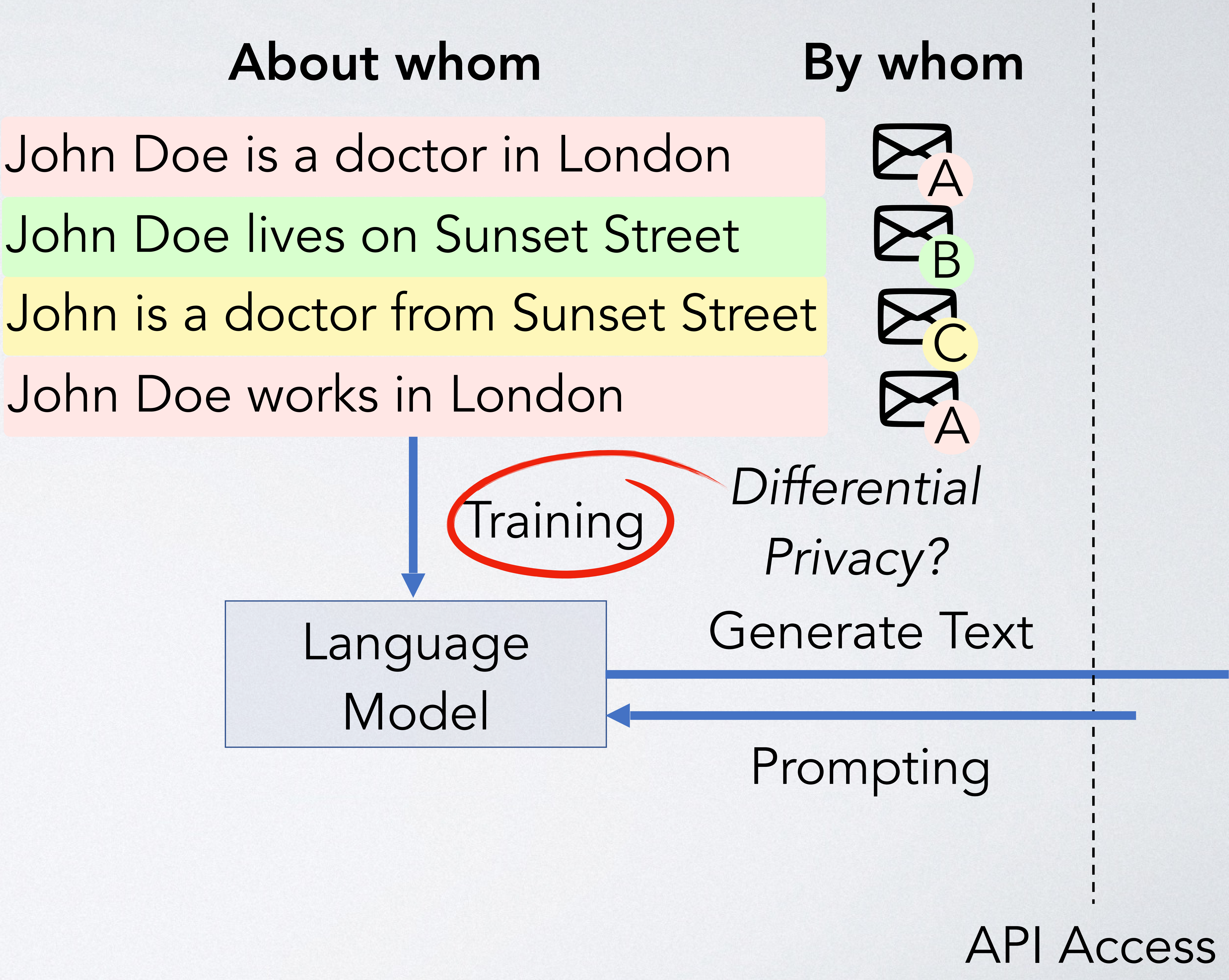


Once upon a time, there existed a tale of medical students John Doe and his girlfriend, Jane Doe. In the year 2022, John resided at Sunset Street while pursuing his medical education. Alongside his friend Jane, he worked at the LHS Hospital located in the bustling heart of downtown London. Before donning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

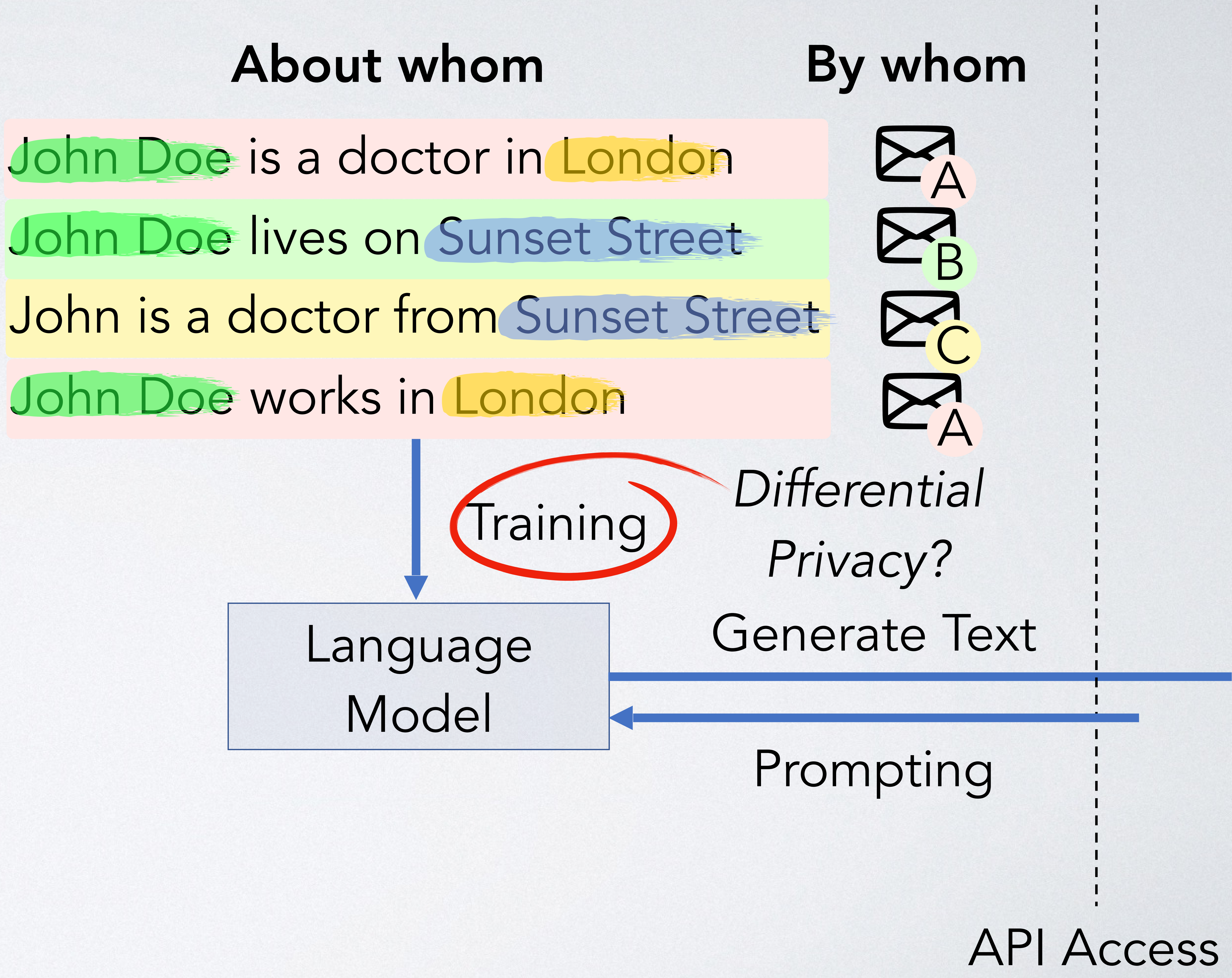


<u>Privacy</u> at the cost of <u>Model Utility</u>



Problems with Differential Privacy

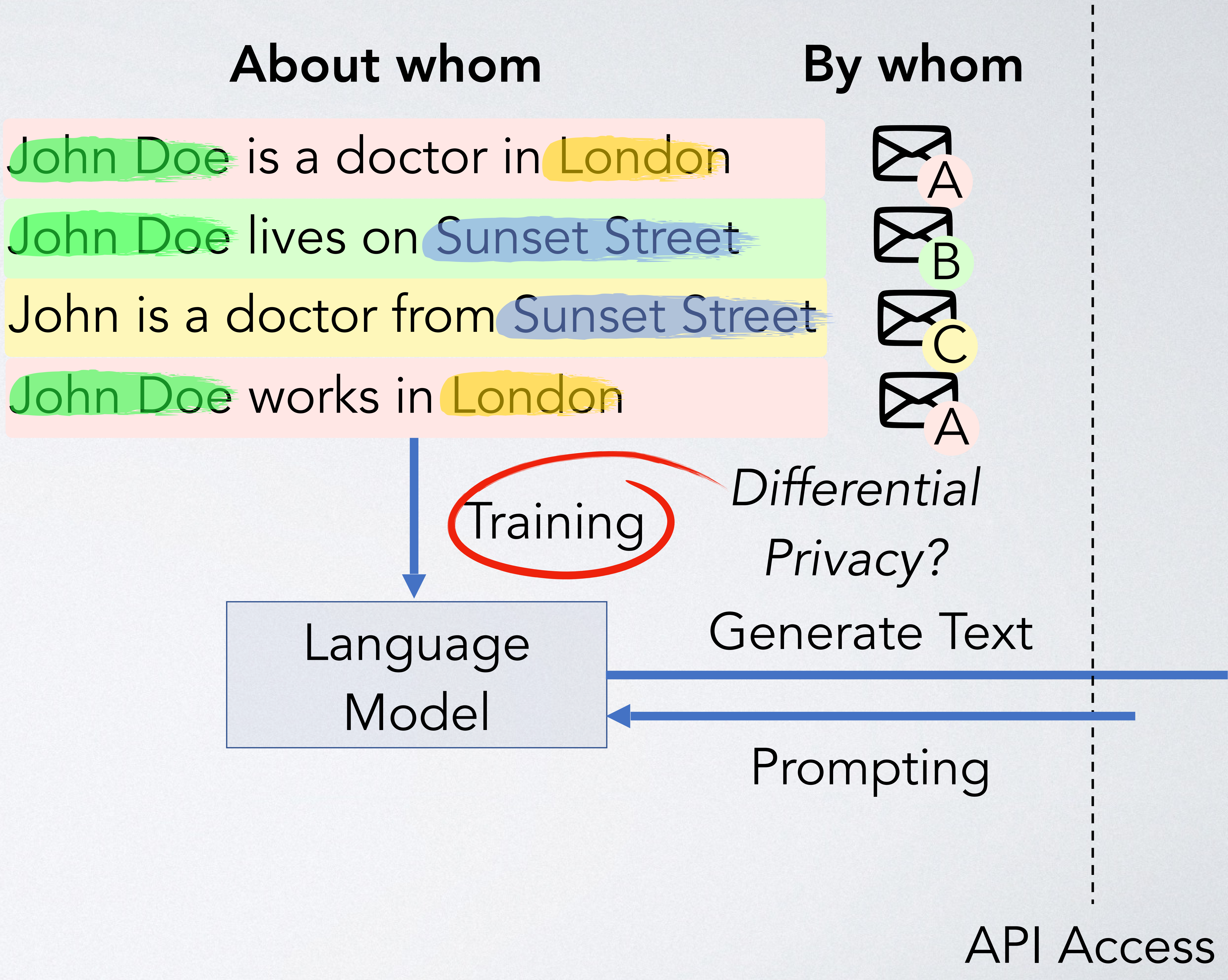
DP protects against an attacker learning **by whom** data was provided, but not **about whom** it contains information.



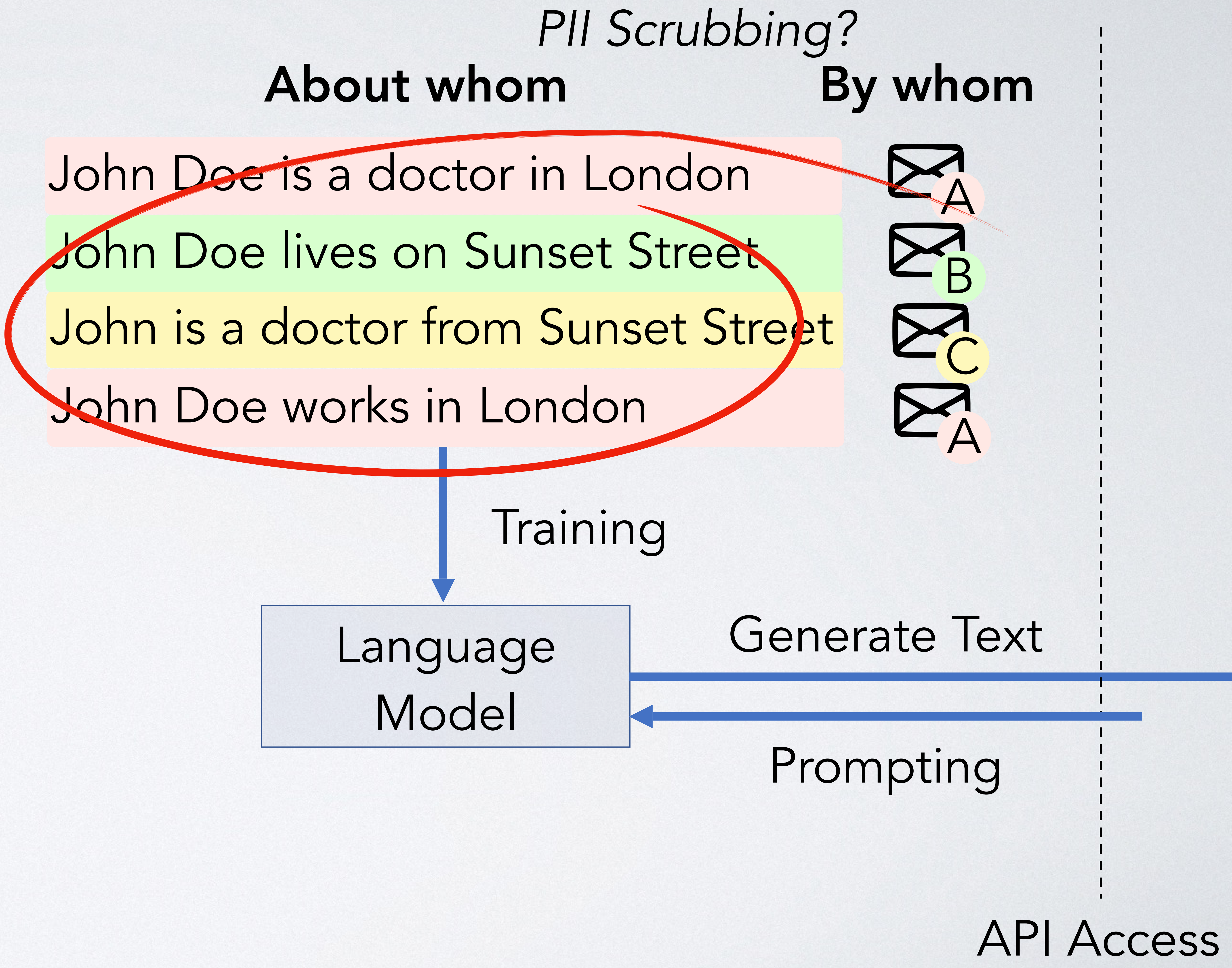
Problems with Differential Privacy

Group-level DP can help but ..

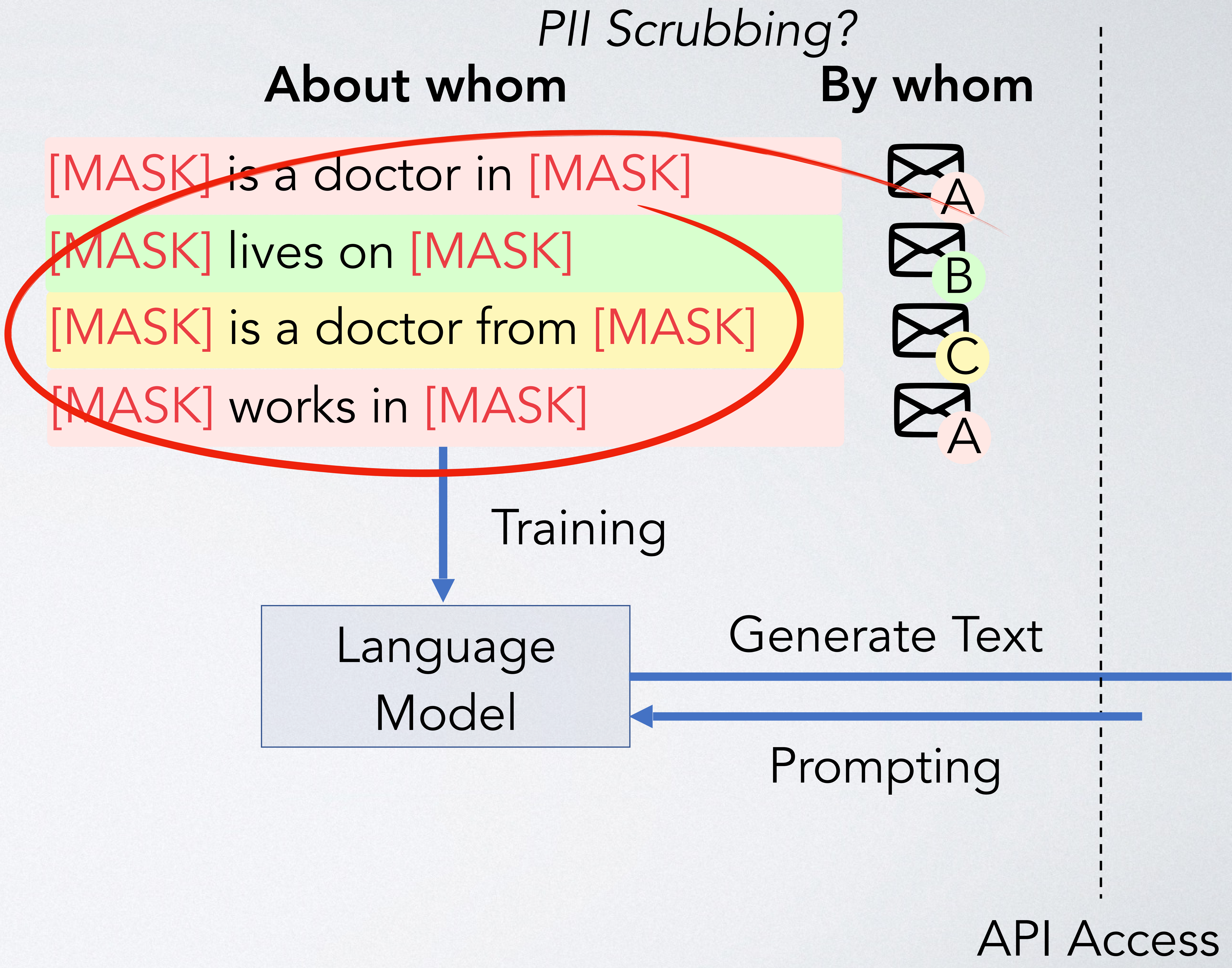
1) Group sizes are not always known a priori and under worst-case assumptions has deleterious impact on model utility.
2) PII Duplication across groups



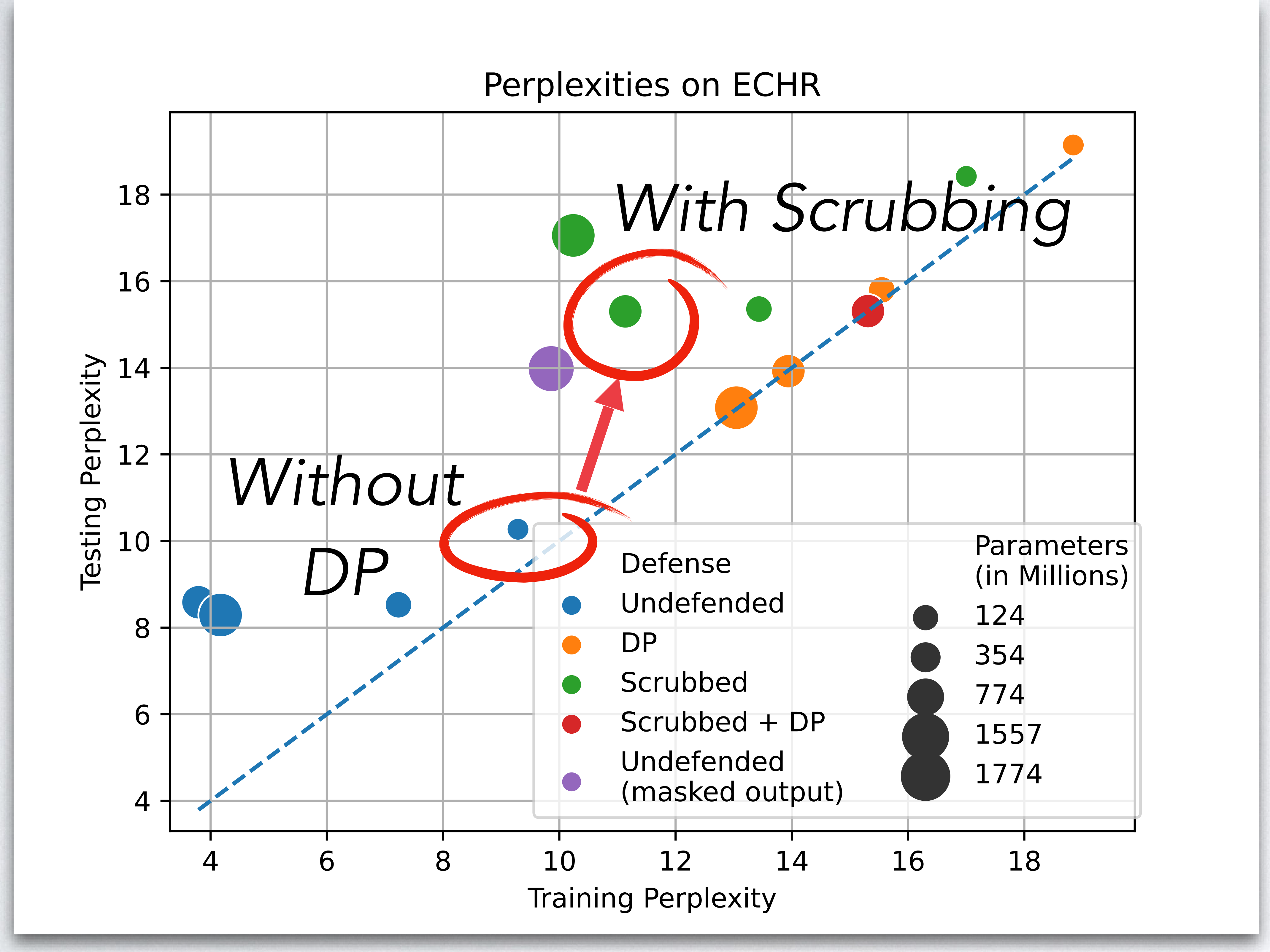
Problems with PII Scrubbing



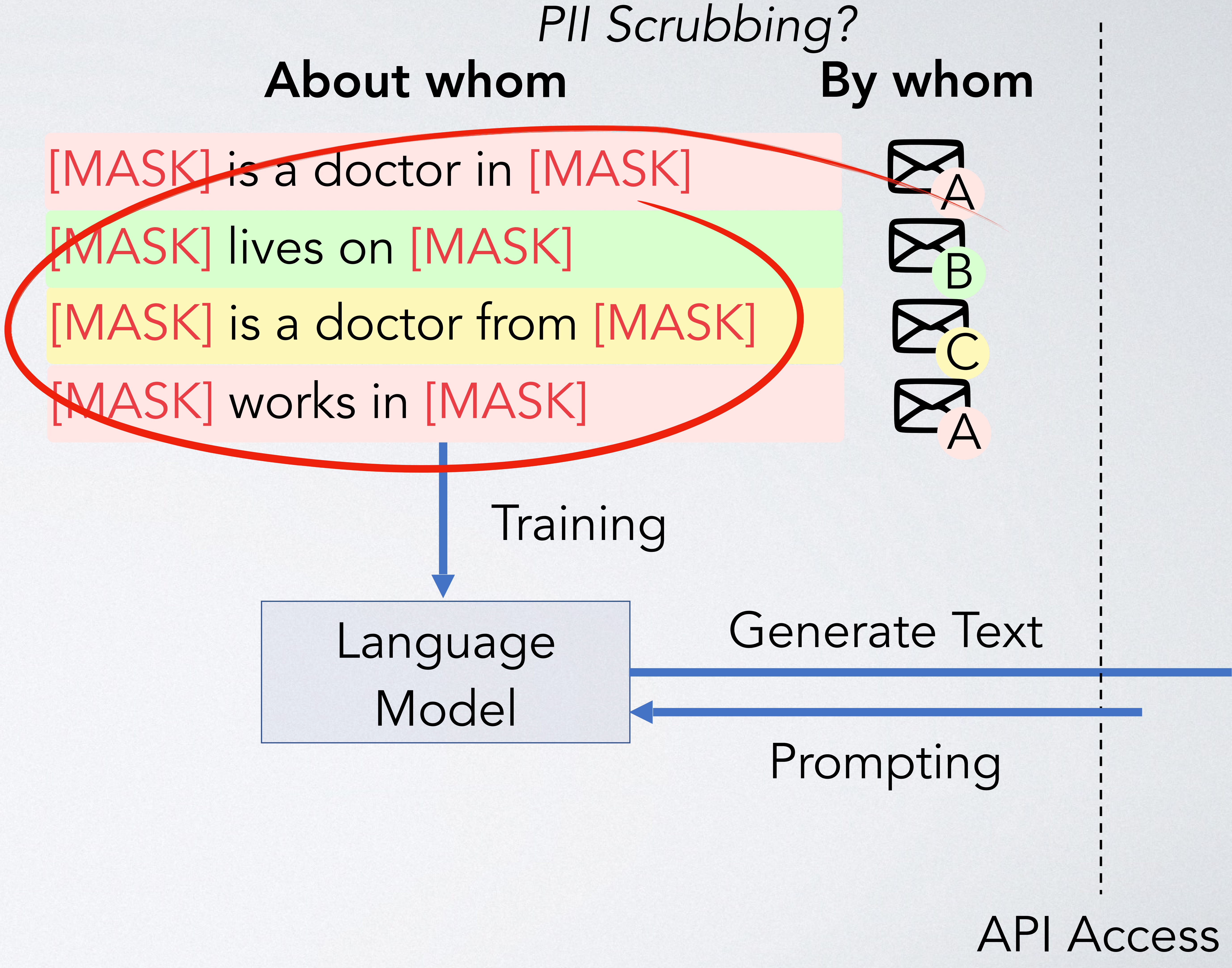
Problems with PII Scrubbing



Problems with PII Scrubbing

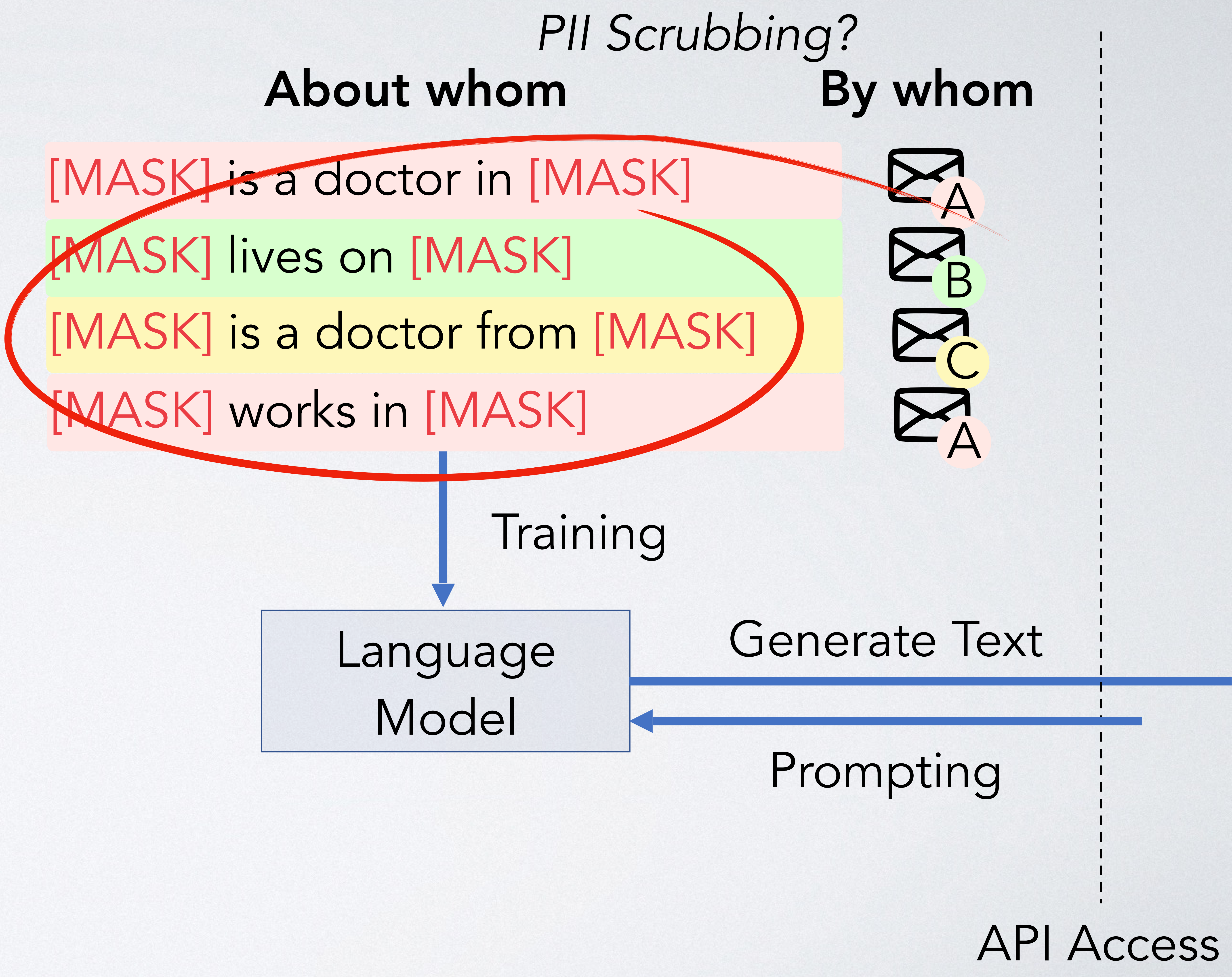


Privacy at the cost of Model Utility



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

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

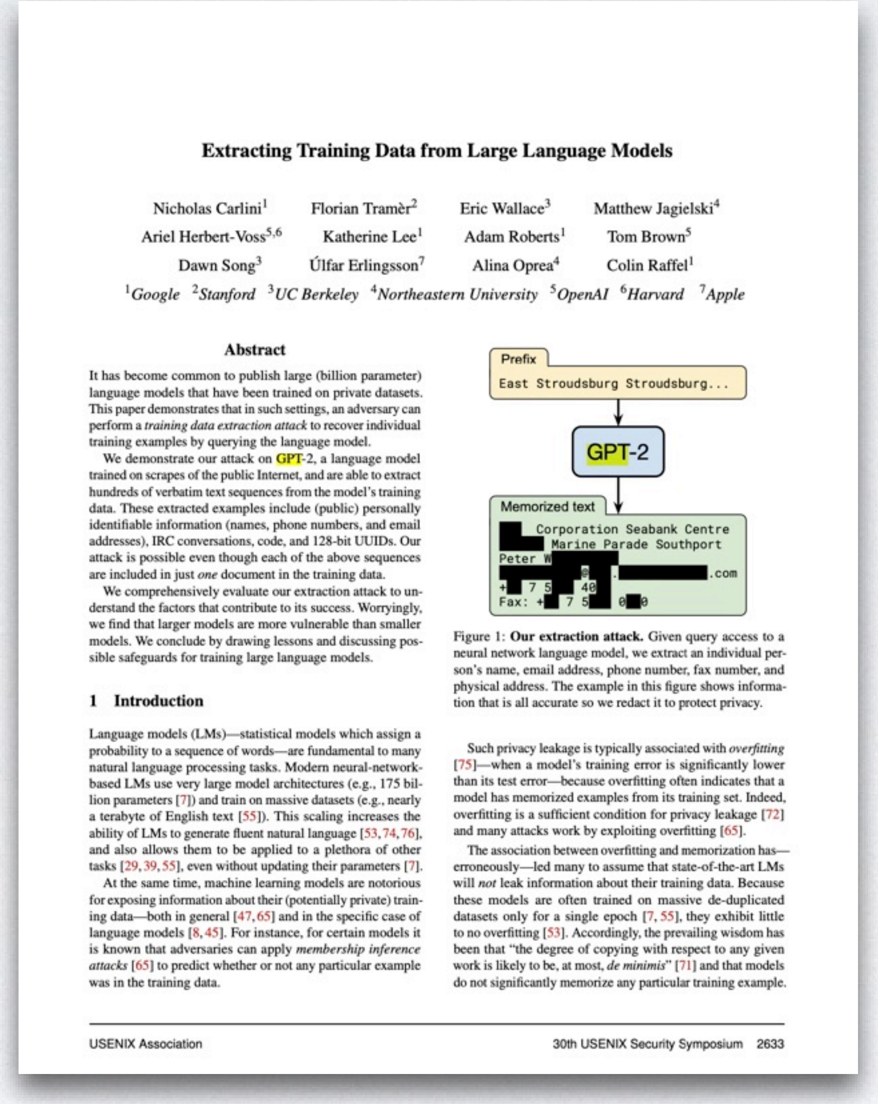
Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, Dawn Song

Carlini et al., 2019

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

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, Asli Celikyilmaz

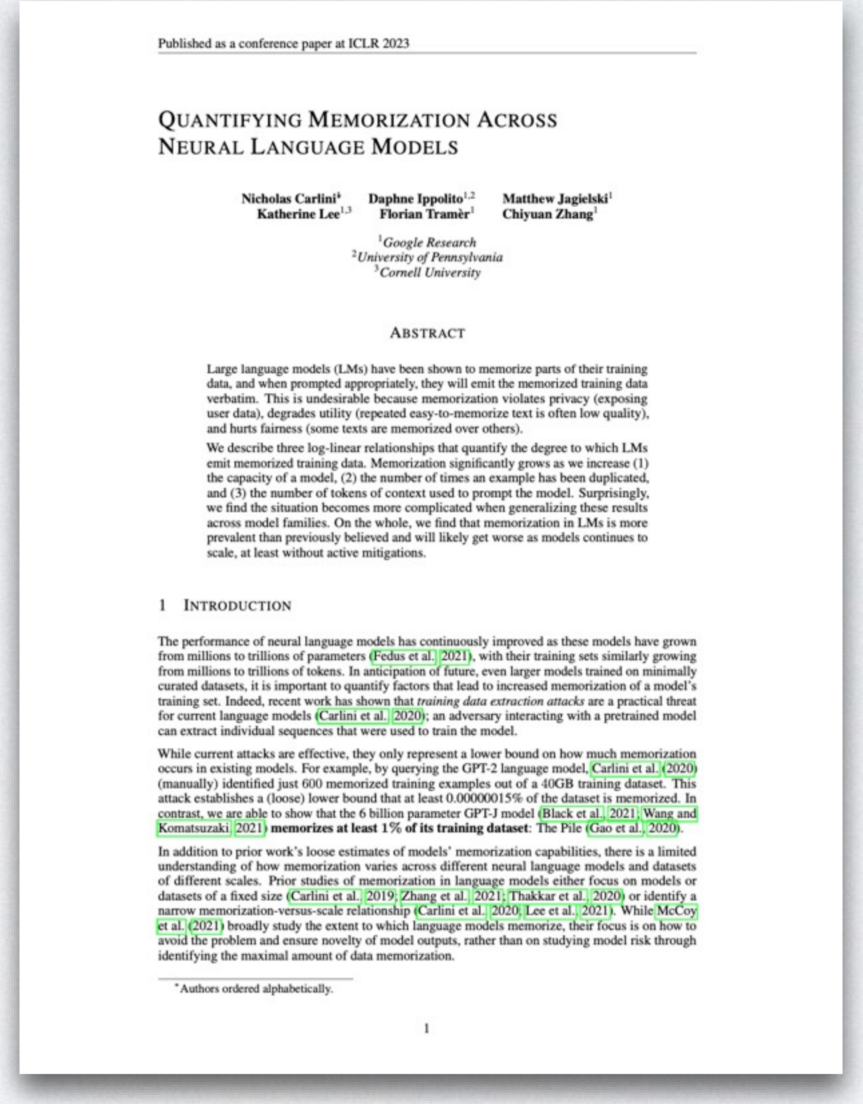
McCoy et al., 2019



Extracting Training Data from Large Language Models

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, Colin Raffel

Carlini et al., 2020



Quantifying Memorization Across Neural Language Models

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, Chiyuan Zhang

Carlini et al., 2022

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

Jie Huang, Hanyin Shao, Kevin Chen-Chuan Chang

Huang et al., 2022

Any Form of Leakage

Our Focus

We study PII leakage in the presence of privacy mechanisms such as **Differential Privacy** or **PII Scrubbing**

Extraction

- Black-box Model Access

Reconstruction

- Black-box Model Access
- Masked Training Data

Inference

- Black-box Model Access
- Masked Training Data
- Auxiliary Information

Is differential privacy alone sufficient to protect PII?

Security Games for PII Leakage

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

experiment IND-INFERENCE($\mathcal{T}, \mathcal{D}, n, \mathcal{A}$)
  $b \sim \{0, 1\}$
  $D \sim \mathcal{D}^n$
  $\theta \leftarrow \mathcal{T}(D)$
  $S_0 \sim D$
  $S_1 \sim \mathcal{D}$
  $\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), S_b)$
  $S \sim \{S \in D | \text{EXTRACT}(S) \neq \emptyset\}$
  $C_0 \sim \text{EXTRACT}(S)$
  $C_1 \sim \mathcal{E}$
  $\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), \text{SCRUB}(\text{SPLIT}(S, C_0)), C_b)$

Algorithm 2 PII Extraction

experiment EXTRACTION($\mathcal{T}, \mathcal{D}, n, \mathcal{A}$)
  $D \sim \mathcal{D}^n$
  $\theta \leftarrow \mathcal{T}(D)$
  $\mathcal{C} \leftarrow \bigcup_{S \in D} \text{EXTRACT}(S)$
  $\tilde{\mathcal{C}} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), |\mathcal{C}|)$

procedure $\mathcal{O}_\theta(S)$
  return $\{w \mapsto \Pr(w|S;\theta)\}_{w \in \mathcal{V}}$

See our paper for more details

Algorithm 5 PII Reconstruction Game

experiment RECONSTRUCTION($\mathcal{T}, \mathcal{D}, n, \mathcal{A}$)
  $D \sim \mathcal{D}^n$
  $\theta \leftarrow \mathcal{T}(D)$
  $S \sim \{S \in D | \text{EXTRACT}(S) \neq \emptyset\}$
  $C \sim \text{EXTRACT}(S)$
  $\tilde{C} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), \text{SCRUB}(\text{SPLIT}(S, C)))$

Algorithm 7 PII Inference Game

experiment INFERENCE($\mathcal{T}, \mathcal{D}, n, m, \mathcal{A}$)
  $D \sim \mathcal{D}^n$
  $\theta \leftarrow \mathcal{T}(D)$
  $S \sim \{S \in D | \text{EXTRACT}(S) \neq \emptyset\}$
  $C \sim \text{EXTRACT}(S)$
  $\mathcal{C} \sim \mathcal{E}^m$
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{C\}$
  $\tilde{C} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}_\theta(\cdot), \text{SCRUB}(\text{SPLIT}(S, C)), \mathcal{C})$

Setup

Training Dataset

John Doe is a doctor in London — A

John Doe lives on Sunset Street — B

Testing Dataset

John is a doctor from Sunset Street — C

John Doe works in London — A

Training Procedure

1. No Defense
2. DP
3. Scrubbing
4. DP & Scrubbing

1. Small
2. Medium
3. Large
4. XL

1. Enron
2. Yelp-Health
3. ECHR

LLM

PII Extraction

PII Reconstruction

[MASK] lives on Sunset Street — B

PII Inference

[MASK] lives on Sunset Street — B

John Doe or Joe Peric

Membership Inference

Datasets with many Detectable PII

	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

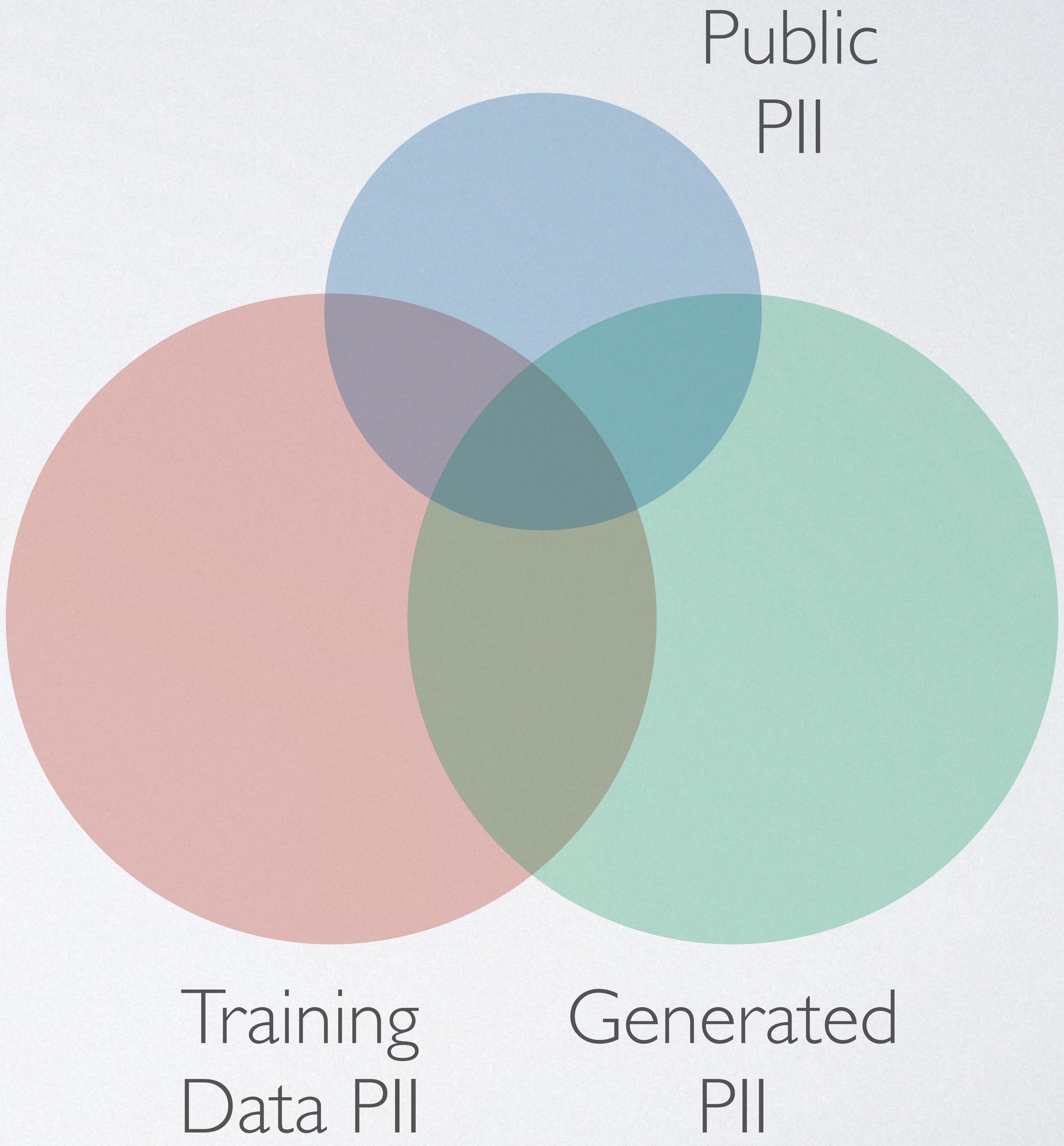
Extraction Attack

Goal: Extract PII from Training data with no auxiliary information

1. Generate N sequences with the model

2. Tag PII generated by the model

3. Calculate Precision & Recall



Reconstruction Attack

🏆 Goal: Reconstruct PII given a masked sentence From the training data

Real Sentence

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

Naïve Attack

Naïve attack ignores the suffix

Prompt

In early September 2023

Language Model

Generated

a group of people went to a conference.

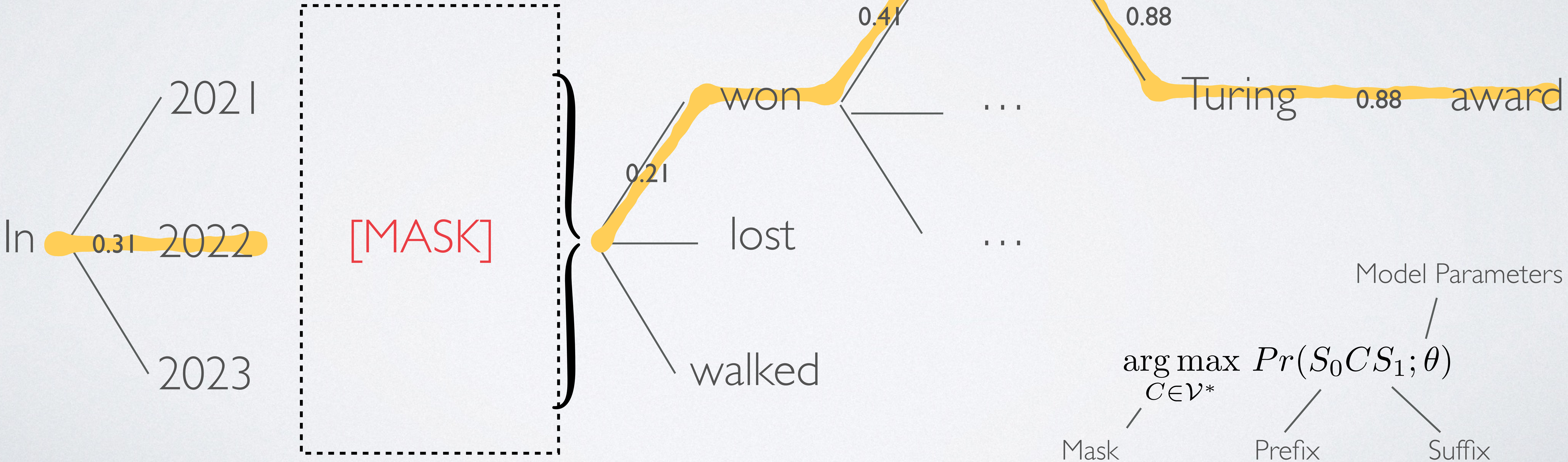
Reconstruction Attack Intuition

🏆 Goal: Reconstruct PII given a masked sentence From the training data

Real Sentence

In 2022 [MASK] won the Turing award.

- Unknown # tokens
- Intractable, must approximate

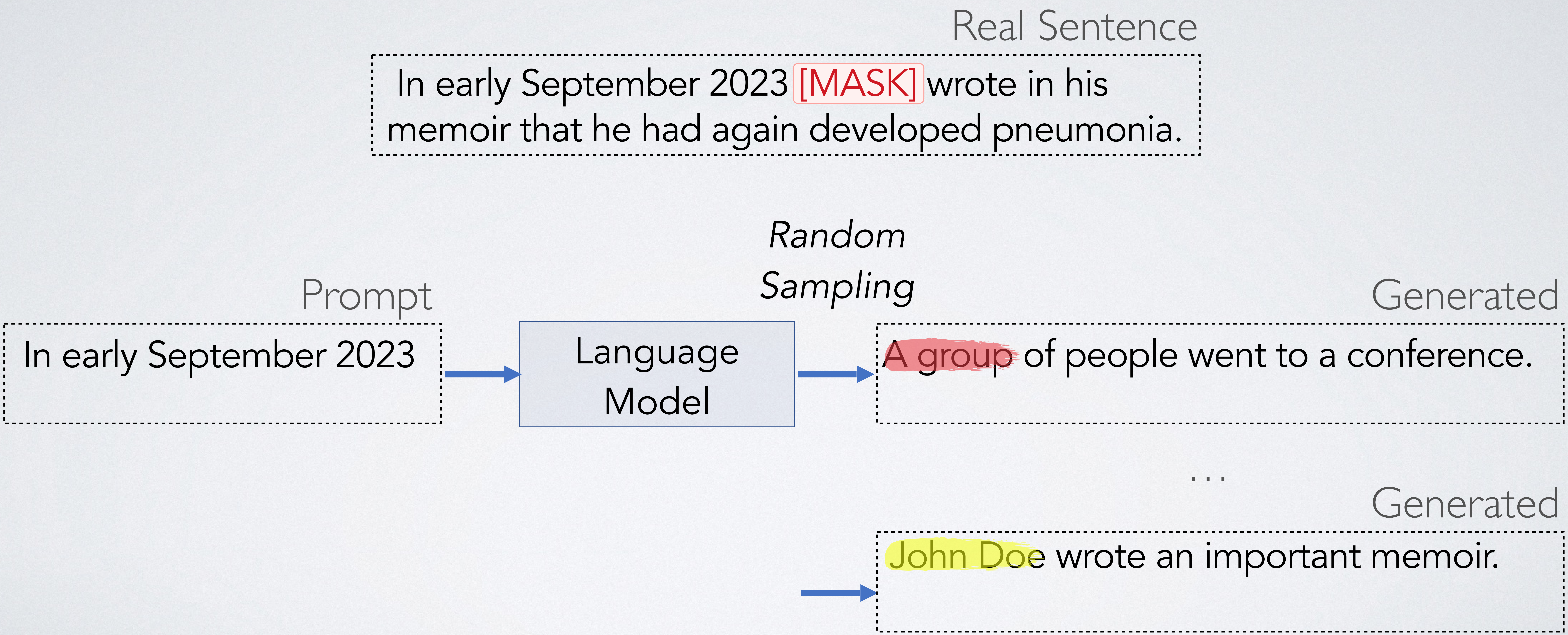


Model Parameters

$$\arg\max_{C \in \mathcal{V}^*} Pr(S_0 C S_1; \theta)$$

Mask Prefix Suffix

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.

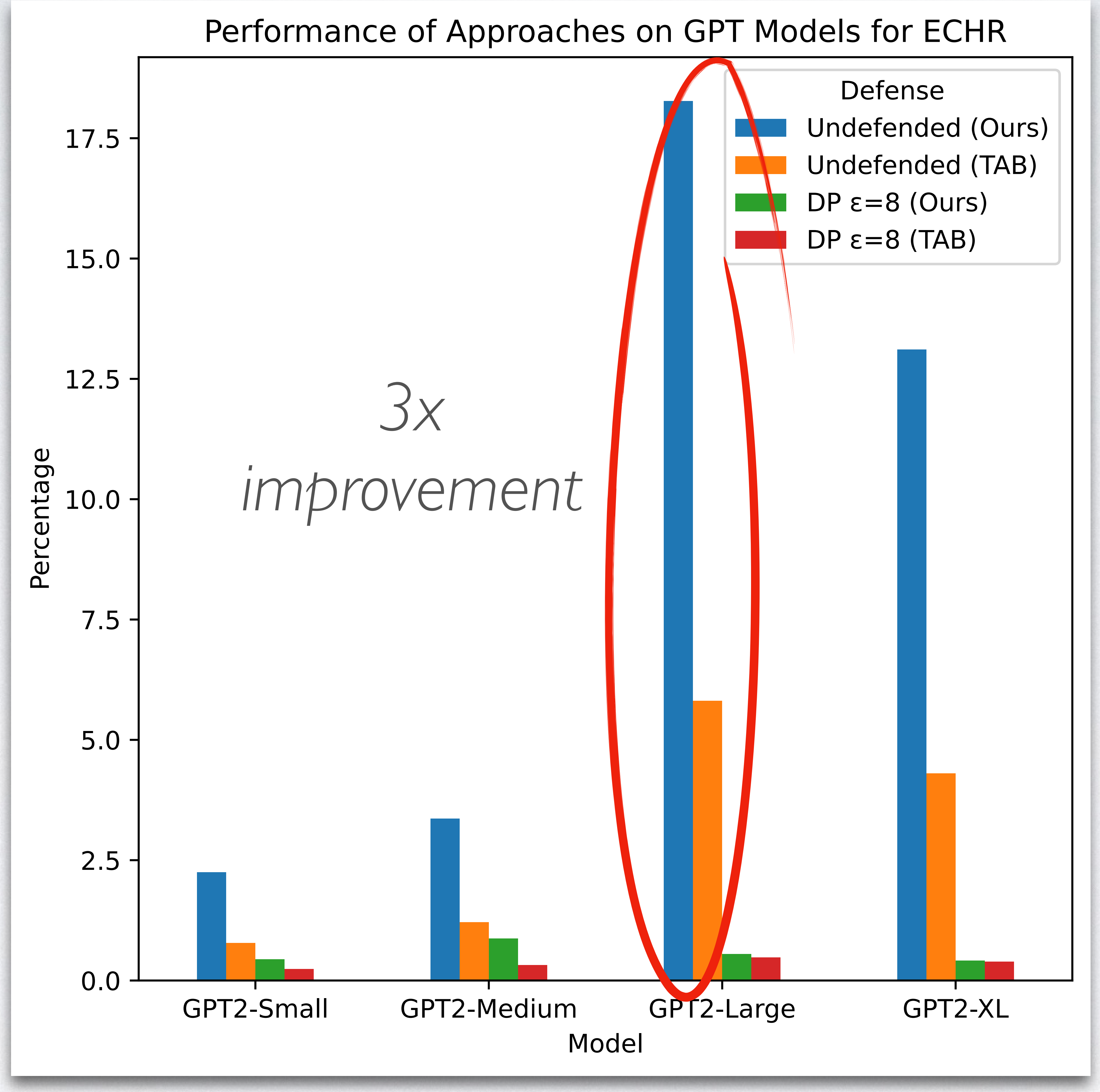
Tag PII

Tag PII & Construct Candidate Set

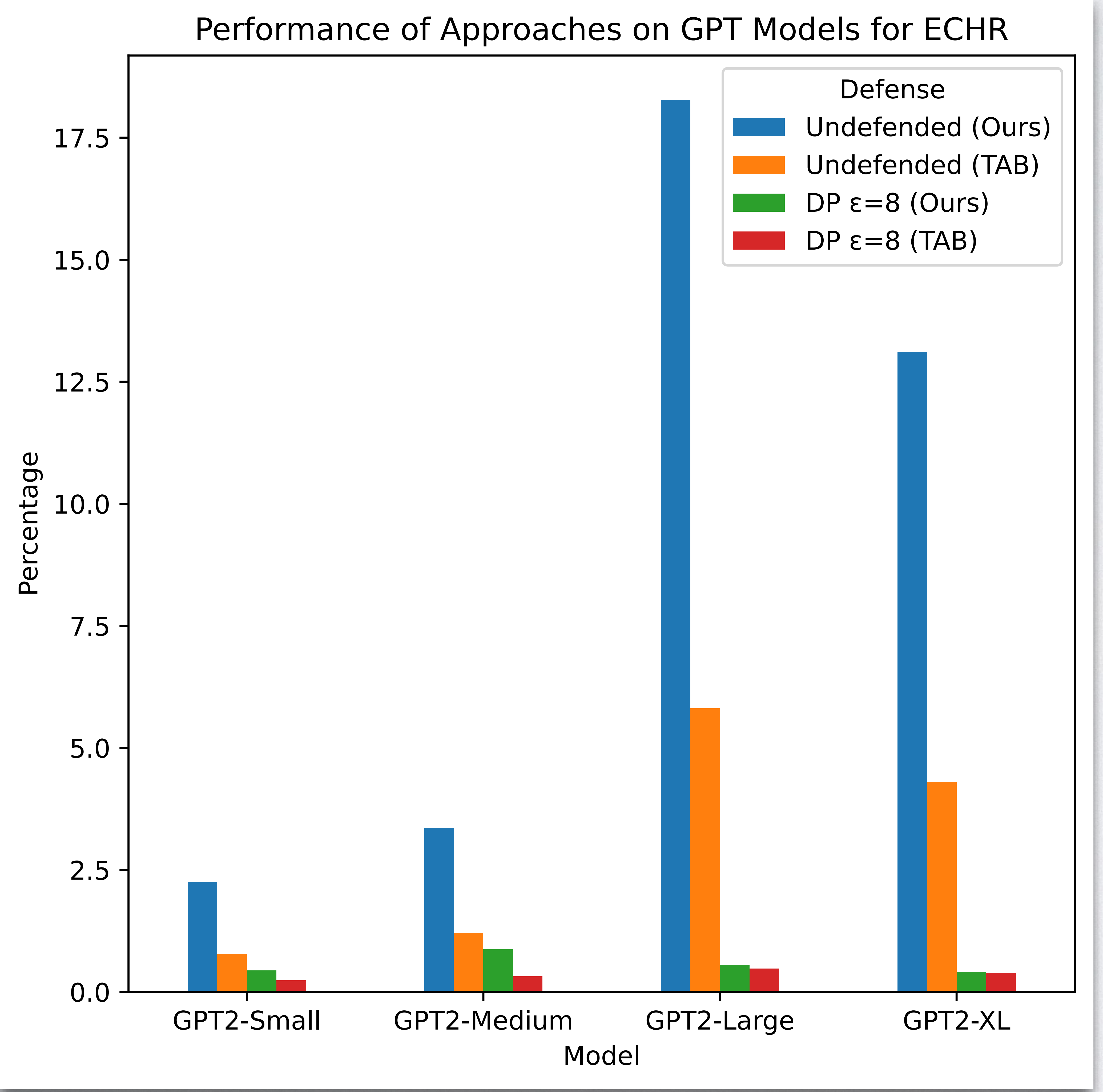
John Doe,
Jane Doe
Teo Peric

Reconstruction Attack

PII Reconstruction



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, $\|\mathcal{C}\| = 64$)	**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

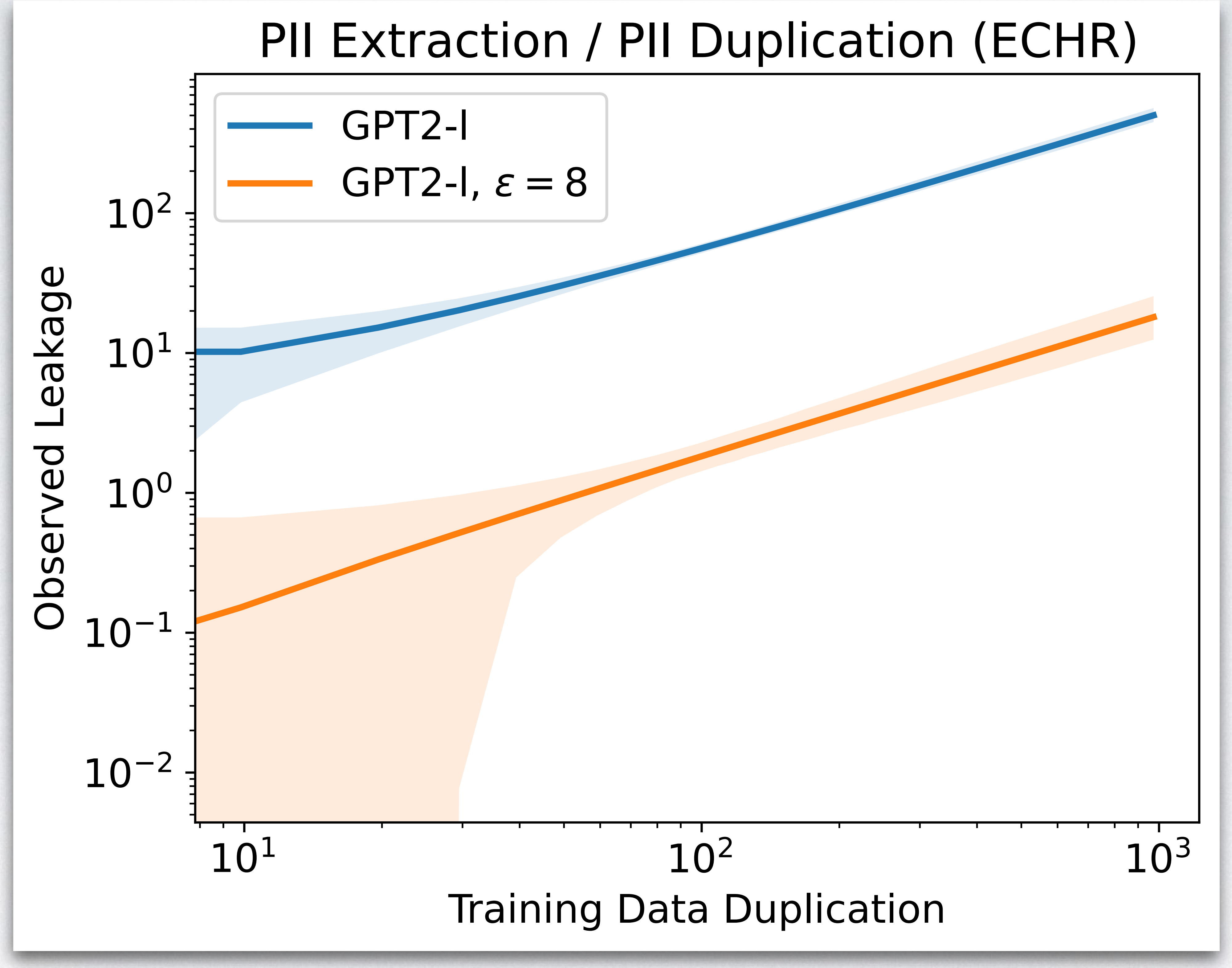
🏆 Goal: Infer PII given a masked sentence from the training data a set of PII candidates

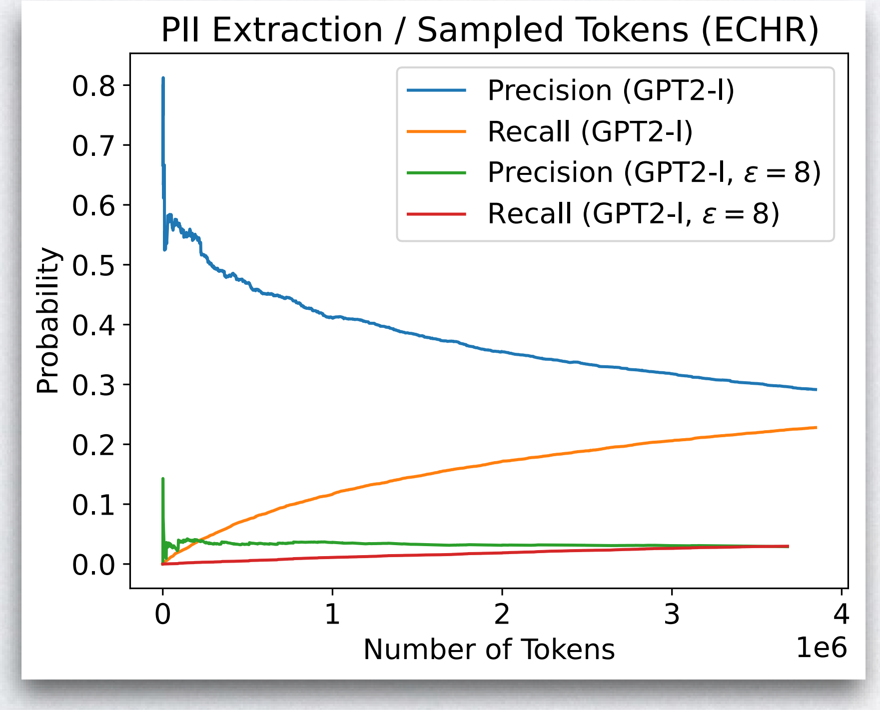
	ECHR		Enron		Yelp-Health	
	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$
$\|\mathcal{C}\| = 100$	70.11%	8.32%	50.50%	3.78%	28.31%	4.29%
$\|\mathcal{C}\| = 500$	51.03%	3.71%	34.14%	1.92%	15.55%	1.86%

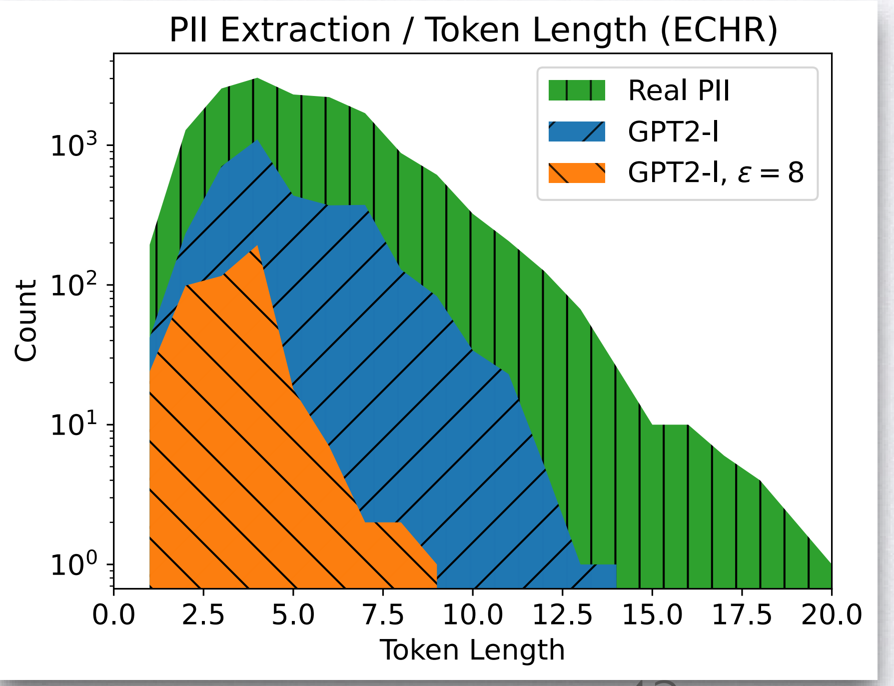
PII Extraction

Duplicated PII are Extractable more often

(Linear scaling)



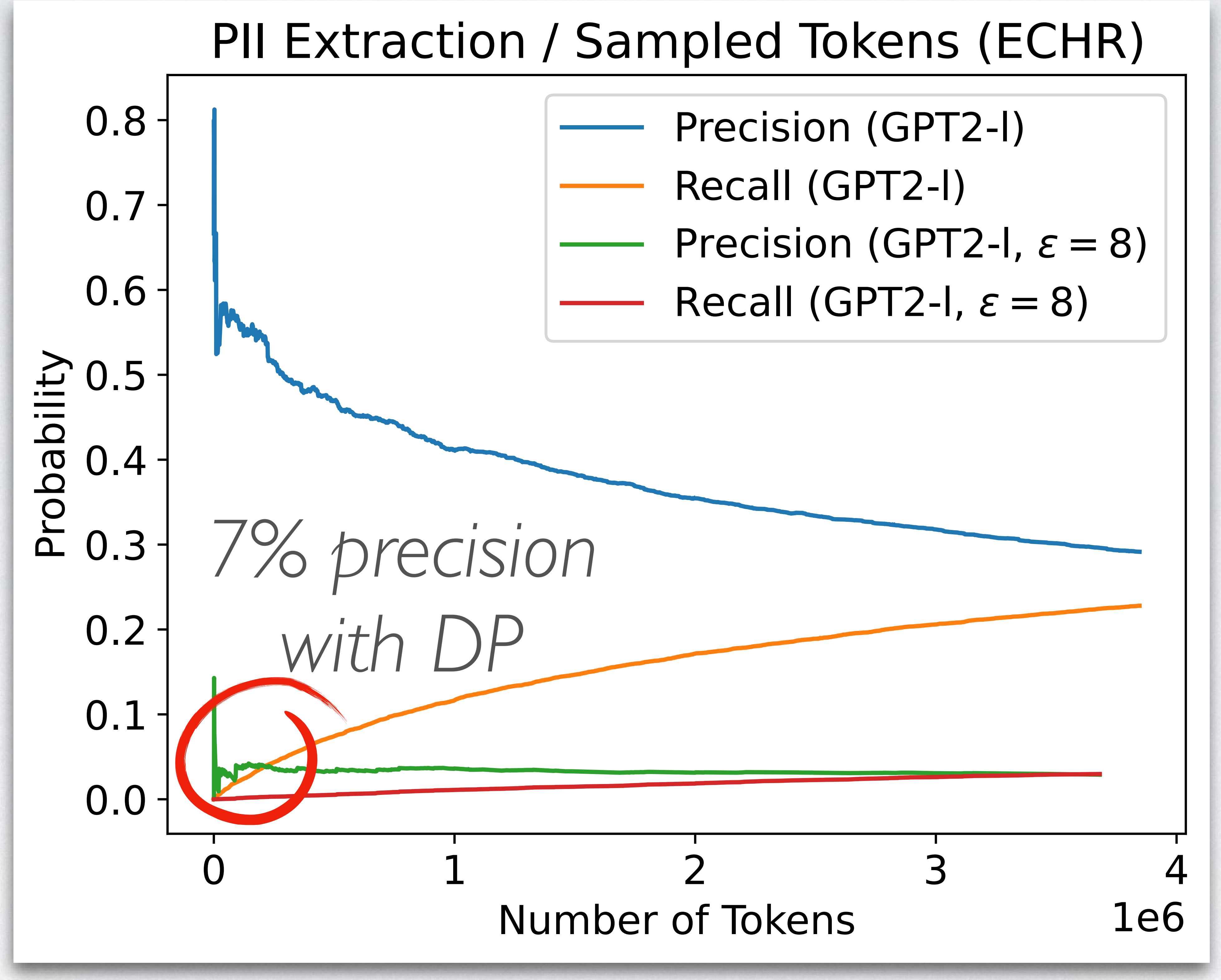


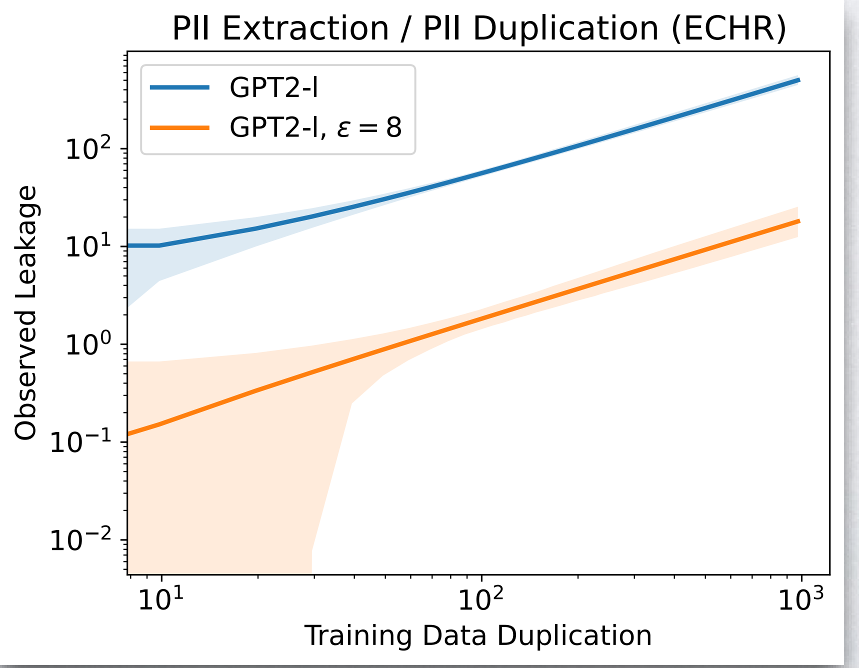


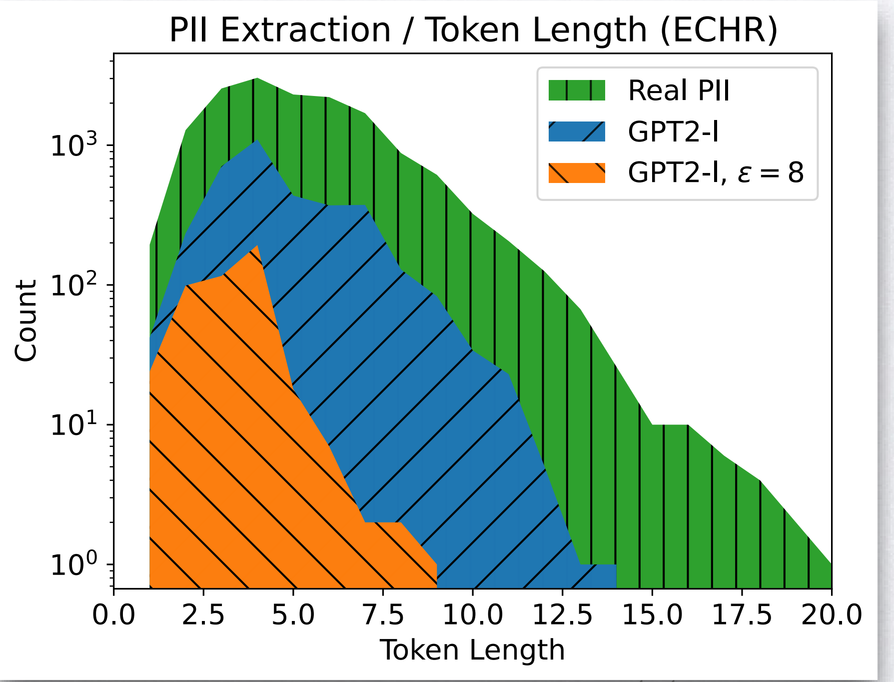
	GPT2-Small		GPT2-Medium		GPT2-Large	
	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$
ECHR						
Prec	24.91%	2.90%	28.05%	3.02%	29.56%	2.92%
Recall	9.44%	2.98%	12.97%	3.21%	22.96%	2.98%
Enron						
Prec	33.86 %	9.37%	27.06%	12.05%	35.36%	11.57%
Recall	6.26%	2.29%	6.56%	2.07%	7.23%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



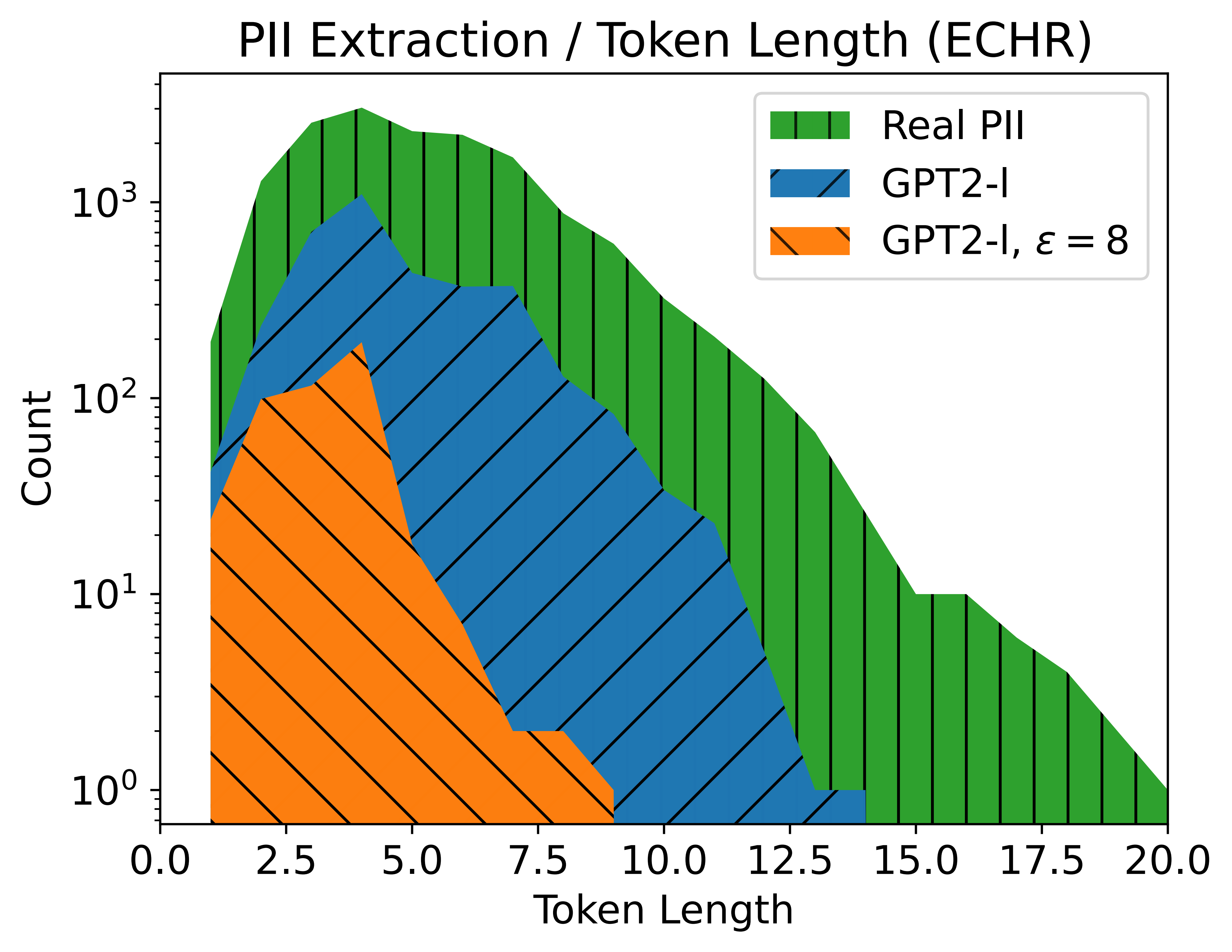


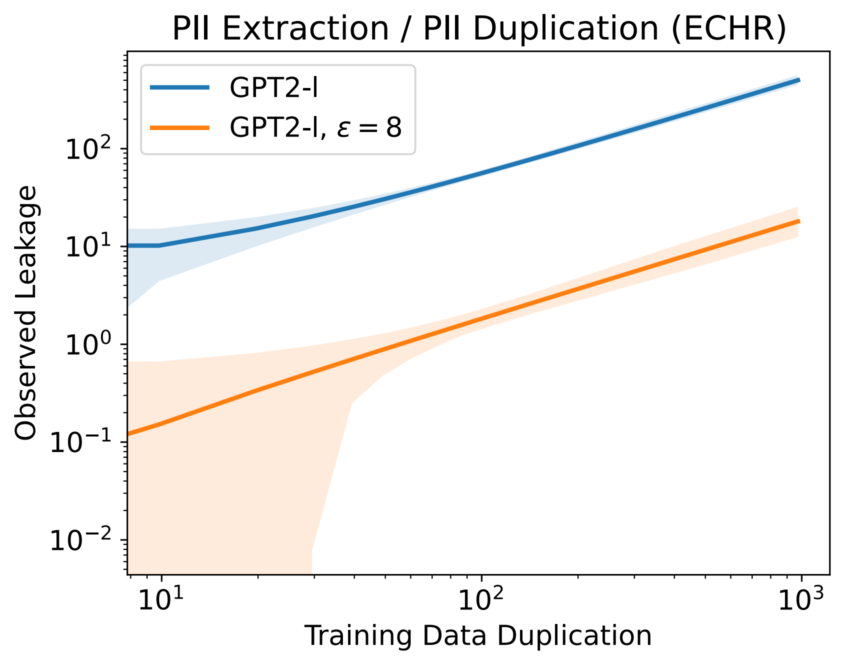


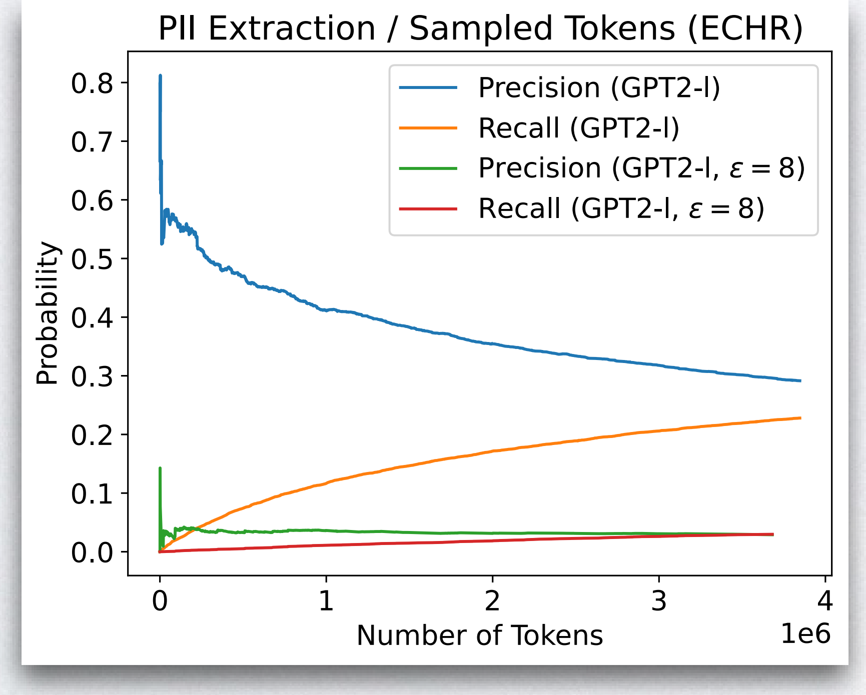
	GPT2-Small		GPT2-Medium		GPT2-Large	
	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$
ECHR						
Prec	24.91%	2.90%	28.05%	3.02%	29.56%	2.92%
Recall	9.44%	2.98%	12.97%	3.21%	22.96%	2.98%
Enron						
Prec	33.86 %	9.37%	27.06%	12.05%	35.36%	11.57%
Recall	6.26%	2.29%	6.56%	2.07%	7.23%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





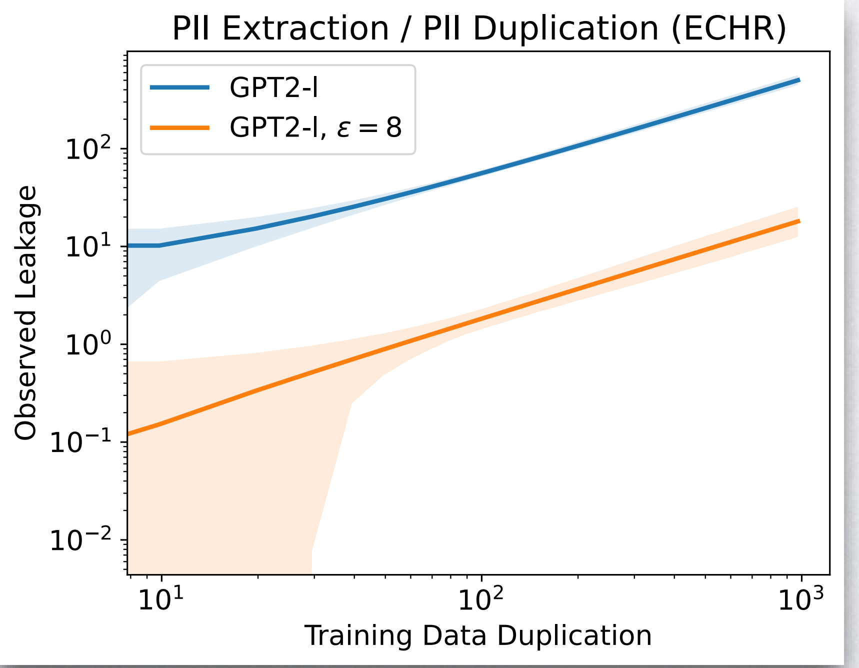


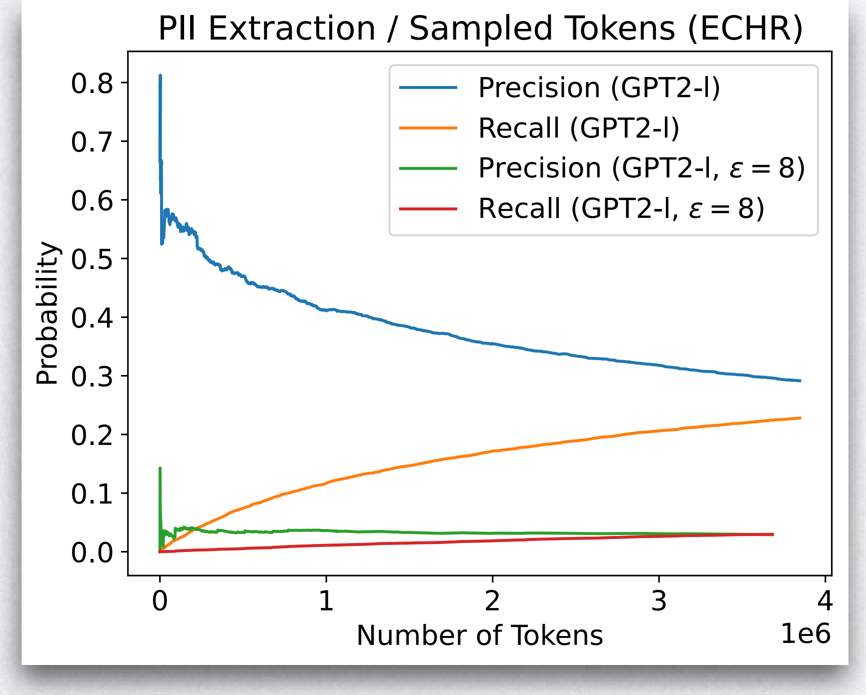
	GPT2-Small		GPT2-Medium		GPT2-Large	
	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$	No DP	$\varepsilon = 8$
ECHR						
Prec	24.91%	2.90%	28.05%	3.02%	29.56%	2.92%
Recall	9.44%	2.98%	12.97%	3.21%	22.96%	2.98%
Enron						
Prec	33.86 %	9.37%	27.06%	12.05%	35.36%	11.57%
Recall	6.26%	2.29%	6.56%	2.07%	7.23%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

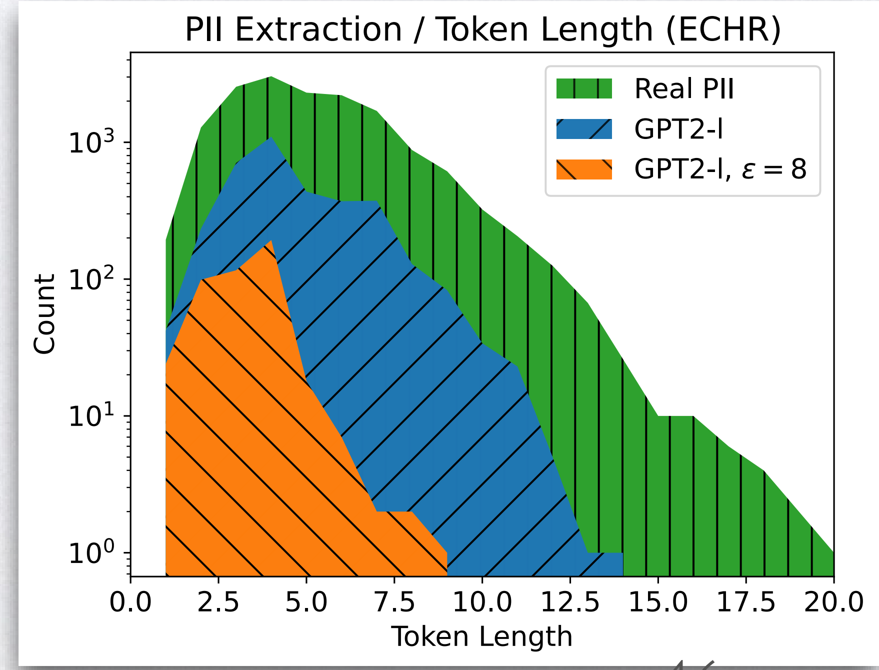
PII Extraction

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%

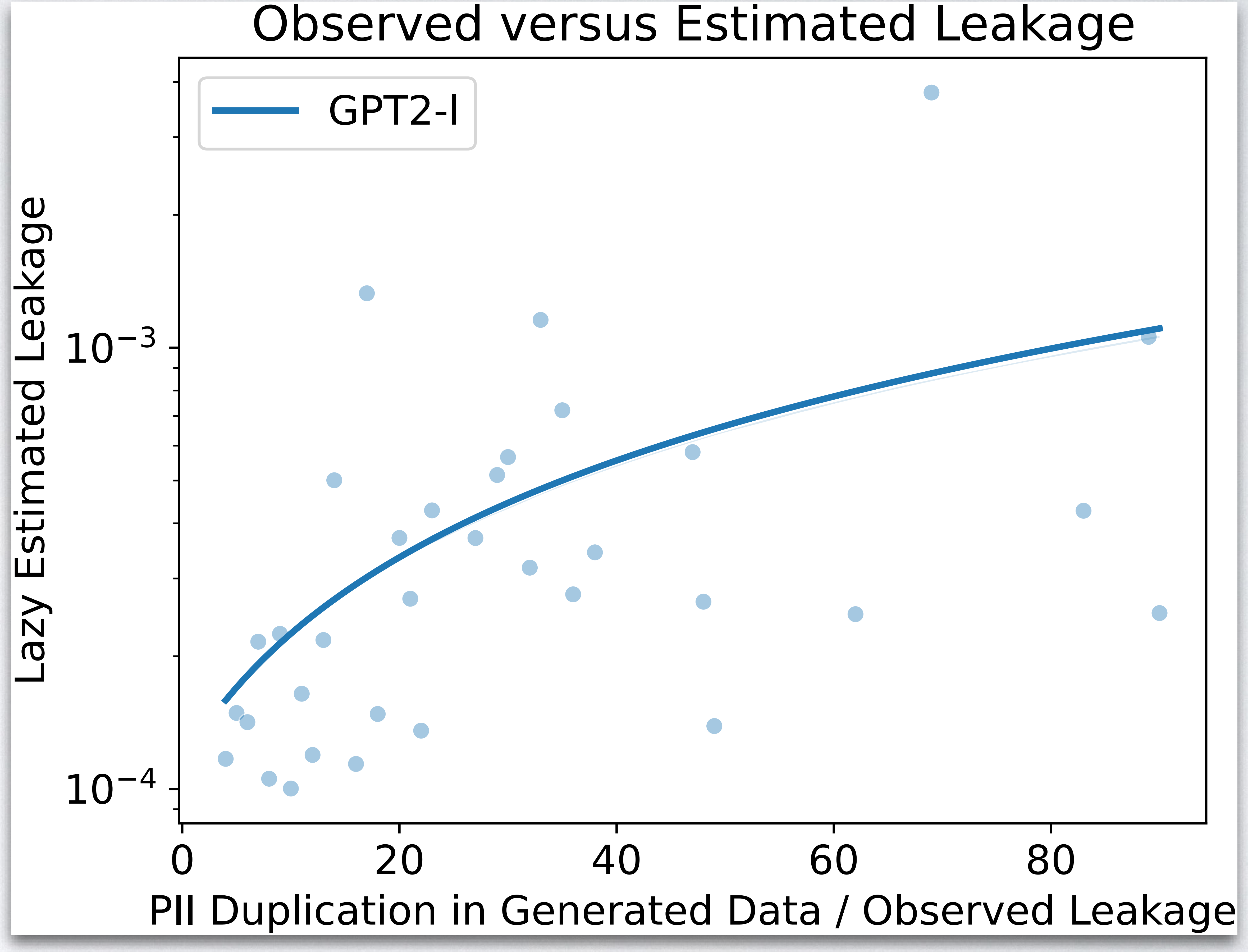






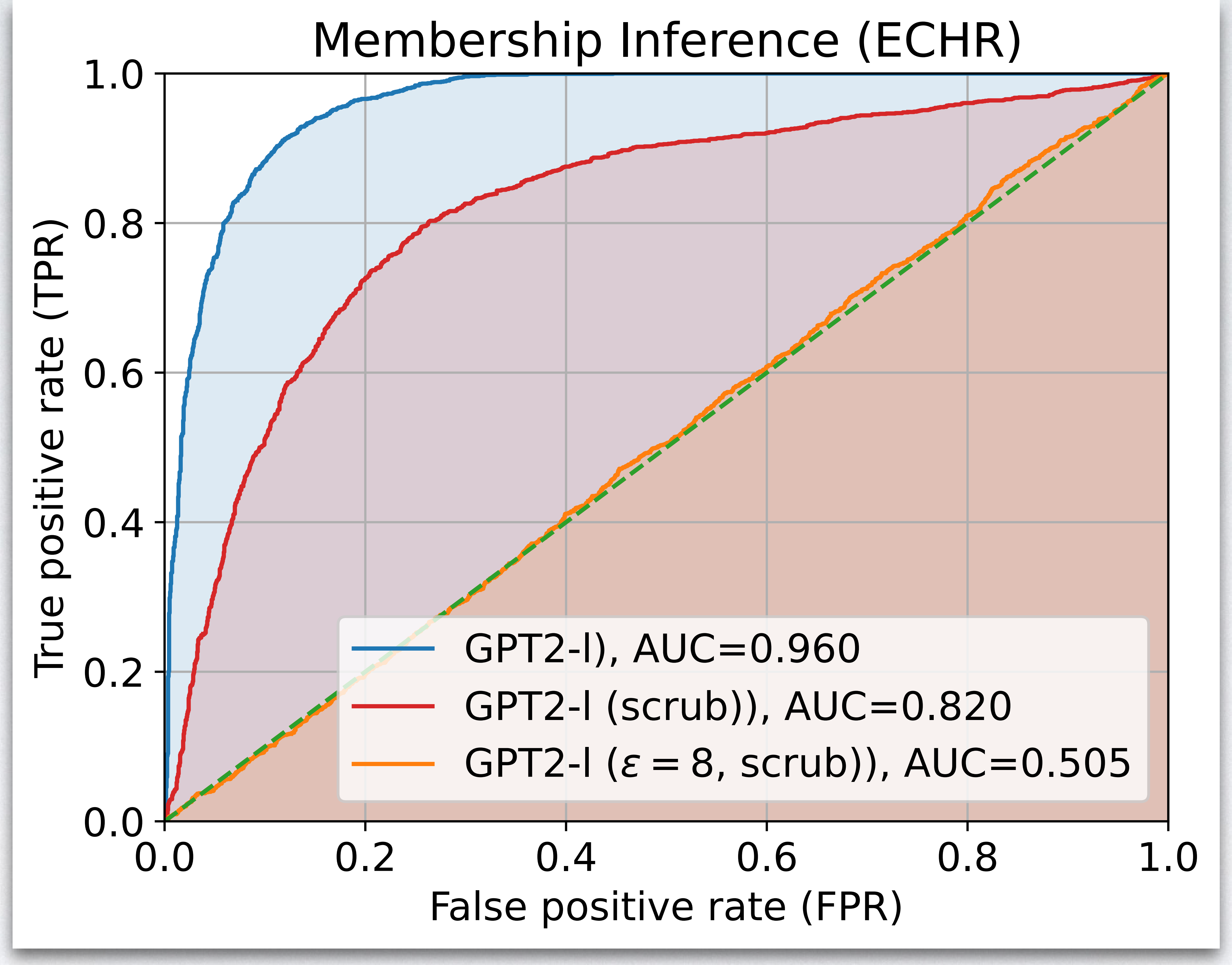
Estimating Extractability

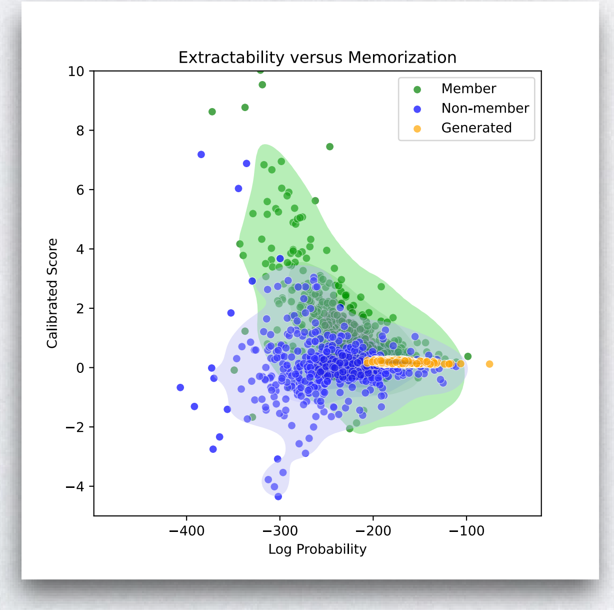
Once upon a time, there existed a tale of two medical students. In the year 2022, they resided at Sunset Street while pursuing their medical education. Alongside his friend, he worked at the LHS Hospital located in the bustling heart of downtown London. Before donning their white coats, both John Doe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

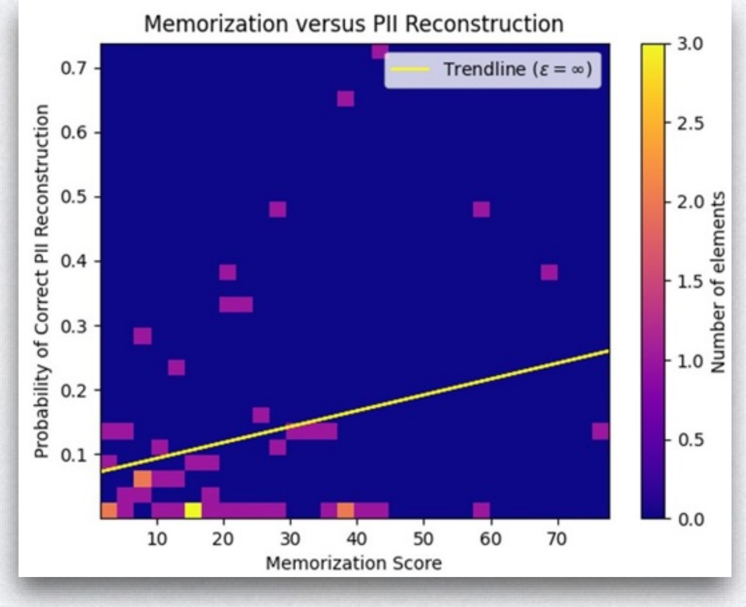


Membership Inference

Scrubbing does not prevent MI

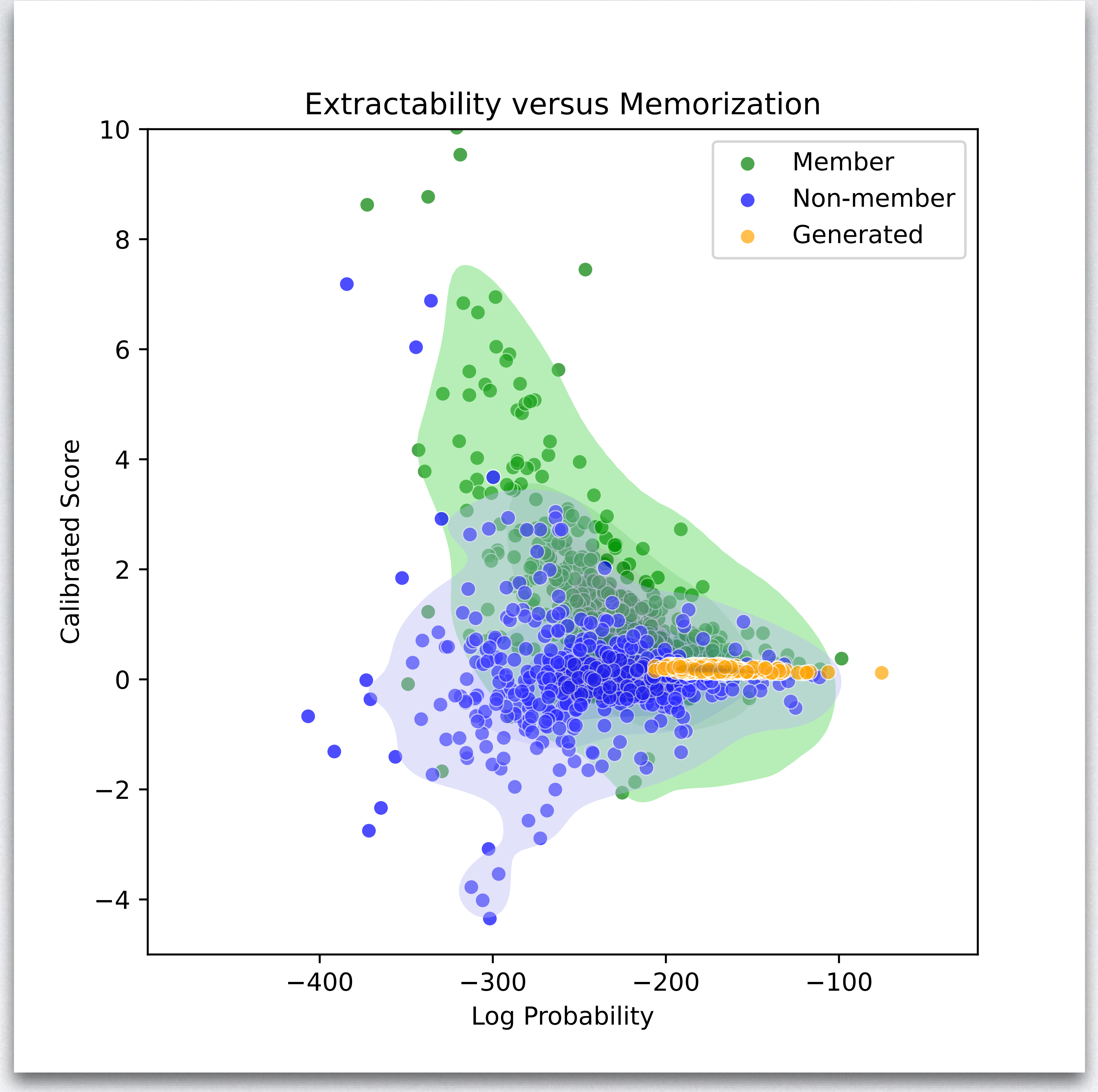


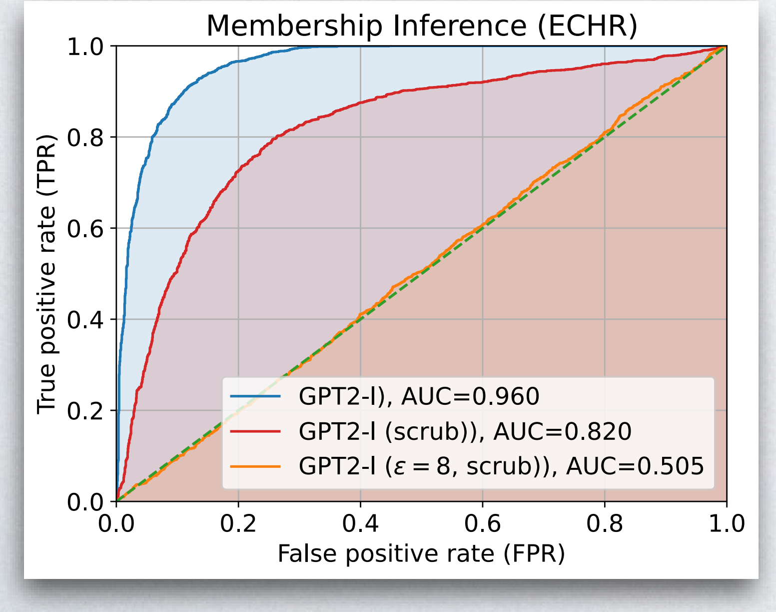


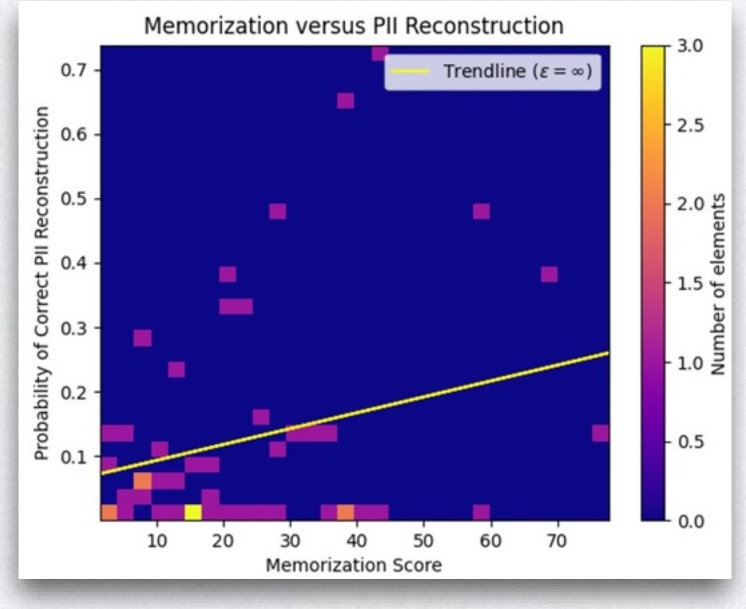


Membership Inference

Randomly generated sequences likely do not contain MI signal

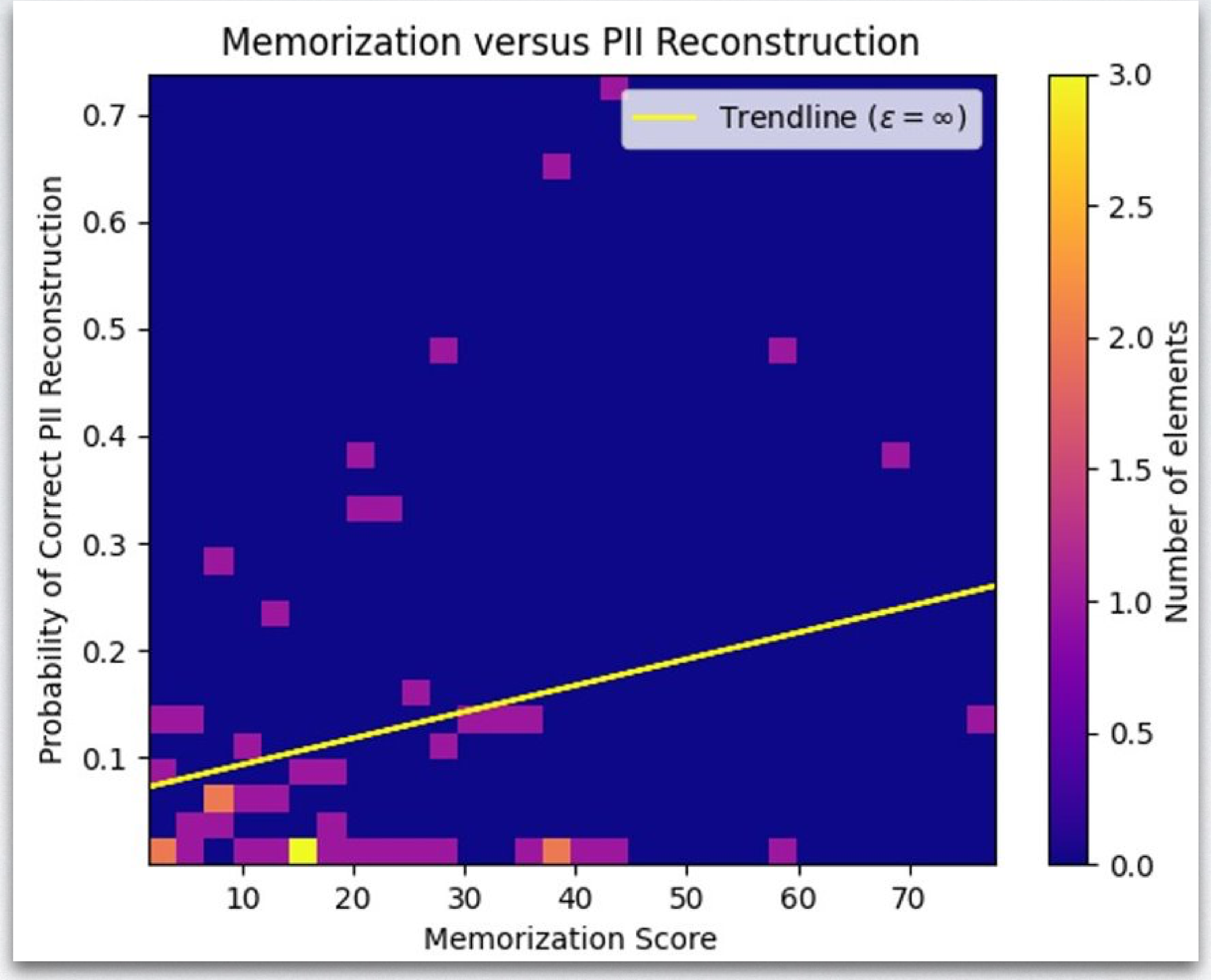


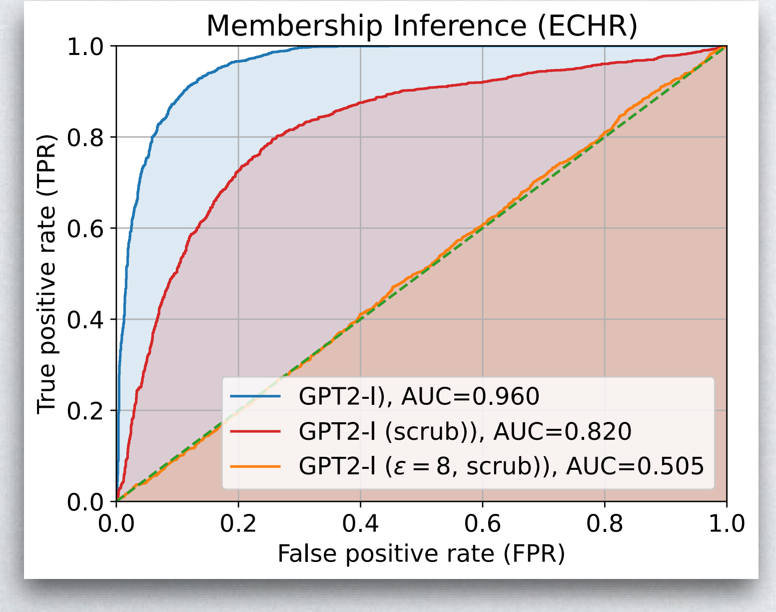


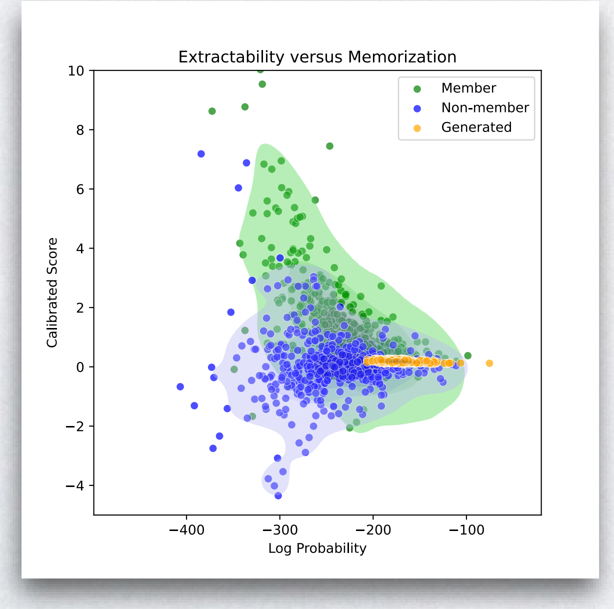


Membership Inference & PII Reconstruction

MI correlates with PII reconstruction







Summary of Results

Metric	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. (|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 between PII)
- **(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]

Scrubbing?

Fake PII?

Stronger attacks / audits?

Unlearning?

Regularization?

1) Data sanitation
2) Alignment
3) Model evaluation
4) Safety filters

Synthetic data?

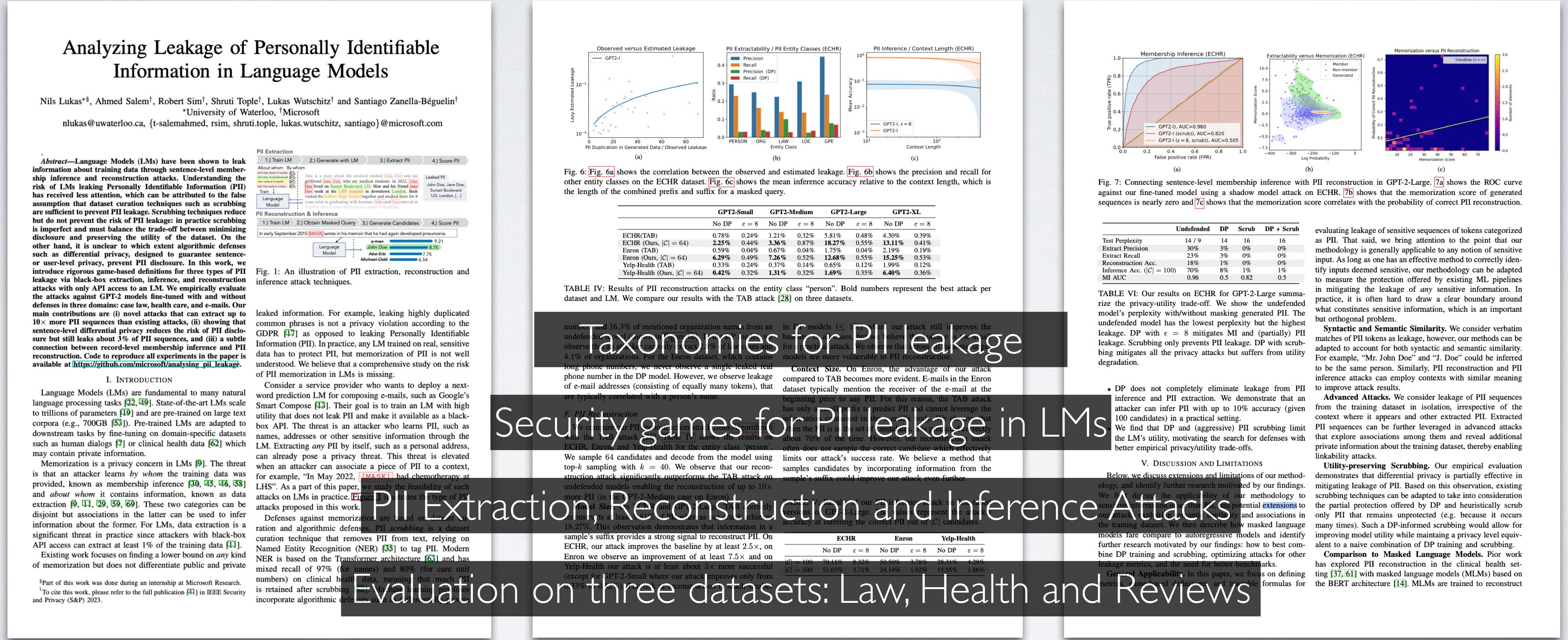
Lower epsilon?

Know your user?

Smaller models?

Red teaming?

Check out our Paper for more Information



Taxonomies for PII leakage

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

[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.

Homepage

https://nilslukas.github.io