Privacy in Language Models

Reza Shokri

Data Privacy and Trustworthy ML Research Lab
National University of Singapore (NUS)
Direct Privacy Risks in Machine Learning

Direct Access to Sensitive Data

Training Set $X$ to $W$ to $x$ through $f(x; W)$
Indirect Privacy Risks in Machine Learning

Indirect Leakage about $X$
Language Models
There is a cat on the table.
There is a cat on the floor.
There is a cat on the chair.
There is a cat on the window sill.
There is a cat on the roof.
There is a cat on the counter.
There is a cat on the bed.
There is a cat on the porch.
There is a cat on the fence.
There is a cat on the car.
Sources
LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT THE DOCKS AT MIDNIGHT ON JUNE 28.

AHA, FOUND THEM!

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.
Real World Attacks against Large Language Models

Members of the training set are identifiable: Presence of any document in a training dataset can be inferred very accurately.

Real World Attacks against Large Language Models

[Carlini, Tramer, et al.] Extracting Training Data from Large Language Models, Usenix security’21
Models are **personal data**
We need a standard method for auditing data privacy in machine learning systems
A world in which data $d$ is not part of training set

Alternative world where data $d$ is part of training set
A world in which data $d$ is **not** part of training set

Alternative world where data $d$ is **is** part of training set
A world in which data $d$ is not part of training set

Alternative world where data $d$ is part of training set
A world in which data $d$ **is not** part of training set

Alternative world where data $d$ **is** part of training set

Can an adversary tell the difference between these two worlds?
Membership Inference

• Given a model, can an adversary infer whether a particular data point is part of its training set?

• Success of attacker is a metric for privacy loss

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP’17
Success rate of adversary indicates information leakage of models about their training data
[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS’22
Membership Inference

Adversary identifies them as 'Member'

[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS’22
Membership Inference

Adversary identifies them as 'Member'

[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS’22
Membership Inference

Privacy Risk Profile

True Positive Rate

False Positive Rate

Adversary identifies them as 'Member'

[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS’22
Actionable Assessment of Privacy Risks

Towards left: risk of being singled out  
Towards right: more anonymity

The fraction of private dataset that are at risk

- **High risk**: Do not deploy
- **Medium risk**: Requires further investigations
- **Low risk**: Safe to deploy
AI Regulations and Guidelines

A Taxonomy and Terminology of Adversarial Machine Learning

Elham Tabassi
Kevin J. Burns
Michael Hadjimichael
Andres D. Molina-Markham
Julian T. Sexton

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8269-draft

Guidance on the AI auditing framework
Draft guidance for consultation

White Paper on Artificial Intelligence: a European approach to excellence and trust

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

November 17, 2020

THE DIRECTOR
M-21-06

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Russell T. Vought
Director

SUBJECT: Guidance for Regulation of Artificial Intelligence Applications
• “… membership inferences show that AI models can inadvertently contain personal data”

• “Attacks that reveal confidential information about the data include membership inference …”

• “… should consider the risks to data throughout the design, development, and operation of an AI system”
Privacy Meter is an open source tool that enables quantifying the privacy risks of statistical and machine learning models.
Example: Generative Model

Chance of correctly inferring if an author’s data was used in training the model

(Speaker Annotated TED talks)
SATED dataset
Examples of Vulnerable Training Data

But it gets worse. And this is very important, what I' generic. It doesn't have anything to do, in specifics, would work as well, for example, in a power plant or don't have -- as an attacker -- you don't have to del the case of Stuxnet. You could also use convention:

This year, Germany is celebrating the 25th anniversary 1989, the Communist regime was moved away, the Bi German Democratic Republic, the GDR, in the East was in the West to found today's Germany. Among many of the East German secret police, known as the Stasi. Old were opened to the public, and historians such as me about how the GDR surveillance state functioned.
We need to “Preserve” Privacy in Language Models
What does it mean to “Preserve” Privacy in Language Models?
Privacy in the Realm of Language

• Language serves as a medium for communication

• Natural language reflects our private lives and identities

• Privacy concerns in natural language are as broad as those of real life
Challenges of Preserving Privacy in Language

• Privacy is *contextual*

• We (as humans) choose whether to disclose or conceal information, based on the context.

• Detecting the context in which text is generated can be challenging (to algorithms) or may exist beyond the text itself.
Challenges of Preserving Privacy in Language

- Private information can take many different forms.
- Private information has no boundaries: information about Alice could be found in Bob’s text data.
- Repeated information can still be private information.
Existing Privacy Preserving Methods
Existing Privacy Preserving Methods

• Scrubbing: Remove secrets from text (before training a model)
Existing Privacy Preserving Methods

- Scrubbing: Remove secrets from text (before training a model)

- Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models
Existing Privacy Preserving Methods

• Scrubbing: Remove secrets from text (before training a model)

• Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models

• Use public data: Crowl the internet to gather all available text
Existing Privacy Preserving Methods

• Scrubbing: Remove secrets from text (before training a model)

  Assumption: Every secret can be found through certain rules.

• Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models

• Use public data: Crowl the internet to gather all available text
Existing Privacy Preserving Methods

• Scrubbing: Remove secrets from text (before training a model)
  Assumption: Every secret can be found through certain rules.

• Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models
  Assumption: A secret is entirely in one user’s data

• Use public data: Crawl the internet to gather all available text
Existing Privacy Preserving Methods

- Scrubbing: Remove secrets from text (before training a model)
  Assumption: Every secret can be found through certain rules.

- Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models
  Assumption: A secret is entirely in one user’s data

- Use public data: Crowl the internet to gather all available text
  Assumption: Publicly available data is public-intended
Existing Privacy Preserving Methods

• Scrubbing: Remove secrets from text (before training a model)
  Assumption: Every secret can be found through certain rules.

• Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models
  Assumption: A secret is entirely in one user’s data

• Use public data: Crowl the internet to gather all available text
  Assumption: Publicly available data is public-intended
Paths Forward

• The problem is hard. Privacy concerns in natural language are as broad as those of real life, and the context in which some information is private is hard if not impossible to infer
Paths Forward

• The problem is hard. Privacy concerns in natural language are as broad as those of real life, and the context in which some information is private is hard if not impossible to infer.

• Nevertheless, we can reduce privacy risks:
  • Make use of publicly-intended data
  • Use private personalization
    • (Only X has access to model trained on X’s data)
  • Use privacy-preserving algorithms
  • Audit privacy risks of algorithms