Analyzing Leakage of Personally Identifiable Information in Language Models





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



Privacy Concerns in ChatBots

| Bloom | berg | | | | | | | | | US | Edition 🔻 | Sign In |
|---------------------|---------|-----------|------------|-----------------------------|----------|--------|----------|---------|-----------------------------|----------|--|---------|
| • Live Now | Markets | Economics | Industries | Technology | Politics | Wealth | Pursuits | Opinion | Businessweek | Equality | Green | CityLat |
| Techno City of I | | | | treet verec | | | | Crac | cking I | Dow | n oi | n |
| | | | | 2 | | U | | · • | ople familia s industrie | | | |
| | | Feb | | Iello, Willia 2023, 7:36 | | | | | 2023, 10:10 | Wa | LIVE ON BLOO atch Live TV sten to Live I | > |
| | | | | Blo | or | nt | Der | rg, | 202 | 23 | ΓΙ |] |

ChatGPT banned in Italy over privacy concerns

🕓 1 April

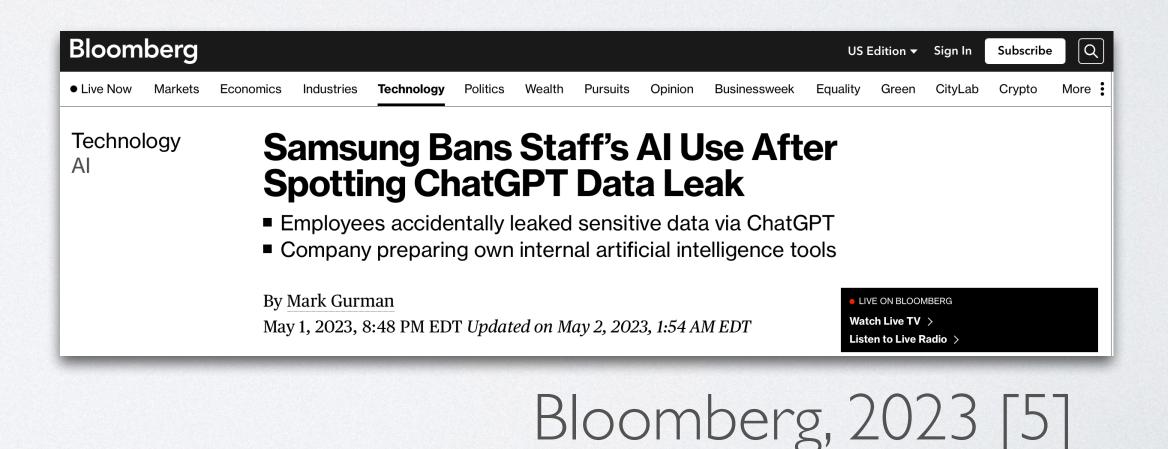
ChatGPT accessible again in Italy

🕓 28 April

BBC News, 2023 [3,4]



Business Insider, 2023 [2]



6. Will you use my conversations for training?

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

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.

Terms of use

ChatGPT, OpenAI [6]

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]

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]

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

Bleedingcomputer, 2023 [11]

Privacy Concerns for Code-Completion

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]





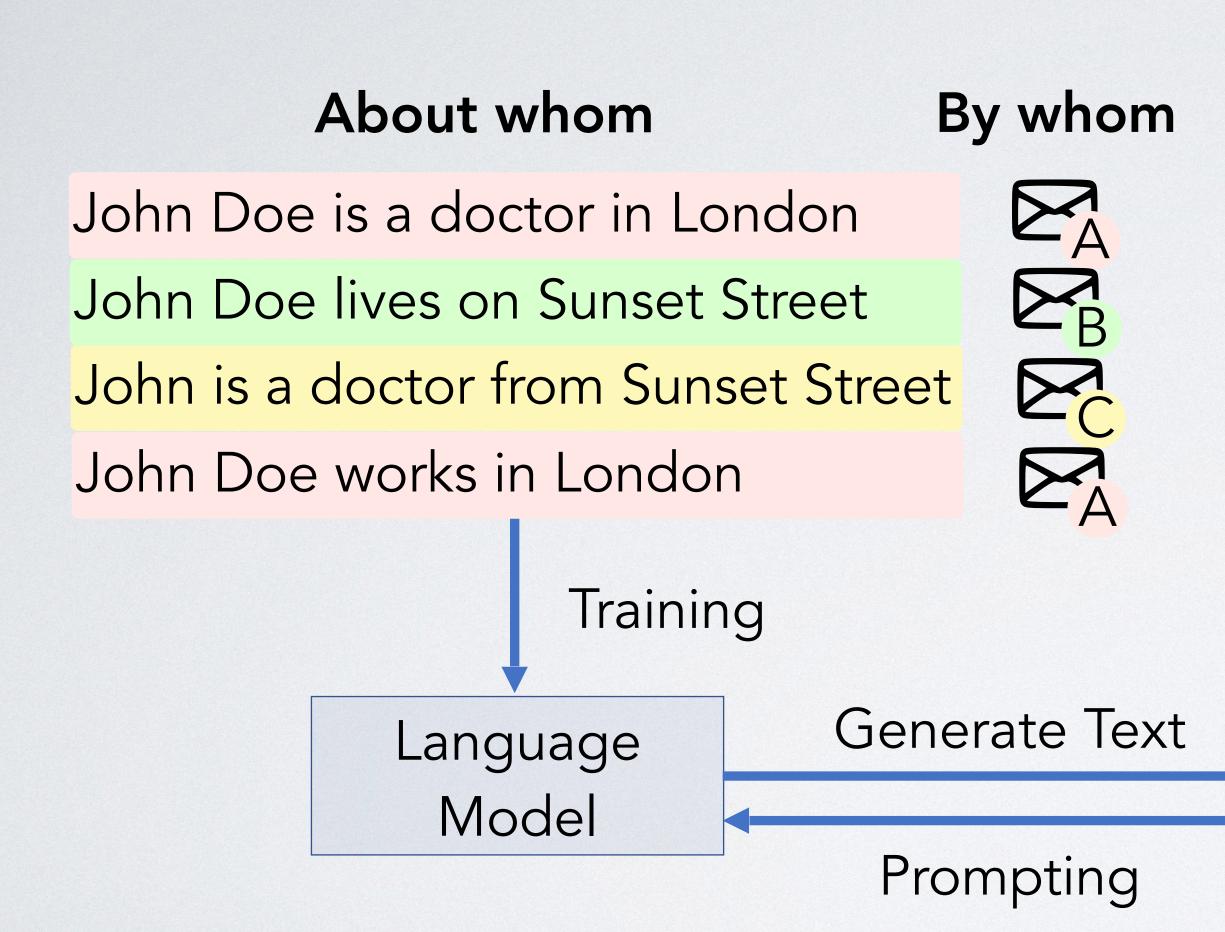
Few-Shot In Context Learning versus Fine-Tuning

-

-

- (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







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.

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.

I.) PII Extraction

John Doe London Sunset Street LHS Hospital

Jane Doe Aubrey High School

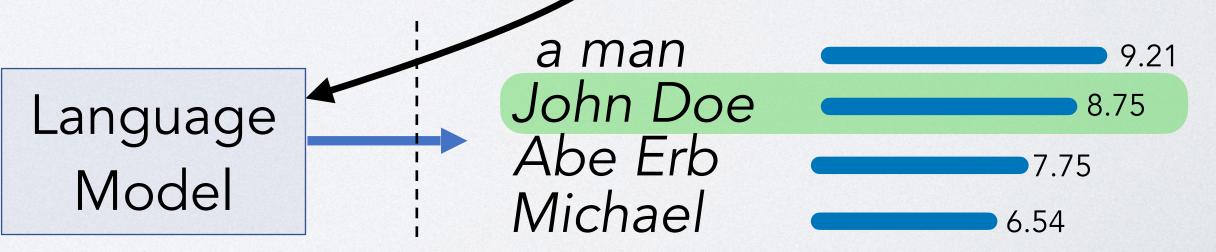
Real

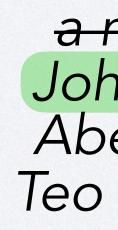
Fictional

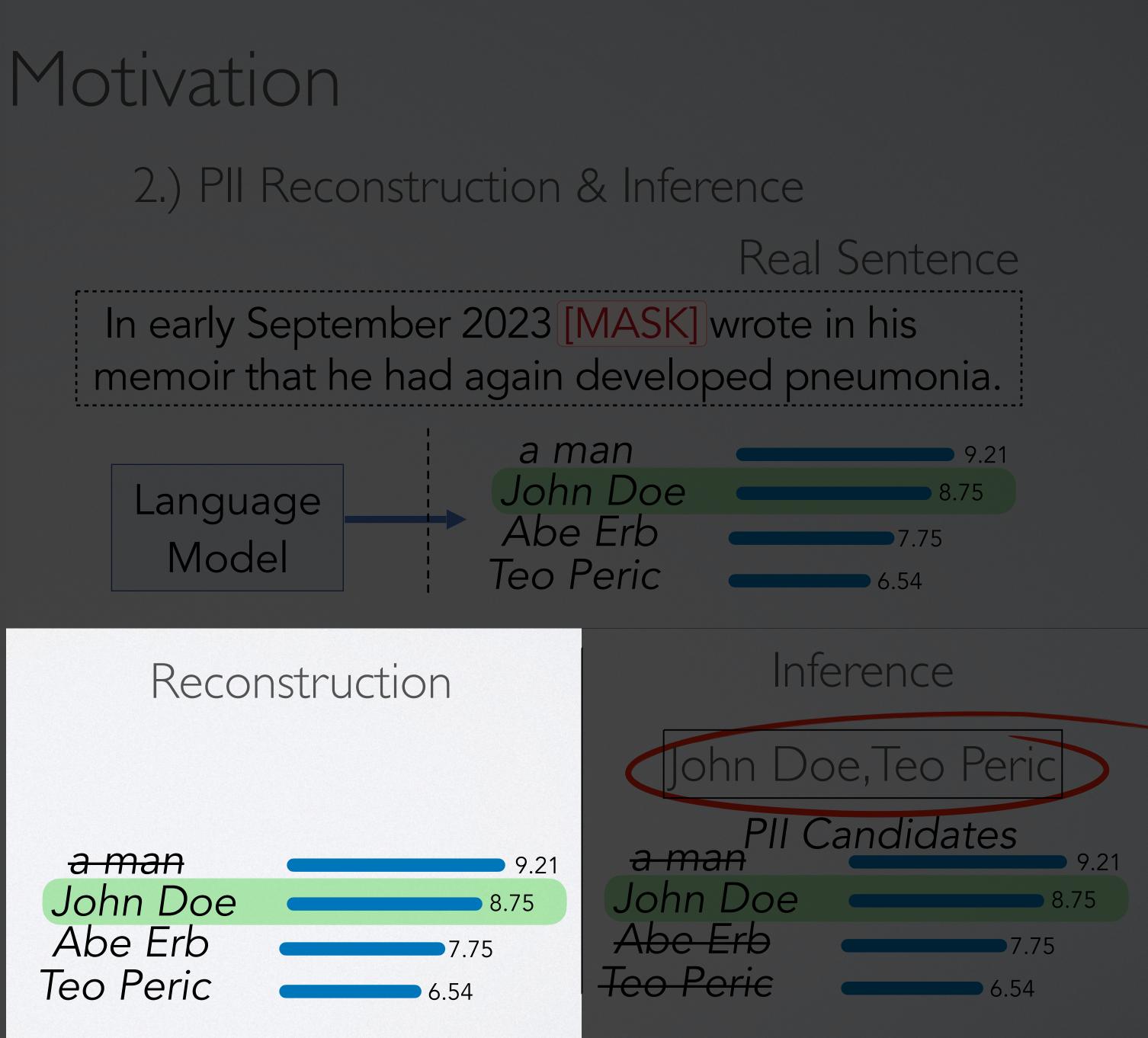
2.) PII Reconstruction & 3.) PII Inference

Real Sentence

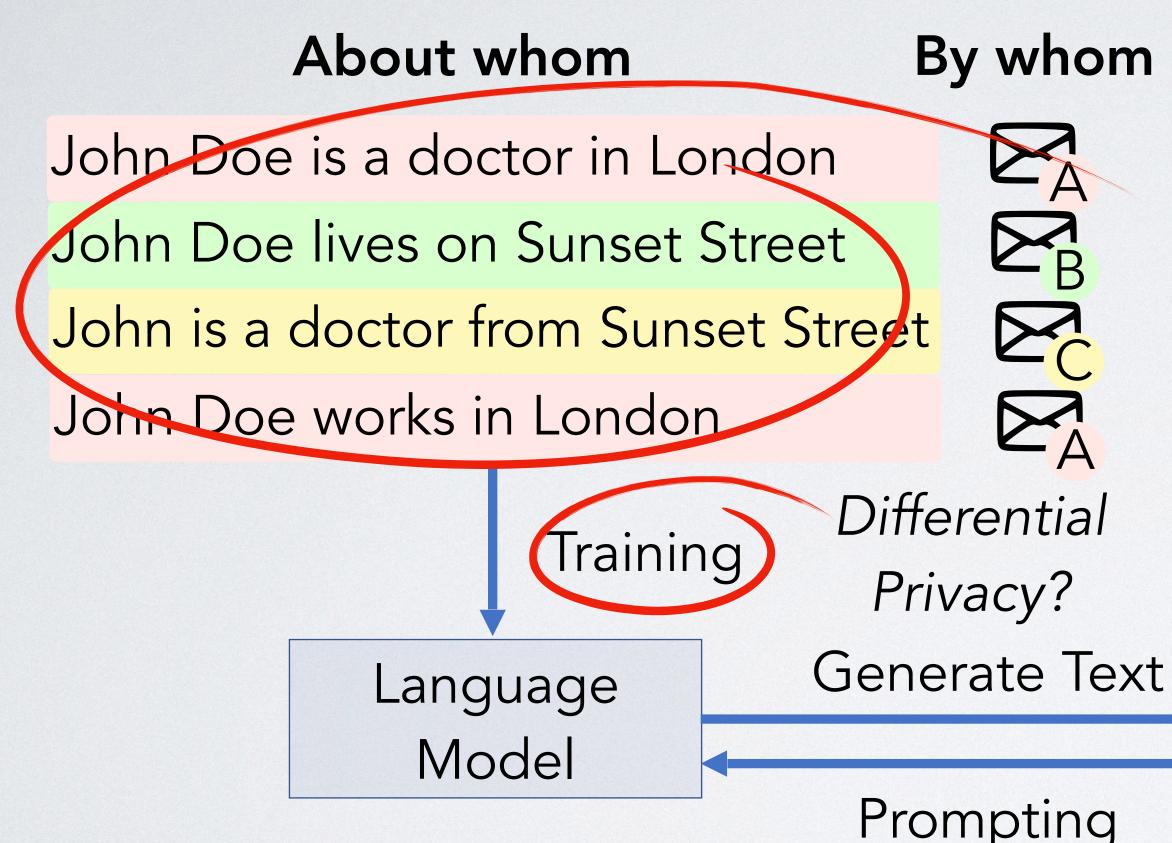
In early September 2023 [MASK] wrote in his memoir that he had again developed pneumonia.







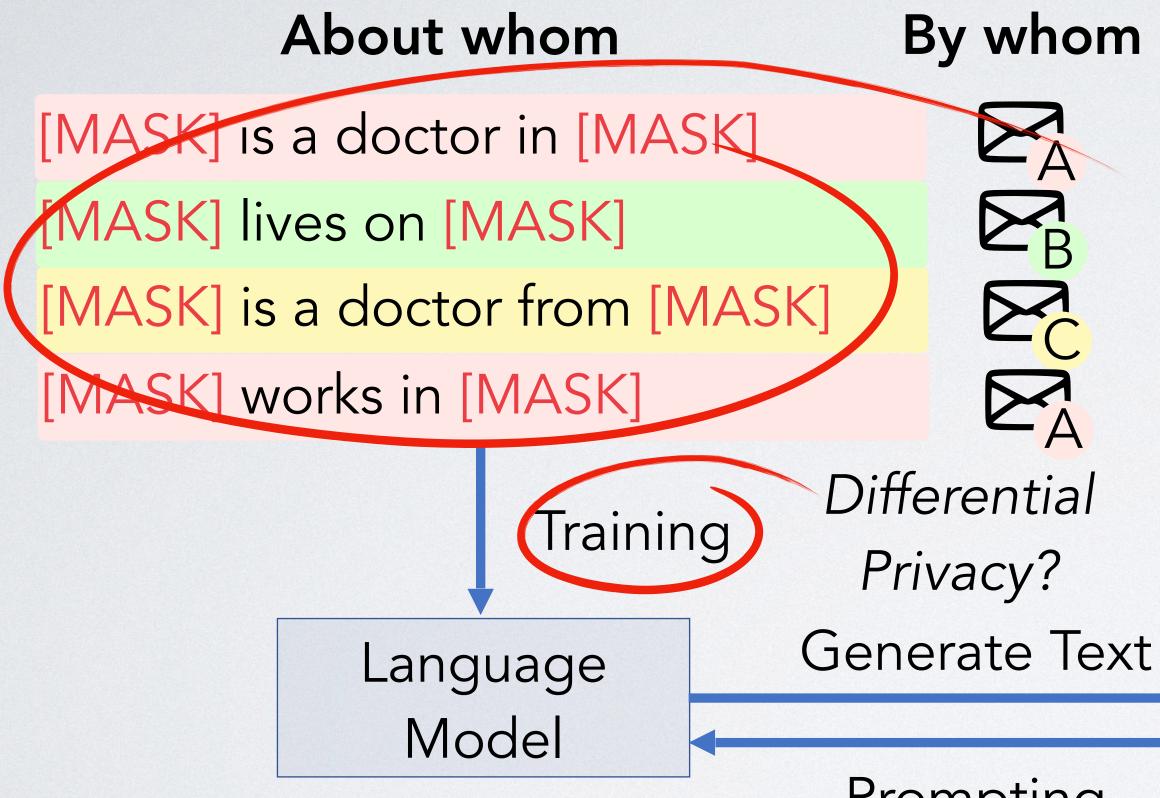
PII Scrubbing?



Motivation



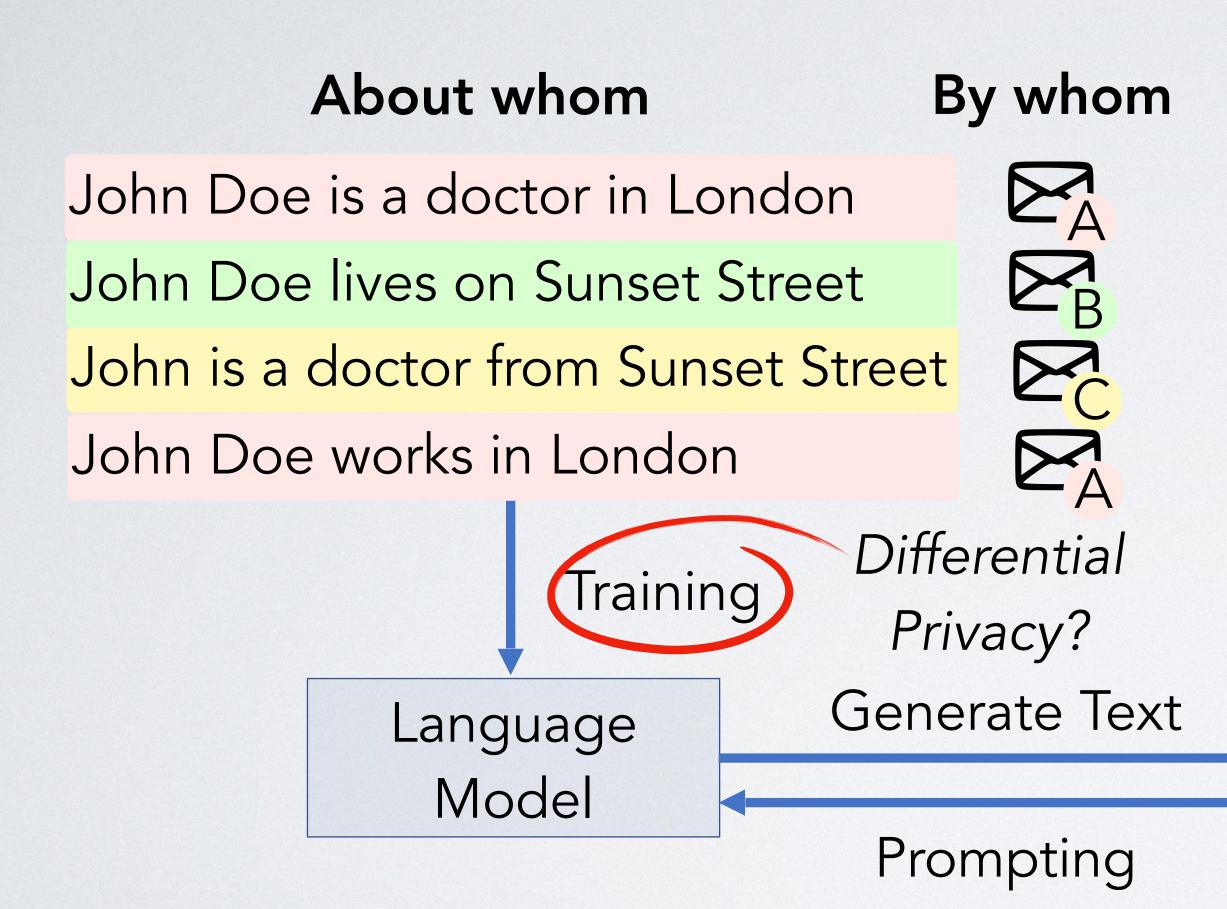
PII Scrubbing?



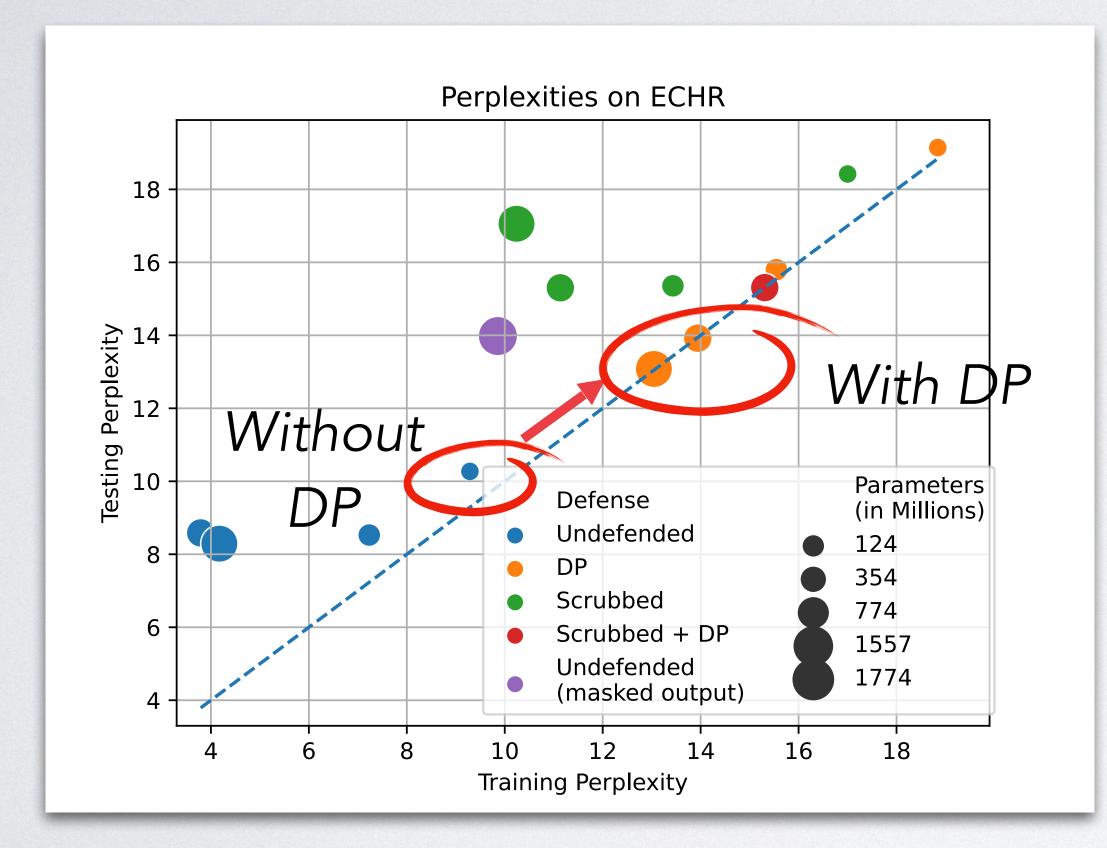
Prompting

Motivation









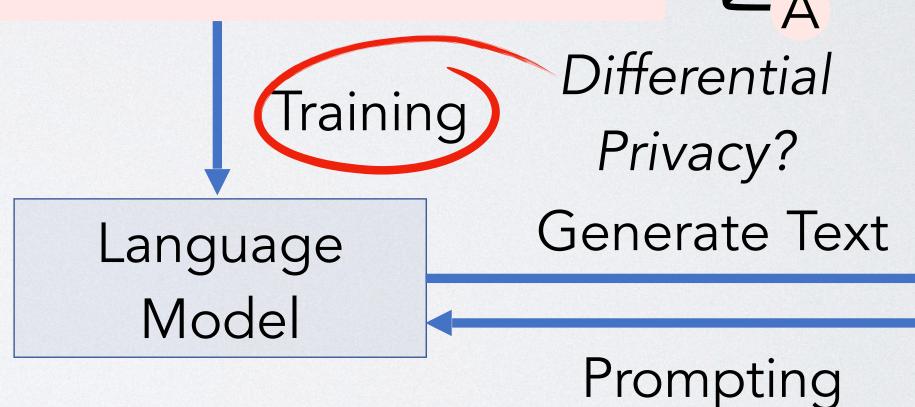
<u>Privacy</u> at the cost of <u>Model Utility</u>

Problems with Differential Privacy

About whom

By whom

E A B C John Doe is a doctor in London John Doe lives on Sunset Street John is a doctor from Sunset Street P_A John Doe works in London



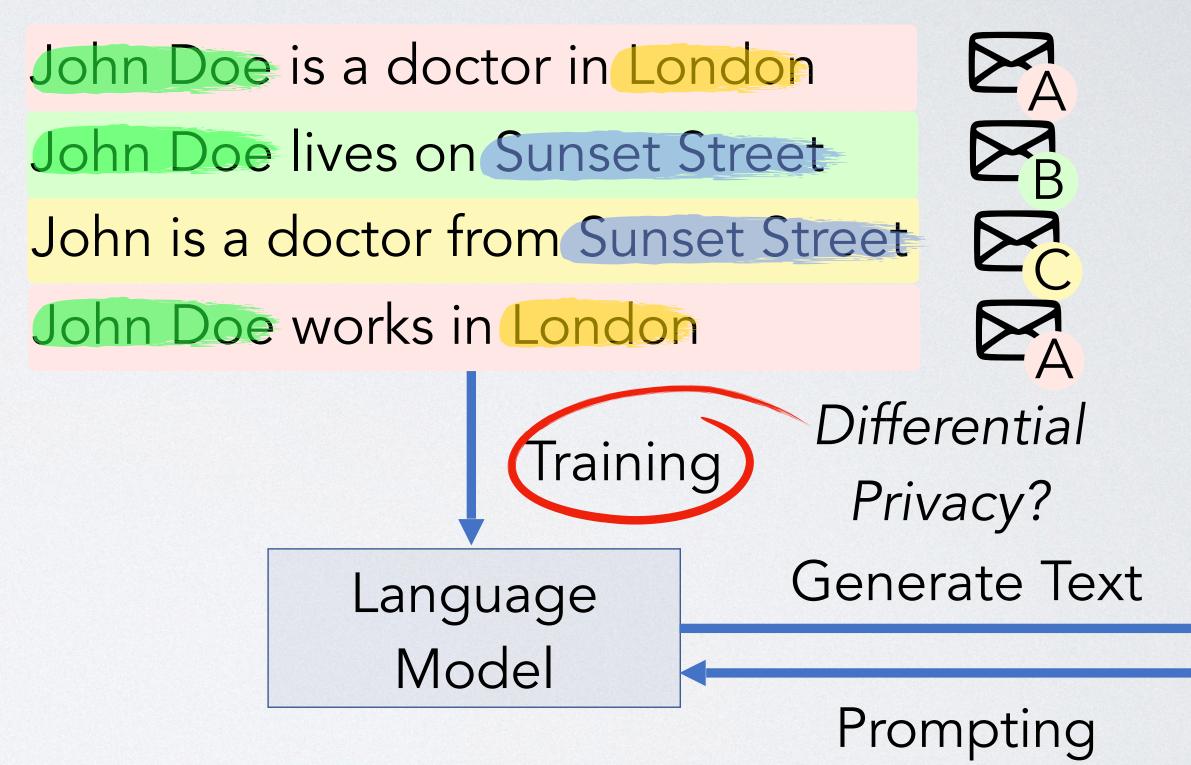


DP protects against an attacker learning by whom data was provided, but not about whom it contains information.

Problems with Differential Privacy

About whom

By whom





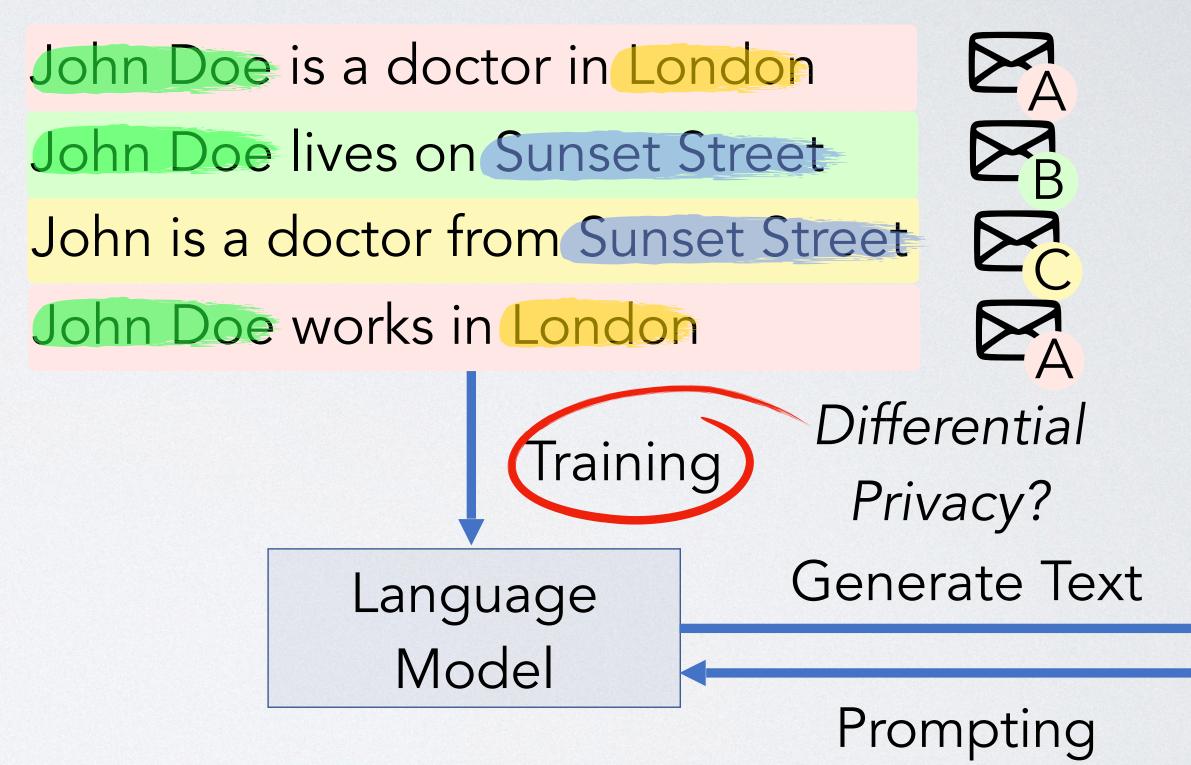
Group-level DP can help but ..

- 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

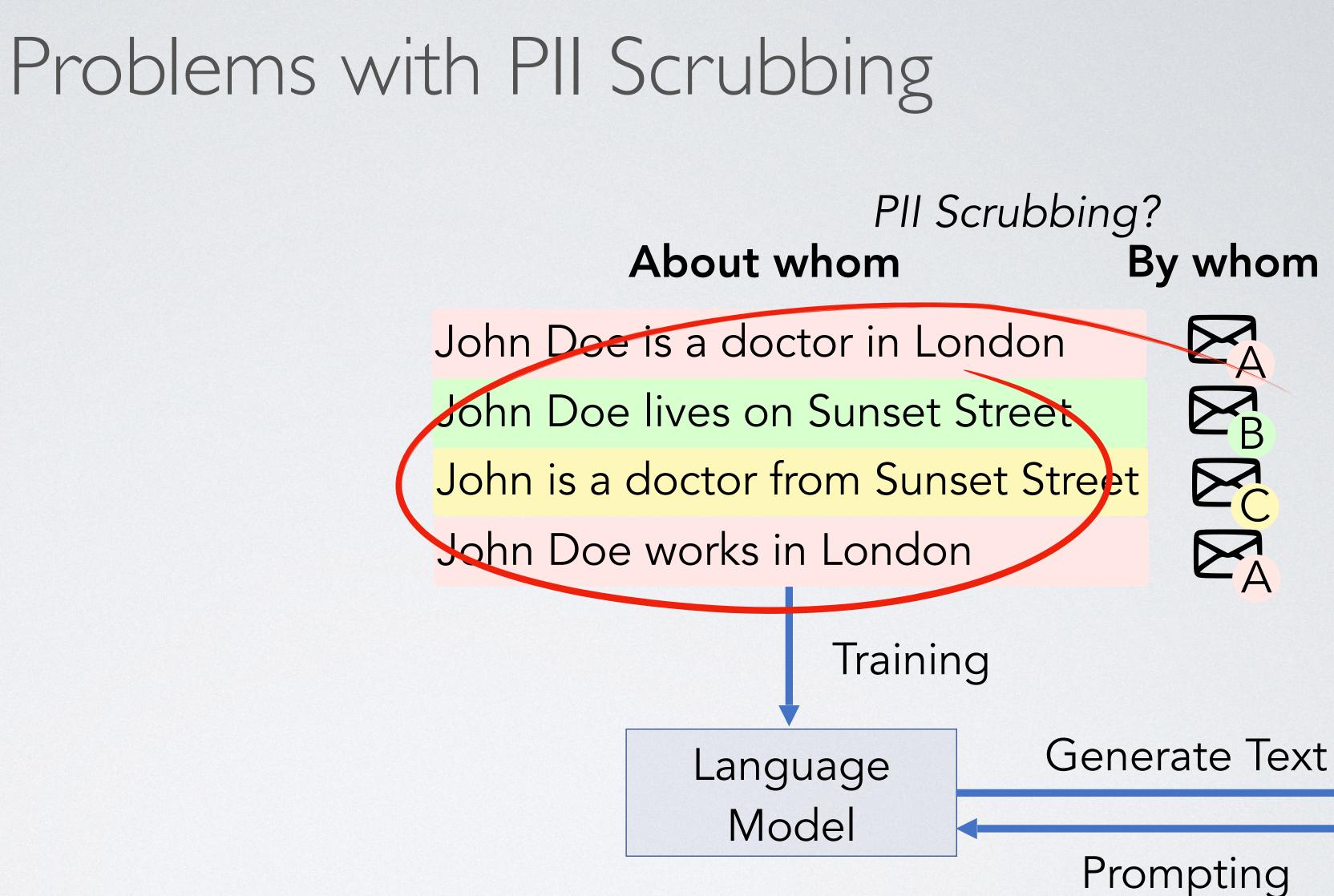
Problems with Differential Privacy

About whom

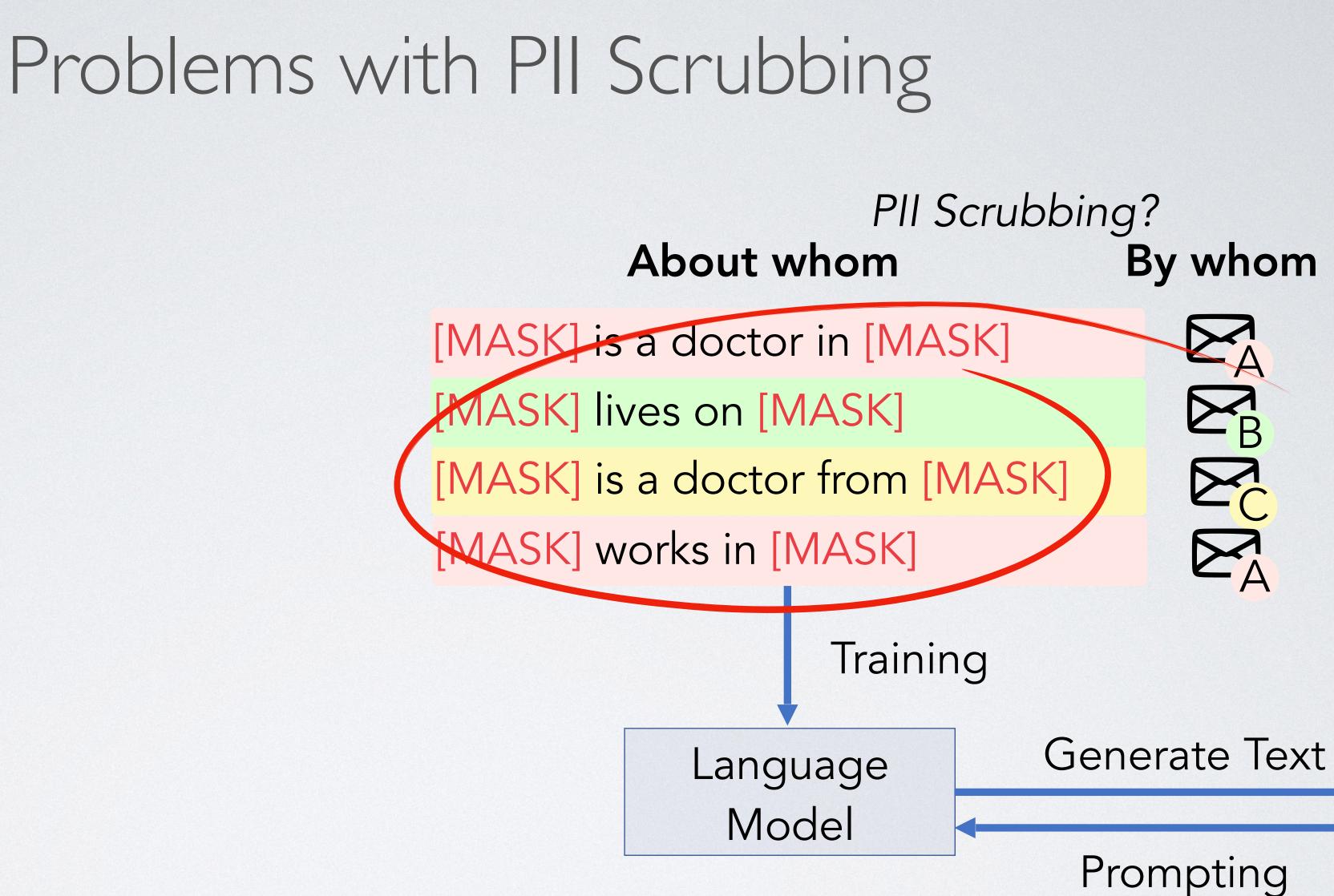
By whom



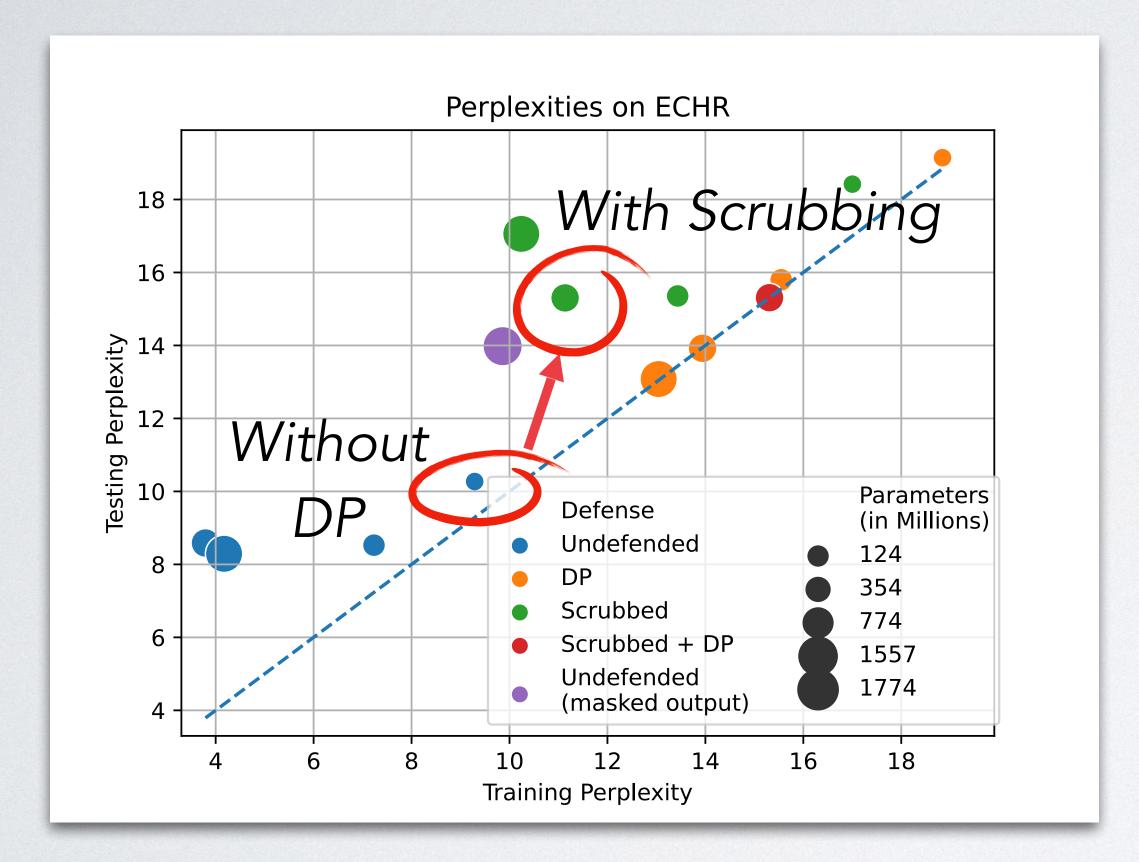




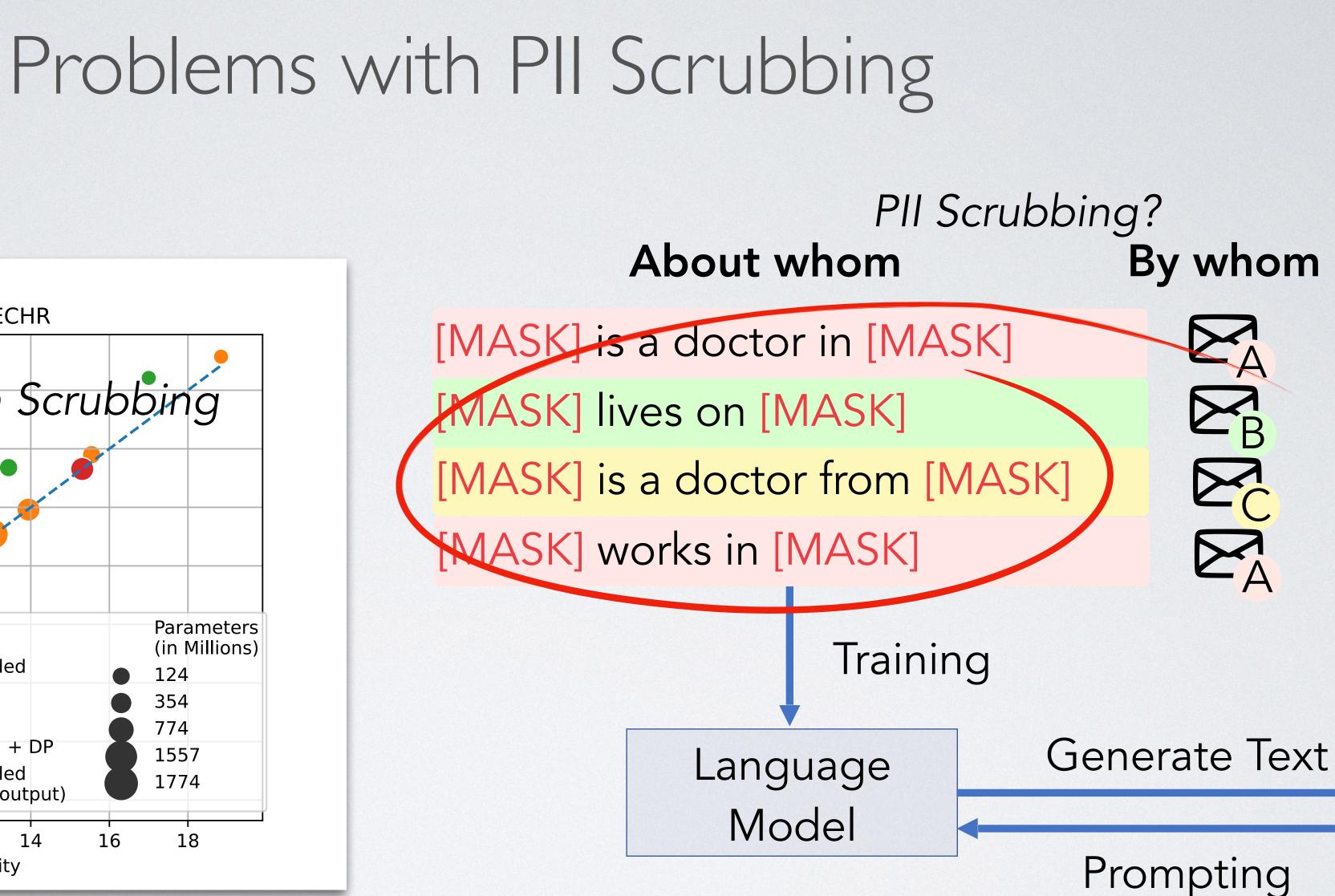






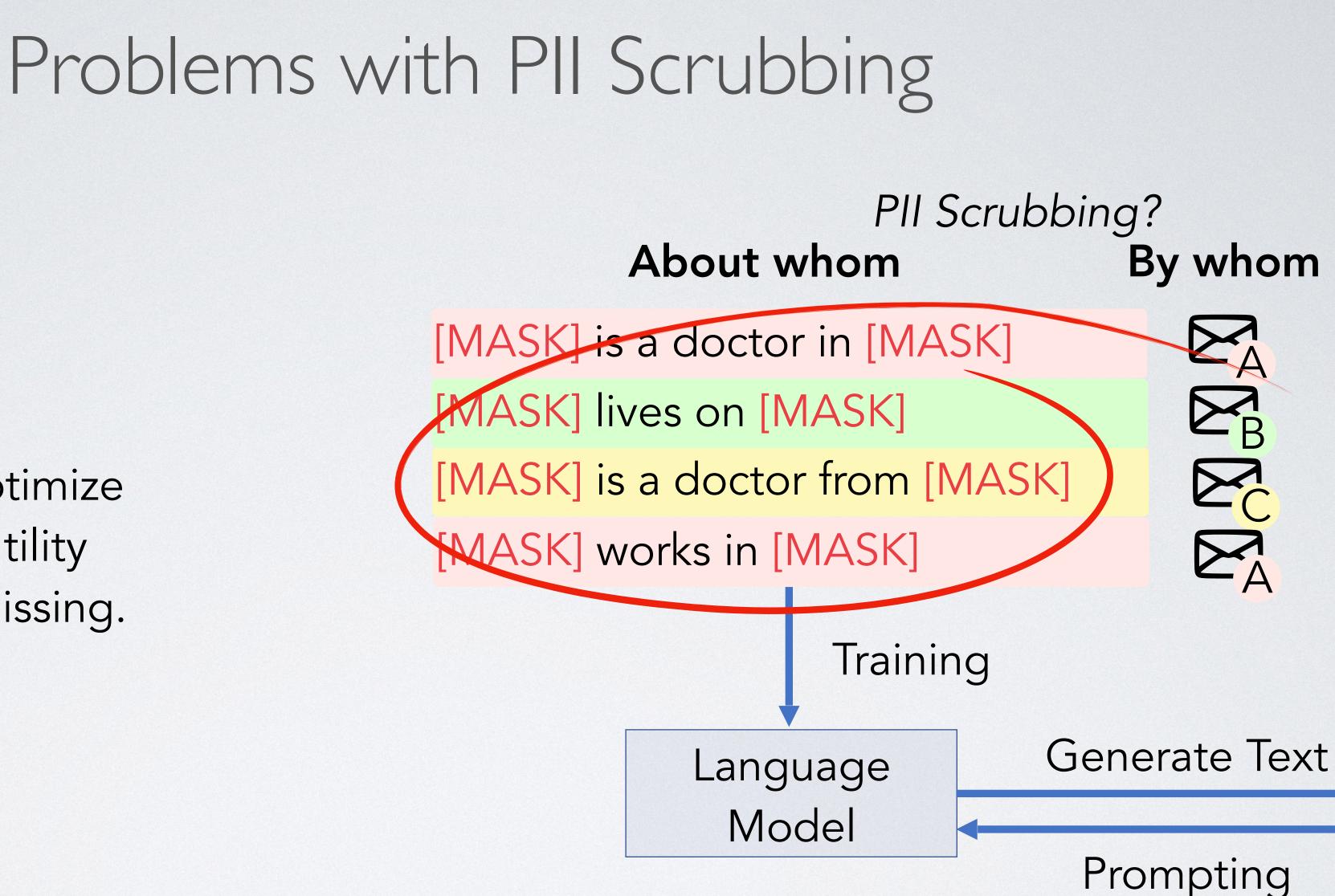


<u>Privacy</u> at the cost of <u>Model Utility</u>





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





Canaries

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

Nicholas Carlini^{1,2} Chang Liu² Úlfar Erlingsson¹ Jernej Kos³ Dawn Song² ¹Google Brain ²University of California, Berkeley ³National University of Singapore

Abstract

This paper describes a testing methodology for quantitatively assessing the risk that rare or unique training-data sequences are *unintentionally memorized* by generative sequence models—a common type of machine-learning model. Because such models are sometimes trained on sensitive data Because such models are sometimes trained on sensitive data (e.g., the text of users' private messages), this methodology can benefit privacy by allowing deep-learning practitioners to select means of training that minimize such memorization. In experiments, we show that unintended memorization a persistent, hard-to-avoid issue that can have serious conse-quences. Specifically, for models trained without considera-tion of memorization, we describe new, efficient procedures that can extrat unious. exerct sequences: such as credit card that can extract unique, secret sequences, such as credit card numbers. We show that our testing strategy is a practical and easy-to-use first line of defense, e.g., by describing its application to quantitatively limit data exposure in Google's mart Compose, a commercial text-completion neural network trained on millions of users' email messages

1 Introduction

When a secret is shared, it can be very difficult to prevent its further disclosure—as artfully explored in Joseph Conrad's The Secret Sharer [10]. This difficulty also arises in machine-The Secret Sharer [10]. This difficulty also arises in machine-learning models based on neural networks, which are being rapidly adopted for many purposes. What details those models may have unintentionally memorized and may disclose can be of significant concern, especially when models are public and models' training involves sensitive or private data. Disclosure of secrets is of particular concern in neural-network models that classify on predict semences of natural.

network models that classify or predict sequences of naturallanguage text. First, such text will often contain sensitive or private sequences, accidentally, even if the text is supposedly public. Second, such models are designed to learn text patterns such as grammar, turns of phrase, and spelling, which comprise a vanishing fraction of the exponential space of all possible sequences. Therefore, even if sensitive or priall possible sequences. Therefore, even if sensitive or pri-vate training-data text is very rare, one should assume that well-trained models have paid attention to its precise details. Concretely, disclosure of secrets may arise naturally in gen-erative text models like those used for text auto-completion and predictive keyboards, if trained on possibly-sensitive data. The users of such models may discover—either by accident or on purpose—that entering certain text prefixes causes the models to output surviviently-revealing text completions [37]. models to output surprisingly-revealing text completions [37].

For example, users may find that the input "my social-security number is..." gets auto-completed to an obvious secret (such as a valid-looking SSN not their own), or find that other inputs are auto-completed to text with oddly-specific details. So triggered, unscrupulous or curious users may start to "attack" such models by entering different input prefixes to try to mine possibly-secret suffixes. Therefore, for generative text mod els, assessing and reducing the chances that secrets may be disclosed in this manner is a key practical concern.

To enable practitioners to measure their models' propensit for disclosing details about private training data, this pape itative metric of exposure. This metric ca the applied during training as part of a testing methodology that empirically measures a model's potential for unintended memorization of unique or rare sequences in the training data.

Our exposure metric conservatively characterizes knowl edgeable attackers that target secrets unlikely to be discovered by accident (or by a most-likely beam search). As validation by accident (or by a most-likely beam search). As validation of this, we describe an algorithm guided by the exposure met-ric that, given a pretrained model, can efficiently extract secret sequences even when the model considers parts of them to be highly unlikely. We demonstrate our algorithm's effectiveness in experiments, e.g., by extracting credit card numbers from a howevee model busined on the Barne around lette. Such a combilanguage model trained on the Enron email data. Such empir cal extraction has proven useful in convincing practition that unintended memorization is an issue of serious, practical concern, and not just of academic interest. Our exposure-based testing strategy is practical, as we

demonstrate in experiments, and by describing its use in removing privacy risks for Google's Smart Compose, a deployed, commercial model that is trained on millions of user email messages and used by other users for predictive text completion during email composition [8].

In evaluating our exposure metric, we find unintended me orization to be both commonplace and hard to prevent. In particular, such memorization is not due to overtrai it occurs early during training, and persists across differe it occurs early during training, and persists across different types of models and training strategies—even when the mem-orized data is very rare and the model size is much smaller than the size of the training data corpus. Furthermore, we show that simple, intuitive regularization approaches such as early-stopping and dropout are insufficient to prevent un-intended memorization. Only by using differentially-private training techniques are we able to eliminate the issue com-nletely. albeit at some loss in utility. pletely, albeit at some loss in utility.

Carlini et al., 2019

N-grams

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

R. Thomas McCoy,^{*1} Paul Smolensky,^{2,1} Tal Linzen,³ Jianfeng Gao,² Asli Celikyilmaz^{*4} ¹Johns Hopkins University ²Microsoft Research ³New York University ⁴Facebook AI Research tom.mccoy@jhu.edu, psmo@microsoft.com, linzen@nyu.edu, jfgao@microsoft.com, aslic@fb.com

Current language models can generate high quality text. Are they simply copying text they have seen before, or have they learned generalizable linguistic abstractions? To ease apart these possibilities, we intr duce RAVEN, a suite of analyses for as-sessing the novelty of generated text, fo-cusing on sequential structure (*n*-grams) and syntactic structure. We apply thes analyses to four neural language models (an LSTM, a Transformer, Transformer-XL, and GPT-2). For local structure-e.g., ind vidual dependencies-model-generated to s substantially less novel than our base line of human-generated text from each model's test set. For larger-scale structuree.g., overall sentence structure-model generated text is as novel or even more novel than the human-generated baseline but models still sometimes copy substan tially, in some cases duplicating passage over 1,000 words long from the training se We also perform extensive manual analyst showing that GPT-2's novel text is usuall well-formed morphologically and syntact cally but has reasonably frequent semanti ssues (e.g., being self-contradictory)

Abstract

1 Introduction

How deep is deep learning? Are neural networks "discovering intricate structures" that support sophisticated generalization (LeCun et al., 2015), or

ext (See et al., 2019; Brown et al., 2020, section Start Goe Cran, 2019, Drivin Ce and 2005, Sector and Construction on advanced period by consonney inter and sector advanced period by constructing inter advanced period by constructing

McCoy et al., 2019

constructed by the model or copied from the train ing set. We argue that it is important to disentangle these possibilities. That is, in addition to evaluat ing the *quality* of generated text, as is already stan-dard (Gatt and Krahmer, 2018; Celikyilmaz et al., 2020), we should also evaluate its novelty. Novelty is important for several reasons. From

a linguistic perspective, one core component of knowing a language is the ability to combine familiar parts in novel ways (Chomsky, 1957; Hock-ett, 1963). From a machine learning perspective models are meant to learn the training distribution tion, not just memorize the training set (Dietterich 1995). Finally, on the more practical side, models that copy training data might leak sensitive information (Carlini et al., 2021) or repeat hate speech (Bender et al., 2021).

In this work, to assess the novelty of gener ated text, we introduce a suite of analyses called RAVEN (RAting VErbal Novelty).^{1,2} These analyses cover both sequential structure (*n*-grams) and syntactic structure. We apply these analyses to text generated by an LSTM, a Transformer, Transformer-XL, and all 4 sizes of GPT-2 (the largest LM for which we had access to the training data). Because there are many ways to generat text from LMs, we test 12 generation methods and 4 prompt lengths. As a baseline, we also analyze human-generated text from each model's test set. We find that models display novelty for all aspects of structure that we analyze: they gener-

histicated generalization (LeCun et al., 2015), or are they "stochastic parrots" that simply memo-rize seen examples and recombine them in shallow ways (Bender et al., 2021)? We focus on this question in the area of open-ended text generation. Neural network language models (I.Ms) can generate grammatical, coherent

Nicholas Carlini¹ Ariel Herbert-Voss5,6

Abstract

It has become common to publish large (billion parameter) language models that have been trained on private datasets. This paper demonstrates that in such settings, an adversary can perform a *training data extraction attack* to recover individual training examples by querying the language model. We demonstrate our attack on **GPT**-2, a language model. trained on scrapes of the public Internet, and are able to extrac hundreds of verbatim text sequences from the model's training data. These extracted examples include (public) personall identifiable information (names, phone numbers, and ema addresses), IRC conversations, code, and 128-bit UUIDs. Out attack is possible even though each of the above sequences are included in just one document in the training data are included in just one document in the training data. We comprehensively evaluate our extraction attack to un-derstand the factors that contribute to its success. Worryingly, we find that larger models are more vulnerable than smaller models. We conclude by drawing lessons and discussing pos-sible safeguards for training large language models.

1 Introduction

Language models (LMs)-statistical models which assign a Language models (LMs)—statistical models which assign a probability to a sequence of words—are fundamental to many natural language processing tasks. Modern neural-network-based LMs use very large model architectures (e.g., 175 bil-lion parameters [7]) and train on massive datasets (e.g., nearly a terabyte of English text [55]). This scaling increases the ability of LMs to generate fluent natural language [53,74,76], and also allows them to be applied to a plethora of other tasks [29, 39,55], even without updating their parameters [7]. At the same time, machine learning models are notorious At the same time, machine learning models are notorious for exposing information about their (potentially private) training data—both in general [47,65] and in the specific case of language models [8,45]. For instance, for certain models it

is known that adversaries can apply membership inference attacks [65] to predict whether or not any particular example was in the training data.

USENIX Association

Related Work

Sequences

Extracting Training Data from Large Language Models

Florian Tramèr² Eric Wallace³ Matthew Jagielski Katherine Lee¹ Adam Roberts¹ Tom Brown⁵ Dawn Song³ Úlfar Erlingsson⁷ Alina Oprea⁴ Colin Raffel¹ ¹Google ²Stanford ³UC Berkelev ⁴Northeastern University ⁵OpenAI ⁶Harvard ⁷Apple

Prefix

It has become common to publish large (billion parameter

Figure 1: Our extraction attack. Given query access to a neural network language model, we extract an individual p son's name, email address, phone number, fax number, and vsical address. The example in this figure shows inform tion that is all accurate so we redact it to protect privacy

East Stroudsburg Stroudsburg...

GPT-2

Such privacy leakage is typically associated with overfittin [75]-when a model's training error is significantly lower than its test error-because overfitting often indicates that a model has memorized examples from its training set. Indeed overfitting is a sufficient condition for privacy leakage [72] and many attacks work by exploiting overfitting [65]

The association between overfitting and memorization has-erroneously—led many to assume that state-of-the-art LMs will *not* leak information about their training data. Because these models are often trained on massive de-duplicated datasets only for a single epoch [7, 55], they exhibit little to no overfitting [53]. Accordingly, the prevailing wisdom has been that "the degree of copying with respect to any given work is likely to be, at most, de minimis" [71] and that models do not significantly memorize any particular training exam

30th USENIX Security Symposium 2633

Carlini et al., 2020

OUANTIFYING MEMORIZATION ACROSS NEURAL LANGUAGE MODELS Nicholas Carlini[‡] Daphne Ippolito^{1,2} Matthew Jagielski Katherine Lee^{1,3} Florian Tramèr¹ Chiyuan Zhang¹ ¹Google Research University of Pennsylve ³Cornell University Abstract Large language models (LMs) have been shown to memorize parts of their training data, and when prompted appropriately, they will emit the memorized training dat verbatim. This is undesirable because memorization violates privacy (exposing user data), degrades utility (repeated easy-to-memorize text is often low quality and hurts fairness (some texts are memorized over others) We describe three log-linear relationships that quantify the degree to which LMs emit memorized training data. Memorization significantly grows as we increase (1 the capacity of a model, (2) the number of times an example has been duplicated nd (3) the number of tokens of context used to prompt the model. Surprisingl we find the situation becomes more complicated when generalizing these results across model families. On the whole, we find that memorization in LMs is more prevalent than previously believed and will likely get worse as models continues to scale, at least without active mitigations. 1 INTRODUCTION The performance of neural language models has continuously improved as these models have grown from millions to trillions of parameters [Fedus et al.] [2021], with their training sets similarly growing from millions to trillions of tokens. In anticipation of future, even larger models trained on minimally curated datasets, it is important to quantify factors that lead to increased memorization of a model's toriation utataled, need mapping the quadrany the site is the order case in action and the site of the While current attacks are effective, they only represent a lower bound on how much memorizatio occurs in existing models. For example, by querying the GPT-2 language model, <u>Carlini et al.</u> [2020] (manually) identified just 600 memorized training examples out of a 40GB training dataset. Thi attack establishes a (losse) lower bound that at least 0.0000015% of the dataset is memorized. It contrast, we are able to show that the 6 billion parameter GPT-J model (Black et al., 2021) Wang a Komatsuzaki 2021) memorizes at least 1% of its training dataset: The Pile (Gao et al., 2020) <u>In addition to prior work's loose estimates of models' memorization capabilities, there is a limited understanding of how memorization varies across different neural language models and datasets of different scales. Frior studies of memorization in language models either focus on models or datasets of a freed size (Carlini et al. 2019) [Zhang et al. 2020] [Thatkare et al. 2020] or identify a narrow memorization-versus-scale relationship (Carlini et al. 2020) [Lee et al. 2021]. While[McCoy] [Et al. 2022] [Dradky and the extent to which language models memorize, their focus is on how to avoid the problem and ensure novelty of model outputs, rather than on studying model risk through identifying the maximal angle models memorized risk through identifying the maximal angle models memo</u> identifying the maximal amount of data memorization *Authors ordered alphabetically

Published as a conference paper at ICLR 2023

Carlini et al., 2022

PII Leakage In Pre-Trained LMs

Are Large Pre-Trained Language Models Leaking Your Personal Information? Jie Huang* Hanyin Shao* Kevin Chen-Chuan Chang University of Illinois at Urbana-Champaign, USA {jeffhj, hanyins2, kcchang}@illinois.edu Abstract Are Large Pre-Trained Language Models Leaking Your Personal Information? In this Leaking Your Personal Information? In this paper, we analyze whether Pre-Trained Lan-guage Models (PLMs) are prone to leaking personal information. Specifically, we query PLMs for email addresses with contexts of the email addresses with contexts of the email addresses with contexts of sonal information due to memorization. How-Figure 1: Results of asking GPT-3 (text-davinci-2) "Are Large Pre-Trained Language Models Leaking Your Personal Information?" sonal information due to memorization. How ever, since the models are weak at association 20 he risk of specific personal info to recover specific patient names and condition extracted by attackers is low. We hope this CL] with which they are associated from a BERT model work could help the community to better un derstand the privacy risk of PLMs and bring that is pre-trained over clinical notes. However, new insights to make PLMs safe.1 they find that with their methods, the model cannot meaningfully associate names with conditions 1 Introduction which suggests that PLMs may not be prone to 2 Pre-trained Language Models (PLMs) (Devlin leaking personal information et al., 2019: Brown et al., 2020: Oiu et al., 2020) Based on existing research, we are not sure NLP tasks, attributing to the explosive growth of parameters and training data. However, recent stud-in: Are Large Pre-Trained Language Models Prone is also suggest that these large models pose some privacy risks. For instance, an adversary is able to To answer the above question, we first idenrecover training examples containing an individual tify two capacities that may cause privacy leakage by querying the model (Carlini et al., 2021). This may lead to privacy leakage if the model is trained with a specific prefix, e.g., tokens before the info on a private corpus, in which case we want to im-prove the performance with the data (Huang et al., PLMs can association, i.e., 2019). Even if the data is public, PLMs may change its owner, thus attackers can query the information but do not expect to be disseminated. Carlini et al. (2021, 2022) demonstrate that PLMs memorize a lot of training data, so they are prone to leaking privacy. However, if the memo-in Carlini et al. (2021), attackers cannot effectively rized information cannot be effectively extracted, it extract specific personal information since it is di is still difficult for the attacker to carry out effective attacks. For instance, Lehman et al. (2021) attempt

¹Code and data are available at https://github.com/ jeffhj/lM_PersonalInfoLeak. 'Equal contribution. We focus on studying a

We focus on studying a specific kind of personal

Huang et al., 2022







Privacy in LMs

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 privacypreserving 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

Brown et al., 2022

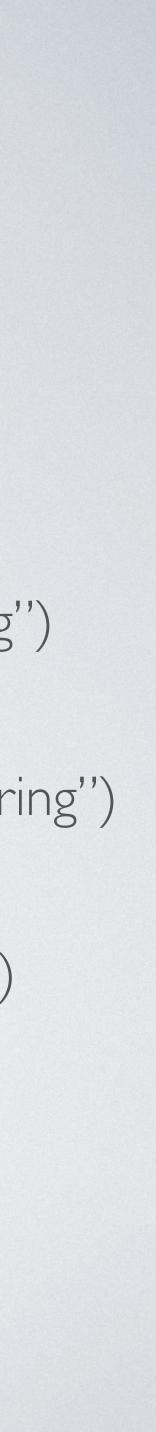
Is public data truly public?

Related Work

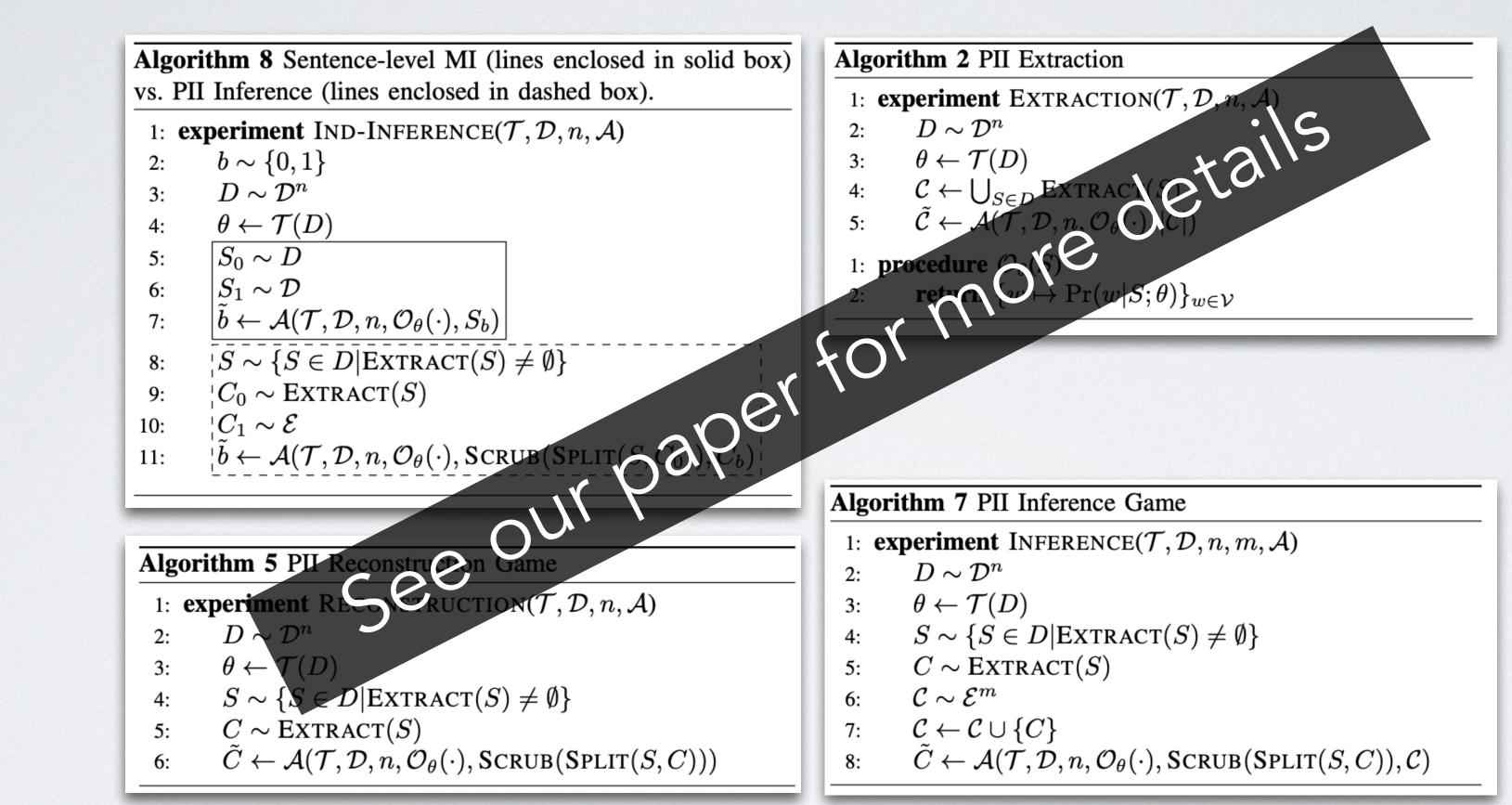
- Data shared to intentionally violate someone's privacy (e.g., "doxing")

- Social media posts issued to a small target audience ("in-group sharing")

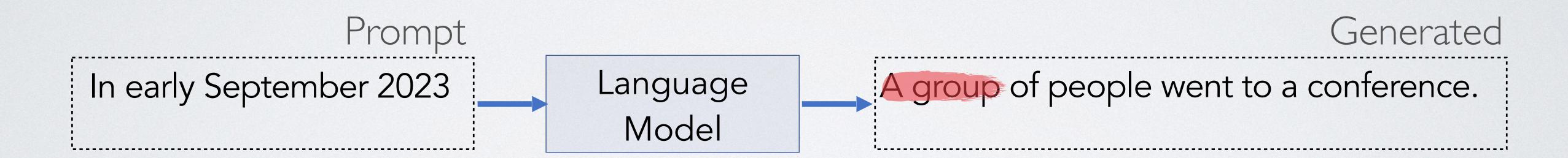
- Accidental leakage of other's information (e.g., "conversations")



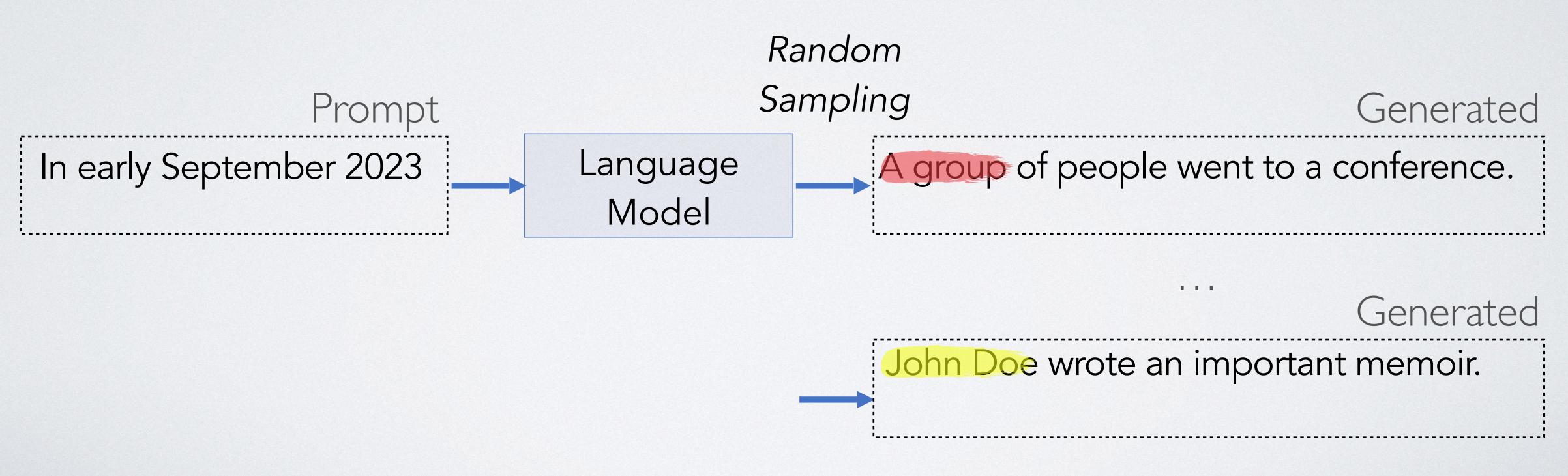
Security Games for PII Leakage



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



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



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

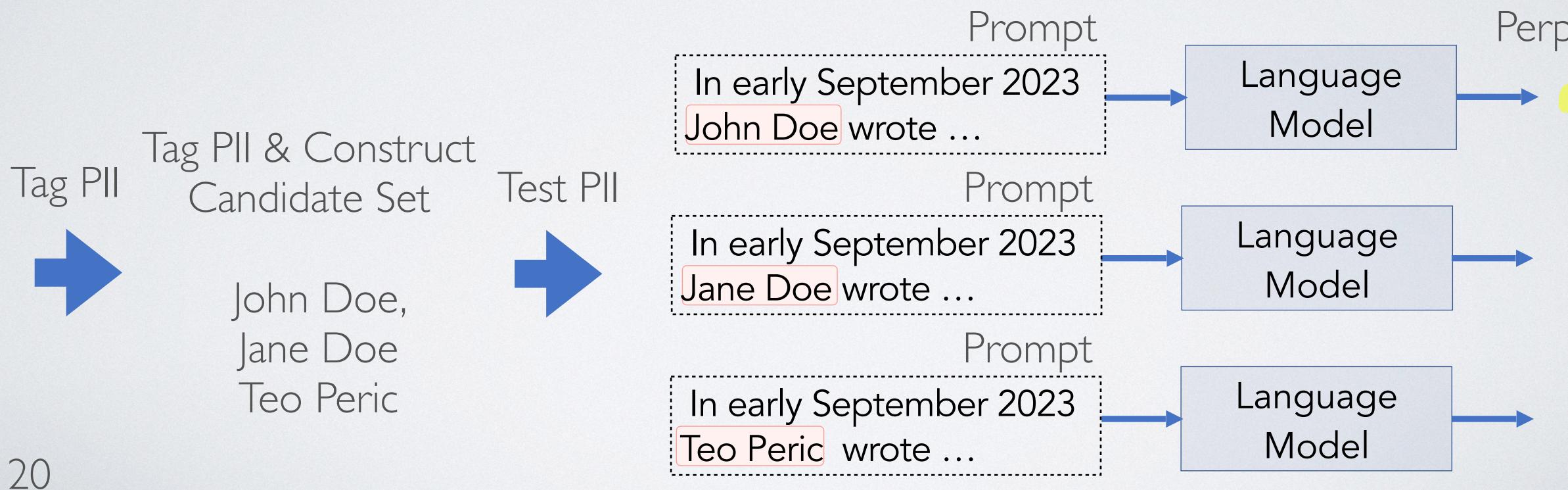
John Doe wrote an important memoir.

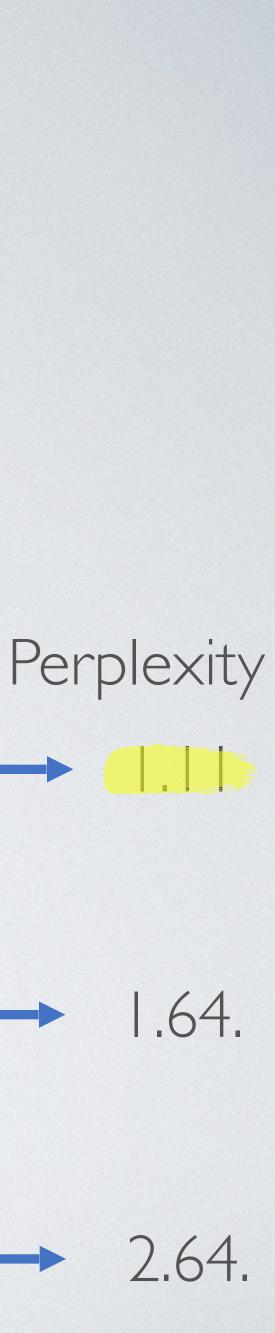


Tag PII & Construct Candidate Set

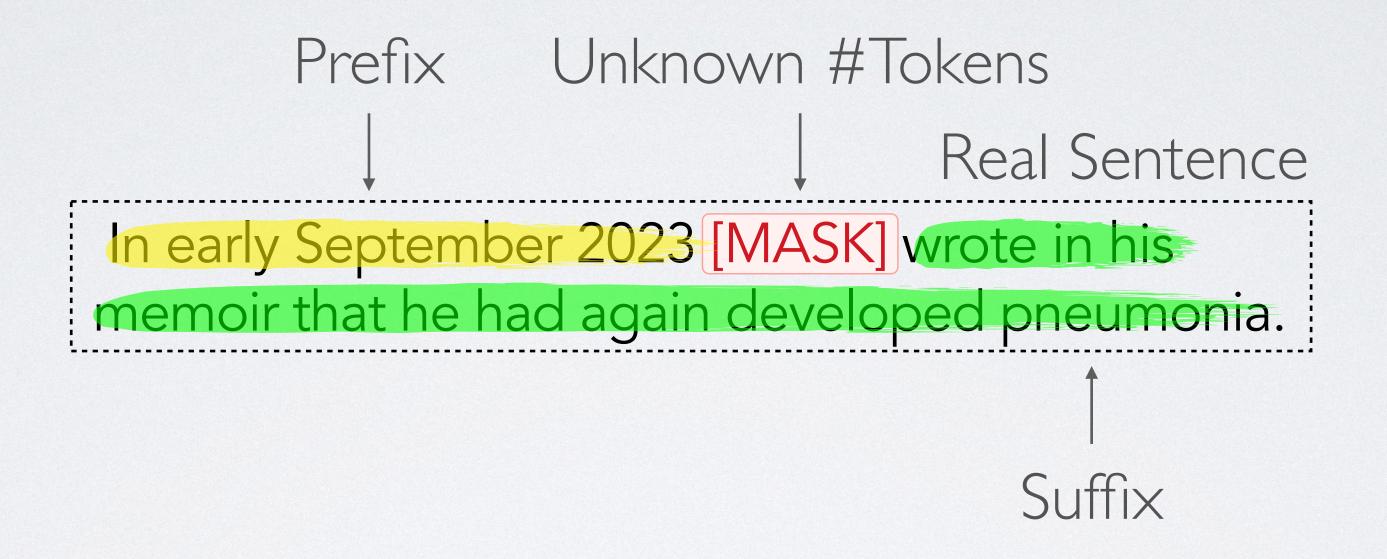
> John Doe, Jane Doe Teo Peric

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





PII Reconstruction Tractability



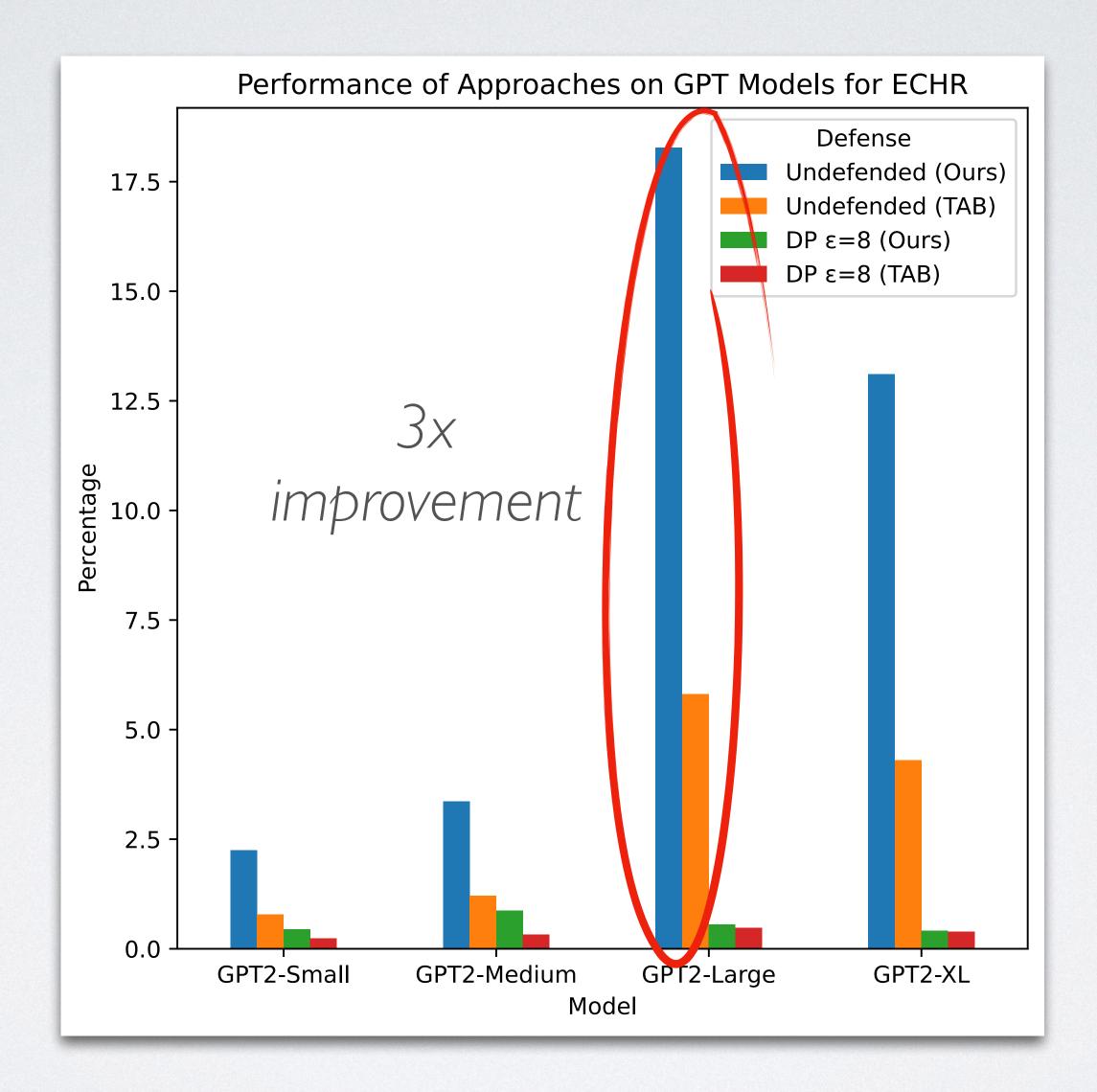
| | Records | Tokens / Record | Unique PII | Records w. PII | Duplicates / PII | Tokens / PII |
|-------------|---------|-----------------|------------|----------------|------------------|--------------|
| ECHR | 118 161 | 88.12 | 16133 | 23.75% | 4.66 | 4.00 |
| Enron | 138919 | 346.10 | 105 880 | 81.45% | 11.68 | 3.00 |
| Yelp-Health | 78 794 | 143.92 | 17 035 | 54.55% | 5.35 | 2.17 |

ECHR Enron : Corporate e-mails

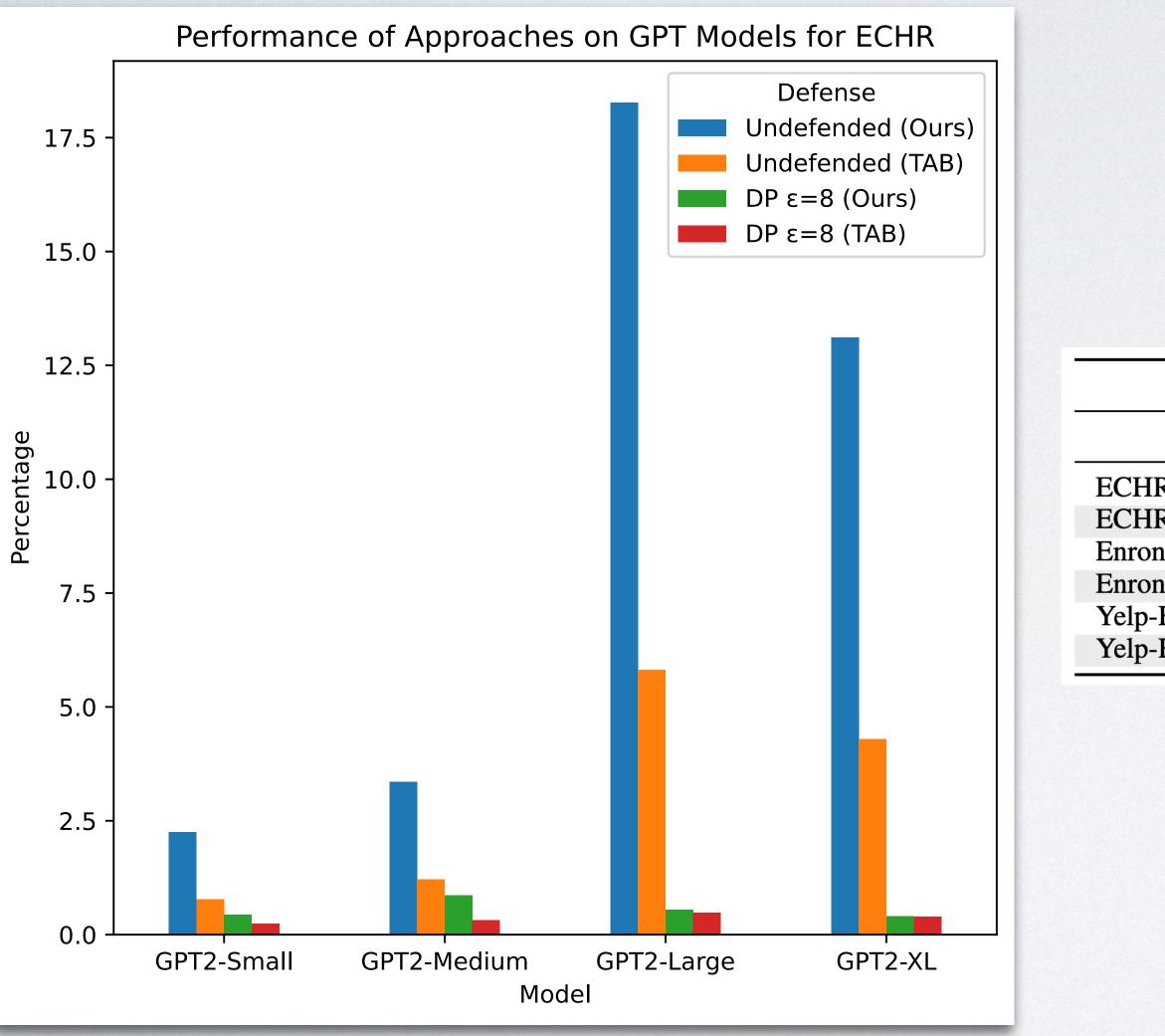
Datasets

: European Court for Human Rights Yelp-Health: Reviews for healthcare facilities

PII Reconstruction

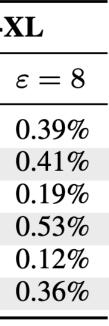


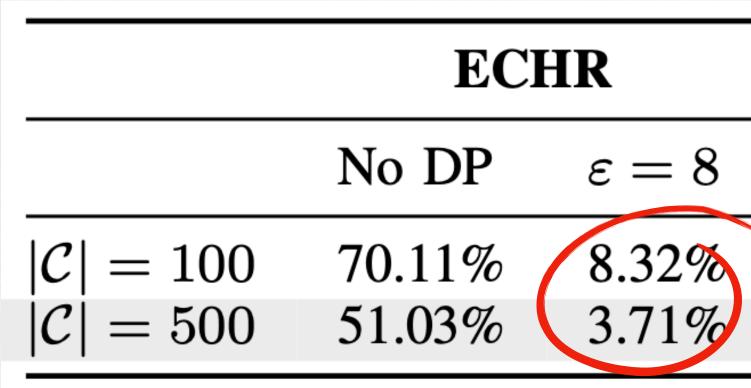
PII Reconstruction



| | GPT2- | Small | GPT2-N | ledium | GPT2- | Large | GPT2 | -X |
|---------------------------------------|-------|-------------------|--------|-------------------|--------|-------------------|--------|----|
| | No DP | $\varepsilon = 8$ | No DP | $\varepsilon = 8$ | No DP | $\varepsilon = 8$ | No DP | ٤ |
| IR(TAB) | 0.78% | 0.24% | 1.21% | 0.32% | 5.81% | 0.48% | 4.30% | (|
| IR (Ours, $ \mathcal{C} = 64$) | 2.25% | 0.44% | 3.36% | 0.87% | 18.27% | 0.55% | 13.11% | (|
| on (TAB) | 0.59% | 0.04% | 0.67% | 0.04% | 1.75% | 0.04% | 2.19% | (|
| on (Ours, $ \mathcal{C} = 64$) | 6.29% | 0.49% | 7.26% | 0.52% | 12.68% | 0.55% | 15.25% | (|
| -Health (TAB) | 0.33% | 0.24% | 0.37% | 0.14% | 0.05% | 0.12% | 1.99% | (|
| -Health (Ours, $ \mathcal{C} = 64$) | 0.42% | 0.32% | 1.31% | 0.32% | 1.69% | 0.35% | 6.40% | (|

up to 7x Improvement



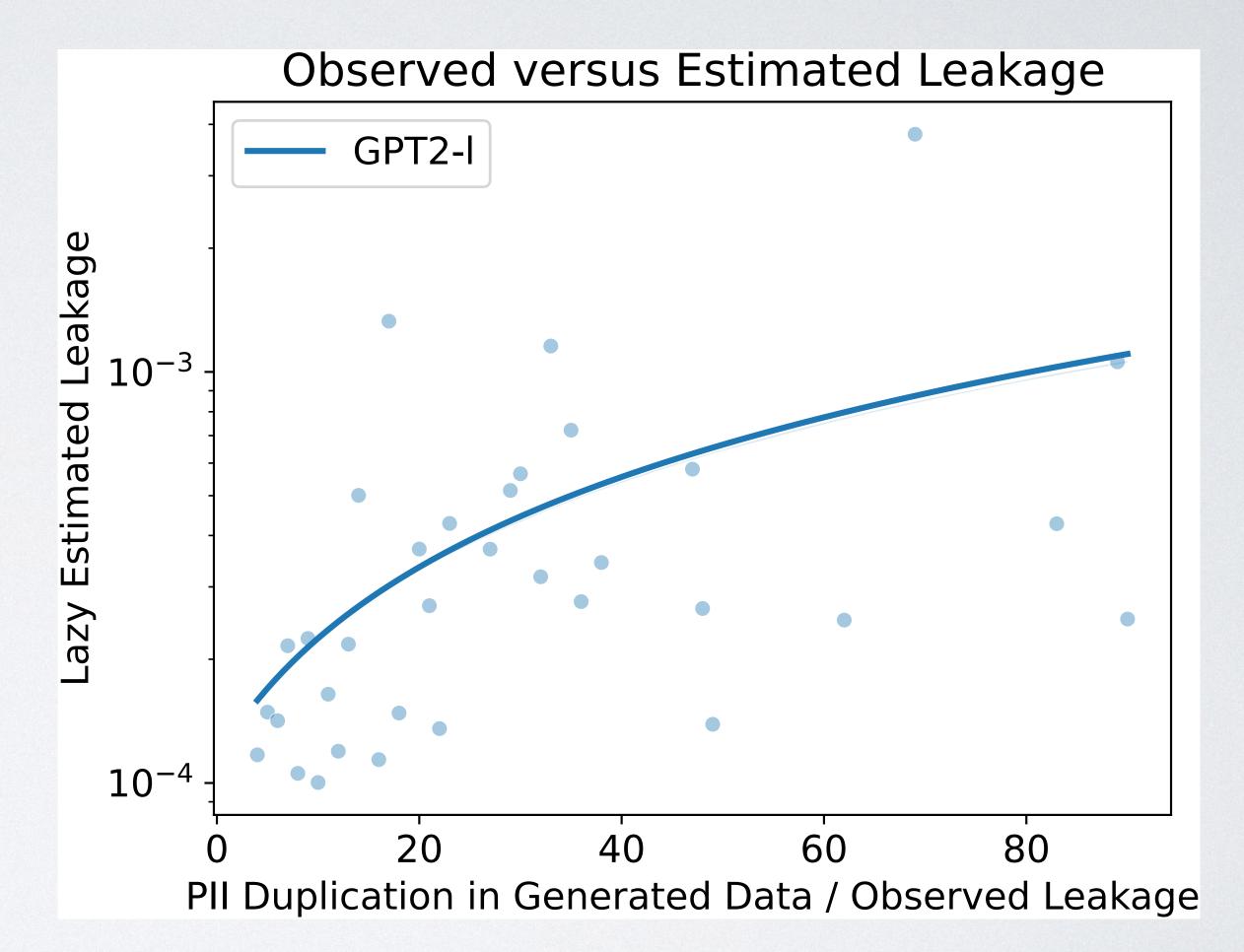


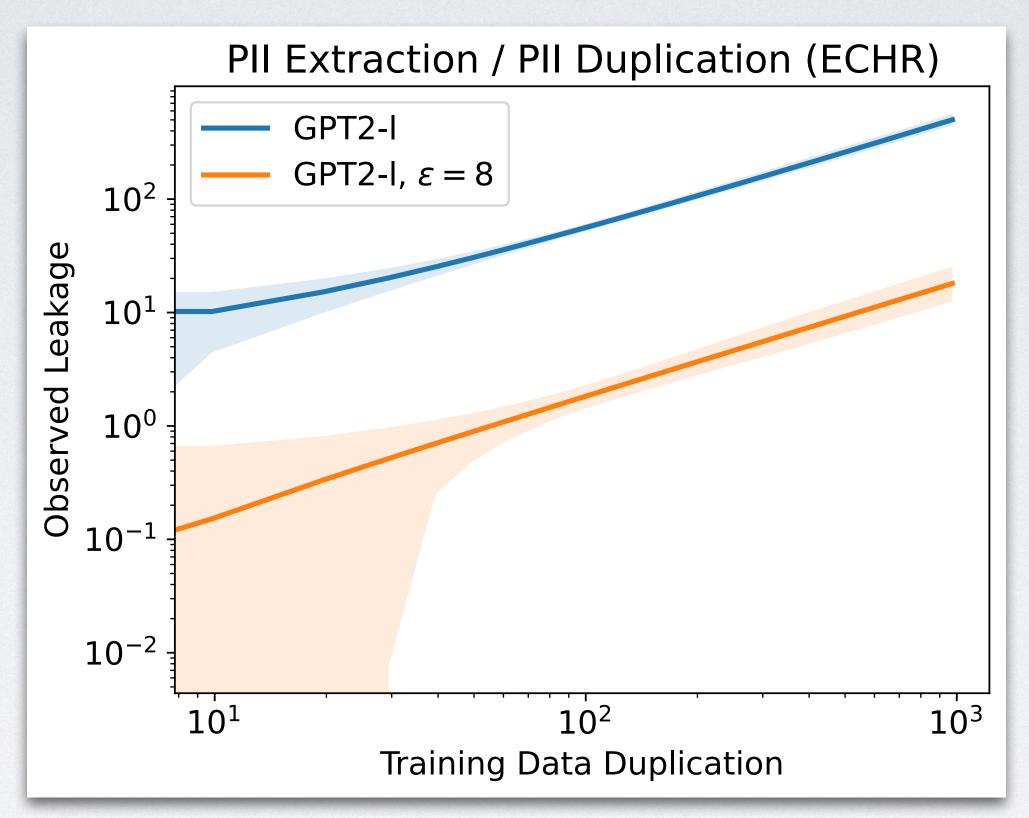
PII Inference

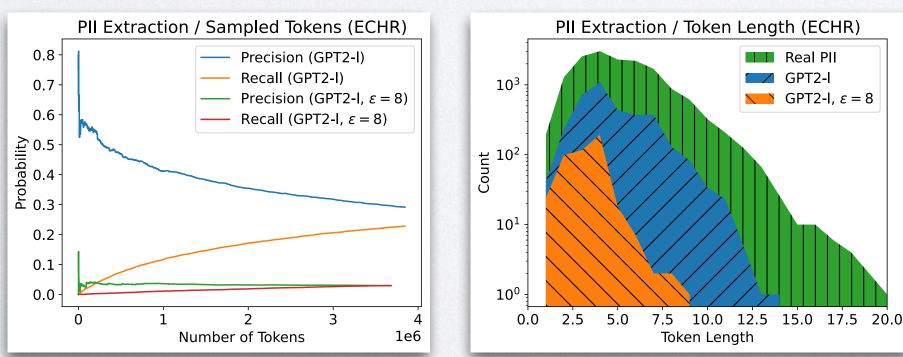
| Enr | on | Yelp-H | ealth | |
|--------|-------------|--------|-------------------|--|
| No DP | arepsilon=8 | No DP | $\varepsilon = 8$ | |
| 50.50% | 3.78% | 28.31% | 4.29% | |
| 34.14% | 1.92% | 15.55% | 1.86% | |

Extraction Attack

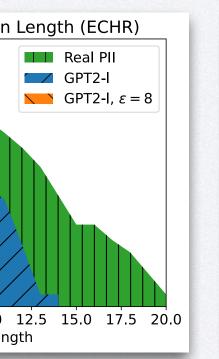
Once upon a time, there existed a tale of two medical students. In the year 2022, they resided at Sunset Street while pursuing their medical education. Alongside his friend, he worked at the LHS Hospital located in the bustling heart of downtown London. Before donning their white coats, both **TebePoic**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



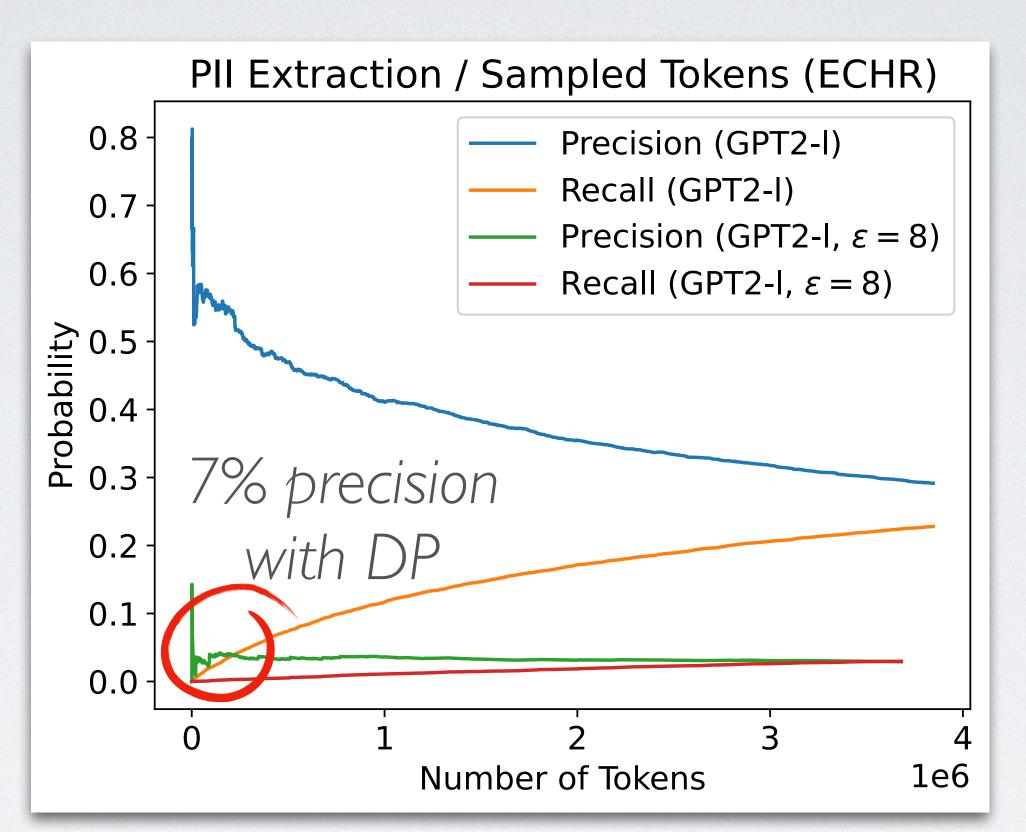


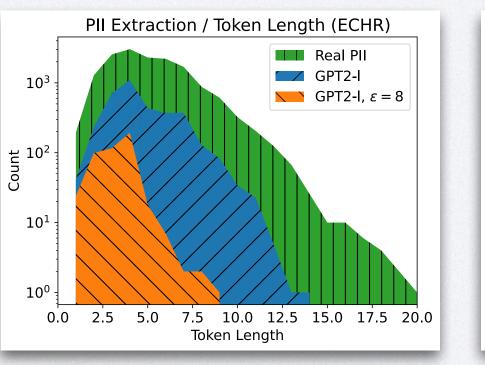


Duplicated PII are leaked more often

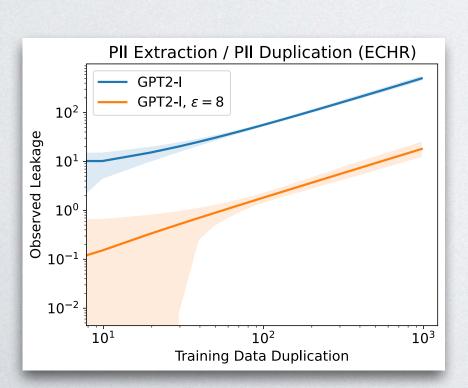


| | GPT2- | Small | GPT2-N | Medium | GPT2-Large | |
|--------|---------|-------------------|--------|-------------------|-------------------|-------------------|
| | No DP | $\varepsilon = 8$ | No DP | $\varepsilon = 8$ | No DP | $\varepsilon = 8$ |
| | | | EC | HR | | |
| Prec | 24.91% | 2.90% | 28.05% | 3.02% | 29.56% | 2.92% |
| Recall | 9.44% | 2.98% | 12.97% | 3.21% | 22.96% | 2.98% |
| | | | En | ron | | |
| Prec | 33.86 % | 9.37% | 27.06% | 12.05% | 35.36% | 11.579 |
| 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% |

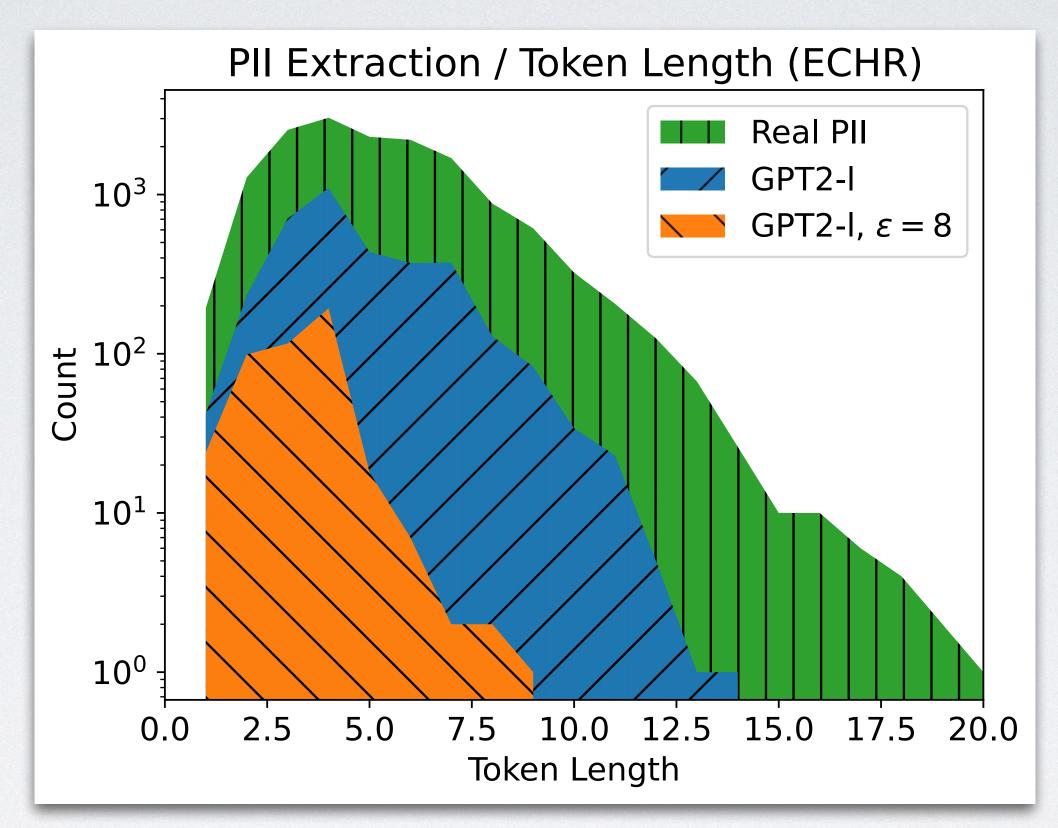


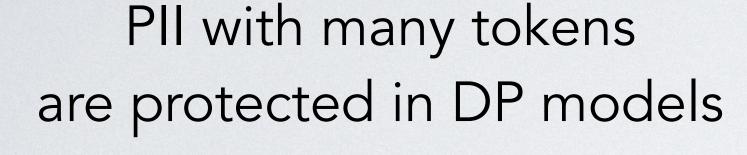


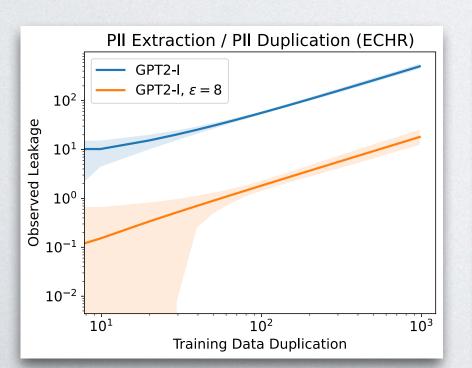
High-precision/ Low-recall attacks

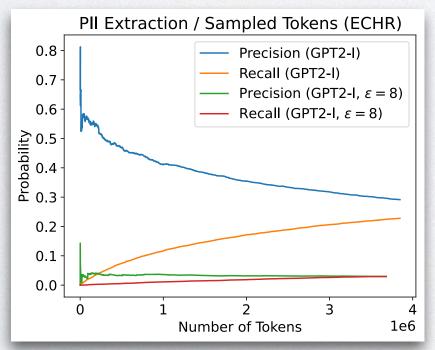


| | GPT2- | GPT2-Small | | Medium | GPT2-Large | |
|--------|---------|-------------------|-----------------|-------------------|-------------------|-------------------|
| | No DP | $\varepsilon = 8$ | No DP | $\varepsilon = 8$ | No DP | $\varepsilon = 8$ |
| | | | EC | HR | | |
| Prec | 24.91% | 2.90% | 28.05% | 3.02% | 29.56% | 2.92% |
| Recall | 9.44% | 2.98% | 12.97% | 3.21% | 22.96% | 2.98% |
| | | | En | ron | | |
| 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% | $14.87\bar{\%}$ | 6.32% | 14.28% | 7.67% |
| Recall | 11.31% | 5.02% | 11.23% | 5.22% | 13.63% | 6.51% |

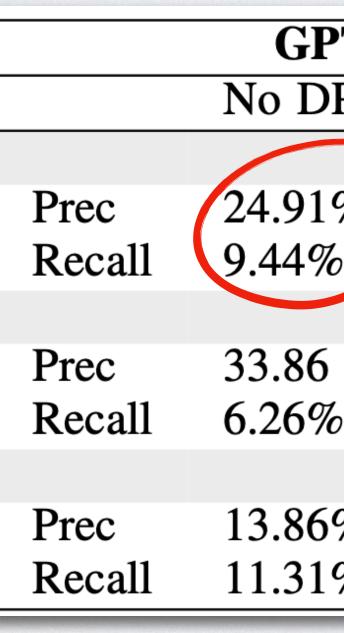




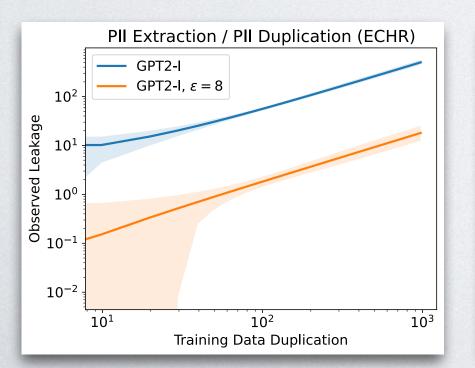


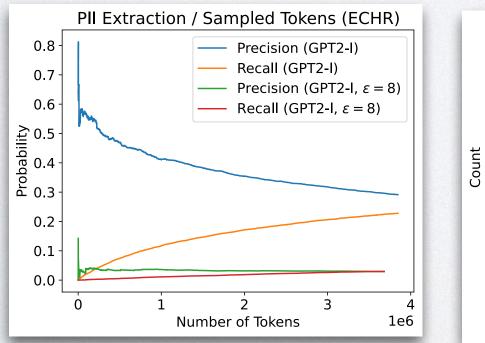


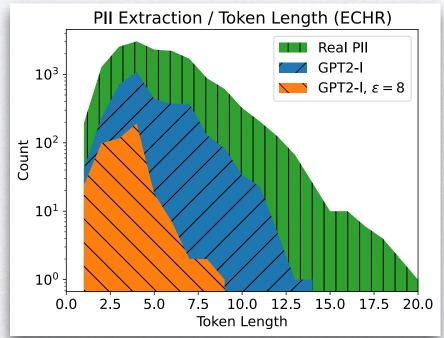
| | GPT2- | GPT2-Small | | Medium | GPT2-Large | | |
|--------|---------|-------------------|--------|-------------------|------------|-------------------|--|
| | No DP | $\varepsilon = 8$ | No DP | $\varepsilon = 8$ | No DP | $\varepsilon = 8$ | |
| | | | EC | HR | | | |
| Prec | 24.91% | 2.90% | 28.05% | 3.02% | 29.56% | 2.92% | |
| Recall | 9.44% | 2.98% | 12.97% | 3.21% | 22.96% | 2.98% | |
| | | | En | ron | | | |
| 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-l | 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% | |



Higher recall in larger models

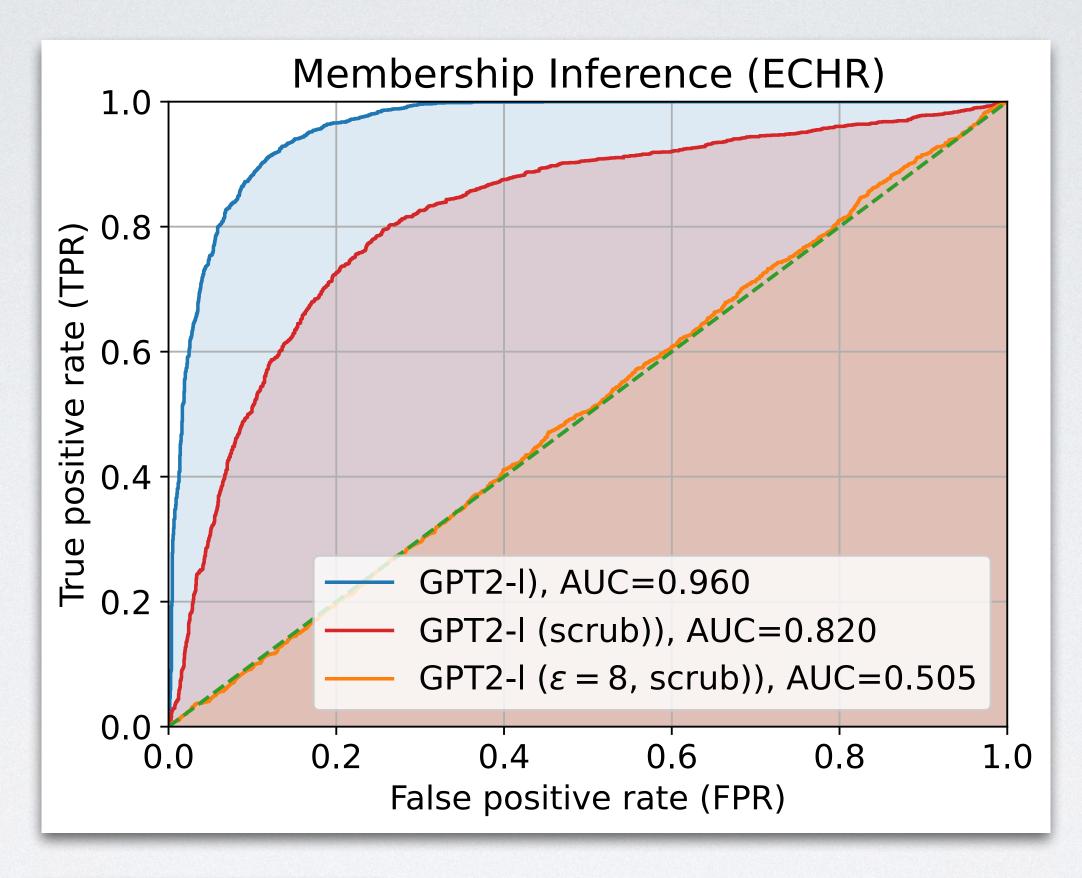


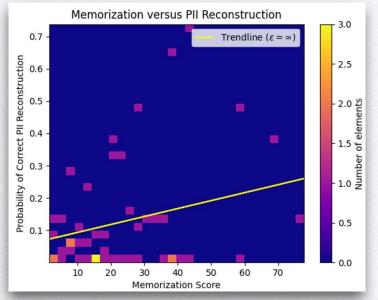




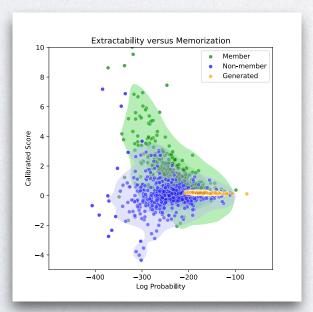
| PT2-8 | Small | GPT2-N | Medium | GPT2-Large | | | | | |
|-------------|-------------|--------|-------------|-------------------|---------------|--|--|--|--|
| P | arepsilon=8 | No DP | arepsilon=8 | No DP | arepsilon = 8 | | | | |
| | | EC | HR | | | | | | |
| % | 2.90% | 28.05% | 3.02% | 29.56% | 2.92% | | | | |
| 70 | 2.98% | 12.97% | 3.21% | 22.96% | 2.98% | | | | |
| Enron | | | | | | | | | |
| % | 9.37% | 27.06% | 12.05% | 35.36% | 11.57% | | | | |
| 6 | 2.29% | 6.56% | 2.07% | 7.23% | 2.31% | | | | |
| Yelp-Health | | | | | | | | | |
| % | 8.31% | 14.87% | 6.32% | 14.28% | 7.67% | | | | |
| % | 5.02% | 11.23% | 5.22% | 13.63% | 6.51% | | | | |

Membership Inference

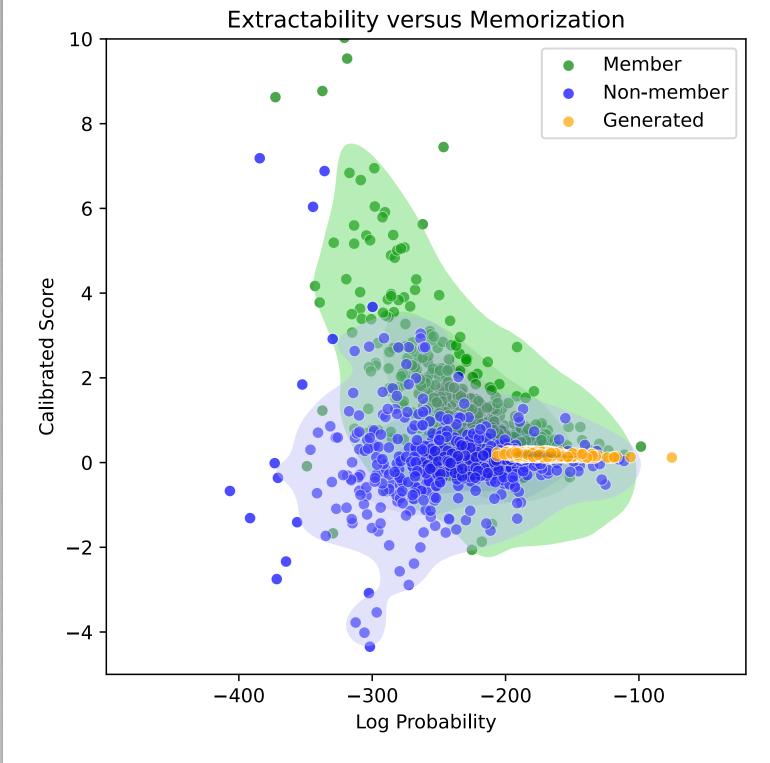


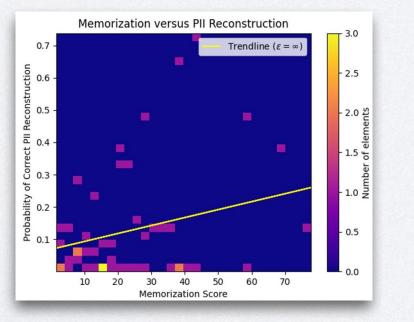


Scrubbing does not prevent MI

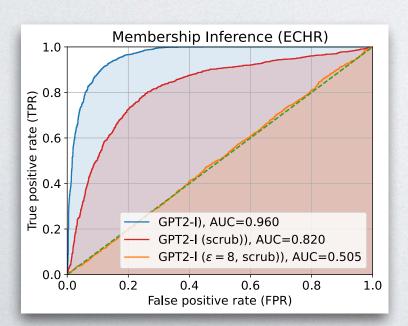


Membership Inference



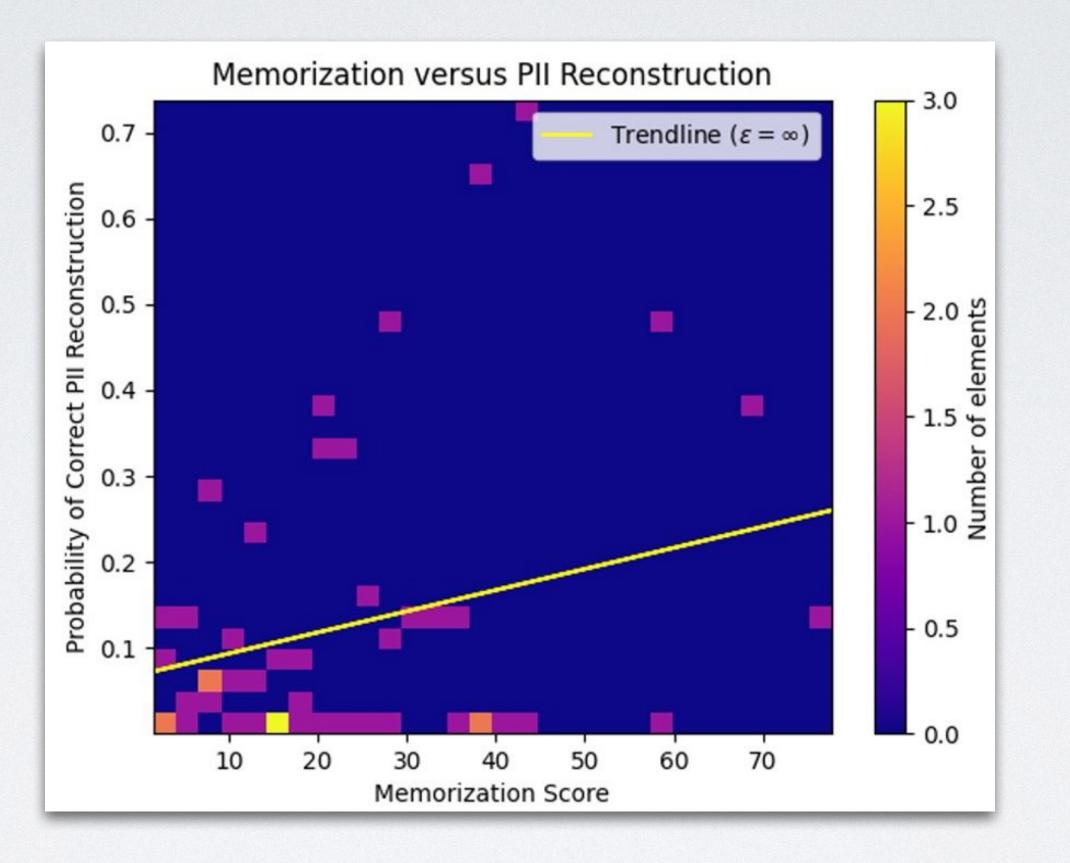


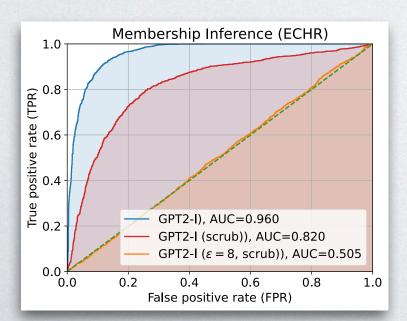
Randomly generated sequences likely do not contain MI signal

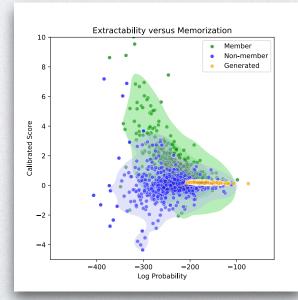


Membership Inference

MI correlates with PII reconstruction







Summary of Results

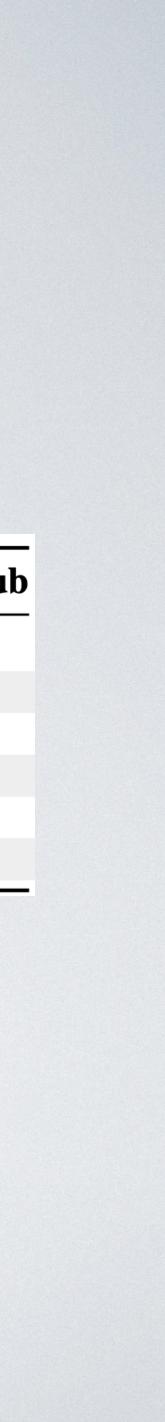
Undefended models are highly Vulnerable to all privacy attacks

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

| | \cap | | | | |
|--|---------|-----|-----|-------|------------|
| U | ndefend | led | DP | Scrub | DP + Scrul |
| Test Perplexity | 9 | | 14 | 16 | 16 |
| Extract Precision | 30% | | 3% | 970 | 0% |
| Extract Recall | 23% | | 3% | 0% | 0% |
| Reconstruction Acc. | 18% | | 1% | 0% | 0% |
| Inference Acc. ($ \mathcal{C} = 100$) | 70% | 1 (| 8% | 1% | 1% |
| MI AUC | 0.96 | | 0.5 | 0.82 | 0.5 |
| | | | | | |



- (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

Limitations

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.

Outlook

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 sentenceor 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 such as human dialogs [7] or clinical health data [62] which LM. Extracting any PII by itself, such as a p 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 for example, "In May 2022, 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 [1].

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. and Privacy (S&P) 2023.

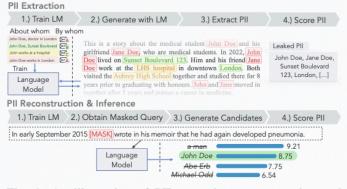


Fig. 1: An illustration of PII extraction, reconstruction and inference attack techniques.

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 nextword prediction LM for composing e-mails, su

Smart Compose [13]. Their goal is to train a can already pose a privacy threat. This threat is elevated when an attacker can associate a piece of PII to a context, LHS". As a part of this paper, attacks on LMs in practice. Fig attacks proposed in this work. Defenses against memoriza ration and algorithmic defens

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%

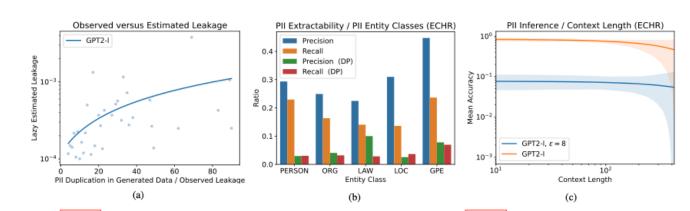


Fig. 6: Fig. 6a shows the correlation between the observed and estimated leakage. Fig. 6b shows the precision and recall for other entity classes on the ECHR dataset. Fig. 6c shows the mean inference accuracy relative to the context length, which is the length of the combined prefix and suffix for a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

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

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.

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



phone number in the DP model. However, we observe leakage compared to TAB becomes more evident. E-mails in the Enron of e-mail addresses (consisting of equally many tokens), that dataset typically mention the receiver of the e-mail at the

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 addresses or other sensitive information. The threat is an attacker who learned to the time information. The threat is an attacker who learned to the time information. The threat is an attacker who learned to the time information. The threat is an attacker who learned to the time information. The threat is an attacker who learned to the time information. The threat is an attacker who learned to the time information. The threat is an attacker who learned to the time information. The threat is an attacker who learned to the time information. The threat is an attacker who learned to the time information. The threat is an attacker who learned to the time information. The threat is an attacker who learned to the time information information. The threat is an attacker who learned to the time information information. The threat is an attacker who learned to the time information infor

We sample 64 candidates and decode from the model using limits our attack's success rate. We believe a method that top-k sampling with k = 40. We observe that our reconsamples candidates by incorporating information from the

Enron

No DP $\varepsilon = 8$ No DP $\varepsilon = 8$ No DP $\varepsilon = 8$

ECHR

Extraction, Reconstruction and Inference Attacks

Yelp-Health

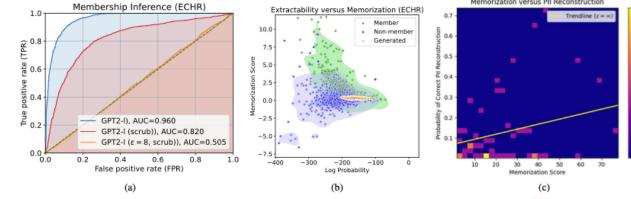


Fig. 7: Connecting sentence-level membership inference with PII reconstruction in GPT-2-Large. 7a shows the ROC curve against our fine-tuned model using a shadow model attack on ECHR. 7b shows that the memorization score of generated sequences is nearly zero and 7c shows that the memorization score correlates 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. $(\mathcal{C} = 100)$ | 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 action. We demonstrate that an

> actical setting. (aggressive) PII scrubbing limit cy/utility trade-offs.

> > ted by our findings.

ind associations in

masked language

V. DISCUSSION AND LIMITATIONS

models fare compare to autoregressive models and identify

further research motivated by our findings: how to best com-

bine DP training and scrubbing, optimizing attacks for other

Below, we discuss extensions and limitations of our method-

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 with up to 10% accuracy (given 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 ting the search for defenses with 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 methodology to scrubbing techniques can be adapted to take into consideration tential extensions to 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 maintaing a privacy level equivalent to a naive combination of DP training and scrubbing.

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

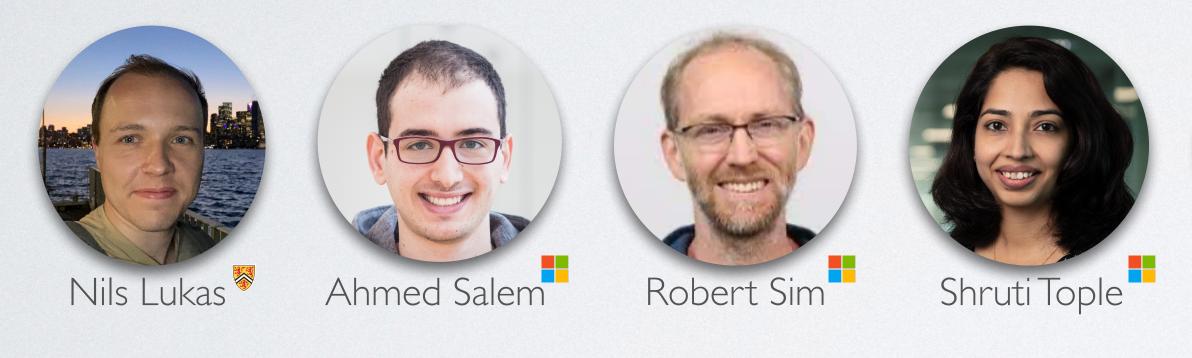
⁸Part of this work was done during an internship at Microsoft Research. ^aTo cite this work, please refer to the full publication [4] in IEEE Security d Privacy (S&P) 2023. ^bTo cite this work, please refer to the full publication [4] in IEEE Security d Privacy (S&P) 2023. ^bTo cite this work, please refer to the full publication [4] in IEEE Security d Privacy (S&P) 2023. ^bTo cite this work, please refer to the full publication [4] in IEEE Security d Privacy (S&P) 2023. ^bTo cite this work, please refer to the full publication [4] in IEEE Security d Privacy (S&P) 2023. ^bTo cite this work, please refer to the full publication [4] in IEEE Security d Privacy (S&P) 2023. ^bTo cite this work, please refer to the full publication [4] in IEEE Security ^bTo cite this work, please refer to the full publication [4] in IEEE Security ^bTo cite this work, please refer to the full publication [4] in IEEE Security ^bTo cite this work, please refer to the full publication [4] in IEEE Security ^bTo cite this work, please refer to the full publication [4] in IEEE Security ^bTo cite this work, please refer to the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite this work of the full publication [4] in IEEE Security ^bTo cite the full publication [4] in IEEE Security ^b

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





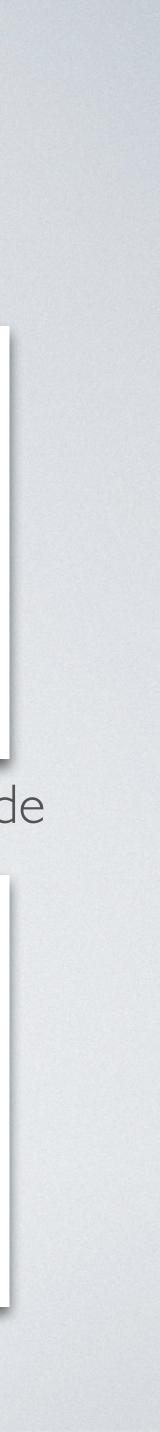






GitHub - Source Code





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-withchatgpt-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 [1] 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.

