Auditing and Mitigating Safety Risks in Large Language Models

Niloofar (Fatemeh) Mireshghallah
Act I: The LLM Takeover?
What are Language Models?

- A language model is a \textit{probability distribution} over sequences of words
- Model what words a given word/context normally appears with

The students opened their _______.

- books
- laptops
- exams
Large Language Models (LLMs)

- Transformer-based language models are often referred to as ‘Large LMs’ due to their parameter count (ranging from 100s of million to billions of parameters)

- Deployed with Pre-train and Fine-tune paradigm
Large Language Models: The Good and the Bad …

- Large language models are very good at generating text.
Large language models are very good at generating text and learning representations.
LLMs: The Bad

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.
LLMs: The Bad
LLMs: The Bad

- LLMs can also *regurgitate data* they have seen before, creating *privacy risks*.

Title:
Hi everyone, my name is Anish Athalye and I'm a PhD student at Stanford University.
LLMs: The Bad

- LLMs can also regurgitate data they have seen before, creating privacy risks.
LLMs are *not ready* to be widely deployed in *safety critical scenarios* as is!
In this talk:

**Question 1:** How can we **audit and quantify safety risks** of LLMs?
- [ACL 2023] Membership Inference Attacks via Neighbourhood Comparison
- [EMNLP2022a] Quantifying Privacy Risks of Masked Language Models Using MIAs
- [EMNLP2022b] Memorization in NLP Fine-tuning Methods
- [FAccT2022] What does it mean for language models to preserve privacy?

**Question 2:** How can we **limit the risks** of LLMs?
- [ACL2023] Privacy-Preserving Domain Adaptation of Semantic Parsers
- [NeurIPS2022] Differentially private model compression
- [NAACL2021] Joint privacy-utility optimization in language models

Don’t repeat this!!
Act II: Auditing LLMs for Privacy
What is information leakage in an ML model?

- ‘Leakage’ is being able to **learn information about the training data**, which cannot be learned from other models/data (from the same distribution)
Measuring Leakage: Membership Inference Attacks

- Can an adversary infer whether a particular data point “x” is part of the training set?

Target sample (x)
Measuring Leakage: Membership Inference Attacks

- Can an adversary infer whether a particular data point “x” is part of the training set?
Measuring Leakage: Membership Inference Attacks

- Can an adversary infer whether a particular data point “x” is part of the training set?
Measuring Leakage: Membership Inference Attacks

- Can an adversary infer whether a particular data point “x” is part of the training set?
- Success of attacker is a metric to quantify information leakage of the model about its individual training data
Background: Membership Inference Attacks

- Membership Inference Attacks (MIAs): Loss-based attack

- Stronger MIAs: Reference-based attacks (MIA) [Mireshghallah2022, Ye2021, Carlini2022]
  - A static, absolute threshold does not control for the intrinsic complexity of each utterance
  - We need to calibrate the threshold for each utterance
Reference-based attack

We propose a reference-based attack:

- Complex training points: points that have higher loss
We propose a reference-based attack:
- Complex training points: points that have **higher loss**

<table>
<thead>
<tr>
<th>Training data point</th>
<th>Target Model Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. Smith has type 2 diabetes.</td>
<td>3</td>
</tr>
<tr>
<td>Mr. Smith has fever .</td>
<td>2</td>
</tr>
<tr>
<td>Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.</td>
<td>7</td>
</tr>
</tbody>
</table>
Reference-based attack

We propose a reference-based attack:

- Complex training points: points that have higher loss

<table>
<thead>
<tr>
<th>Training data point</th>
<th>Target Model Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. Smith has type 2 diabetes.</td>
<td>3</td>
</tr>
<tr>
<td>Mr. Smith has fever.</td>
<td>2</td>
</tr>
<tr>
<td>Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.</td>
<td>7</td>
</tr>
</tbody>
</table>
Reference-based attack

We propose a reference-based attack:

- Complex training points: points that have higher loss
- We use a reference model, to provide an insight into how difficult each data point is

<table>
<thead>
<tr>
<th>Training data point</th>
<th>Target Model Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. Smith has type 2 diabetes.</td>
<td>3</td>
</tr>
<tr>
<td>Mr. Smith has fever.</td>
<td>2</td>
</tr>
<tr>
<td>Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.</td>
<td>7</td>
</tr>
</tbody>
</table>
We propose a reference-based attack:

- Complex training points: points that have higher loss
- We use a reference model, to provide an insight into how difficult each data point is

<table>
<thead>
<tr>
<th>Training data point</th>
<th>Target Model Loss</th>
<th>Reference Model Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. Smith has type 2 diabetes.</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Mr. Smith has fever .</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>
Example: loss-based attack

![Histogram showing the proportion of members and non-members with varying losses. The x-axis represents losses, ranging from -1 to 8, and the y-axis represents proportion. The histogram indicates a higher proportion of members with lower losses compared to non-members.]
Example: loss-based attack
Example: Reference-based attack
Our likelihood ratio-based attack has an AUC of 0.90, vs the 0.66 of the loss-based attack.
However …

- The success of reference-based attacks is contingent upon having a ‘good reference’ model, which is not always feasible:
  - We might have a very small dataset, therefore holding out part of the data to train a reference model on would significantly impact the utility of the final model
  - We might have limited/no information about the training data of the model we are probing, therefore curating non-overlapping, similar data would be non-trivial
  - We might not have access to enough compute to train large reference models

How can we leverage the loss function and its curvature to determine membership?
Proposed: Neighbourhood Comparison-based Attacks

- Instead of **likelihood ratio**, we use **local-optimality** of each point as a signal to determine membership. The intuition is:
  - If a data point is part of the training-set, its likelihood would be **locally optimal**, compared to its neighboring points.
  - If a data point is not part of the training set, then there would be points in its neighborhood with both higher and lower likelihoods.
Stocks fall to end Wall Street’s worst year since 2009, S&P 500 ends 2022 down nearly 20%
Experimental Setup

- We are mounting a membership inference attack on fine-tuned GPT2
- Baseline: Likelihood-ratio based attack
  - Base reference: Pre-trained, non-finetuned model
  - Candidate reference: fine-tuned GPT2, but on a dataset with small distribution shift
  - Oracle reference: fin-tuned GPT2 on a dataset with the same distribution as target model
Does this really work?

<table>
<thead>
<tr>
<th>Attack Method</th>
<th>False Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Reference</td>
<td>0.91</td>
</tr>
<tr>
<td>Candidate Reference</td>
<td>0.95</td>
</tr>
<tr>
<td>Oracle Reference</td>
<td>3.76</td>
</tr>
<tr>
<td><strong>Neighborhoud (Ours)</strong></td>
<td>1.73</td>
</tr>
</tbody>
</table>

As we step into lower false-positive rate (more precise) attack scenarios, we see that our method outperforms the likelihood ratio based attack.
Does this really work?

<table>
<thead>
<tr>
<th>Attack Method</th>
<th>False Positive Rate</th>
<th>0.1</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Reference</td>
<td></td>
<td>0.91</td>
<td>0.16</td>
</tr>
<tr>
<td>Candidate Reference</td>
<td></td>
<td>0.95</td>
<td>0.15</td>
</tr>
<tr>
<td>Oracle Reference</td>
<td></td>
<td>3.76</td>
<td>0.16</td>
</tr>
<tr>
<td>Neighborhood (Ours)</td>
<td></td>
<td>1.73</td>
<td>0.29</td>
</tr>
</tbody>
</table>

As we step into lower false-positive rate (more precise) attack scenarios, we see that our method outperforms the likelihood ratio based attack.
Experimental Results: Other Experiments

1. Other Datasets:

2. Ablations:
   - Number of Generated Neighbours
   - Number of Word Replacements

3. Mitigations
   - Differentially Private SGD
Detour: Relation to Machine-generated Text Detection

- Concurrent work: DetectGPT -- Mitchell et al. demonstrate that the same type of algorithm could be used to distinguish between human written text and machine generated text.

So far …

- We show that using a reference model can improve the performance of existing attacks, and uncover higher levels of memorization.

- We also demonstrate reference-free methods, that can be used in scenarios where access to a reference is infeasible.

- How can we mitigate these privacy risks, specifically by generating synthetic data?
Act III: Limiting the Privacy Risks of LLMs
Task-oriented dialogue systems often assist users with personal or confidential matters

- Data is private and practitioners are not allowed to look at it

- How can we know where the system is failing and needs more training data or new functionality?

Could you tell me what the weather is gonna be like today in New York?

Email everyone who declined the invitation, saying …
Background: Differential Privacy

- DP protects the **membership of every single sample** in the training data
- A randomized algorithm $A$ satisfies $\epsilon$-DP, if for all databases $D$ and $D'$ that differ in data pertaining to one user, and for every possible output value $Y$:

$$\frac{\Pr[A(D) = Y]}{\Pr[A(D') = Y]} \leq e^{\epsilon}.$$
Private Training of Large Language Models: Prior Work

- To limit the leakage of fine-tuning data, prior work [Li et al. 2022, Yu et al. 2022] has used DP-SGD during fine-tuning.
Private Training of Large Language Models: Prior Work

- To limit the leakage of fine-tuning data, prior work [Li et al. 2022, Yu et al. 2022] has used DP-SGD during fine-tuning
  - **Differential Privacy SGD (DP-SGD)** is the gold standard of private training
Private Training of Large Language Models: Prior Work

- To limit the leakage of fine-tuning data, prior work [Li et al. 2022, Yu et al. 2022] has used DP-SGD during fine-tuning
  - **Differential Privacy SGD (DP-SGD)** is the gold standard of private training
    - DP protects the **membership of every single sample** in the training data
Differentially Private SGD

Clip gradients for each example

Problem Definition: Adding New Functionality

- Why not just **fine-tune** on the eyes-off data **privately**?
  - If some users are asking the system to hop up and down, fine-tuning is unlikely to make it grow legs.
Problem Definition: Adding New Functionality

- Why not just **fine-tune** on the eyes-off data **privately**?
  - If some users are asking the system to hop up and down, fine-tuning is unlikely to make it grow legs.

What is the weather like in Seattle Today?
Problem Definition: Adding New Functionality

- Why not just **fine-tune** on the eyes-off data **privately**?
  - If some users are asking the system to hop up and down, fine-tuning is unlikely to make it grow legs.
  - We need to be able to **look at synthesized data** to identify additional needed functions, then **annotate** with new functions and **add** to the training data to **improve the semantic parser**.

How can we privately synthesize data that is distributionally close to eyes-off user data?
Baseline: Private Fine-Tuning of a Generative Model

- Intuitive Baseline: We model $p(x)$, where $x$ is a private utterance.
Proposed: 2-stage Modeling of Intermediate Variables

- Intuitive Baseline: We model $p(x)$, where $x$ is a **private utterance**.
- Proposed: We model $p(y)$ and $p(x|y)$, where $y$ is a **private parse-tree**.
  - one stage models the parse-trees, $p_{\theta_y}$
  - The other stage models an utterance given a parse-tree, $p_{\theta_yx}$

```
“Could you tell me what the weather is gonna be like today in New York?”
```

Dataset of private utterances $D_{priv}$

Dataset of private utterances $D_{priv}$

DP utterance generation model $p_{\theta_x}$

DP parse generation model $p_{\theta_yx}$

Privacy Barrier

Corresponding private parse trees

WeatherQueryApi

Yield

DateTime

Today

AtPlace

New York
Does This Really Work?

We simulated a situation where users are asking about the weather but the original semantic parser was not trained on weather-related functions:

1. We created the original semantic parser by training on $\frac{1}{10}$ of our data (SMCalFlow), excluding any examples that use weather-related functions.
2. We treated the other $\frac{9}{10}$ of the data as private user utterances, including those requesting weather. We created approximate private annotations for the private utterances, using the original semantic parser.
3. We apply the baseline and proposed methods to create public synthesized datasets, which include weather functions.
4. We simulated high-quality human annotation of the public synthetic utterances. We re-train the parser with this additional annotated data.
Our proposed 2-stage method outperforms the baseline in terms of the downstream parser performance improvement on the weather function.
Experimental Results: Other Experiments

1. Effect of the **number of modes in the data** distributions on the gains that the 2-stage method provides

2. Effect of **disrupting the correlation** between the parse-trees and utterances

3. Experimenting with **larger models** (GPT2-Large)

4. Studying the **effect of DP hyperparameters** on the privacy-utility trade-off (the budget split between the two stages, the clipping threshold and the learning rate.)

5. Additional Baseline: **1-stage + Domain Prompt**
So far …

- We propose methods **for privately synthesizing data that can be studied and annotated** to improve the performance of semantic parsers, by characterizing the private users’ data.

- **Future Directions:**
  
  - How can we **incorporate active learning** for a more targeted improvement of the semantic-parser?
  
  - How can we modify the objective to **directly evaluate the marginal distribution** over each function type?
Act IV: Future Directions
What is Privacy in Language?
Differential Privacy

- DP is a guarantee that was first *developed and designed for tabular data*
- What makes DP not suitable for language?
  1. Differential privacy requires a *unified definition* for secret boundaries, which is very hard if not impossible to achieve for language data
  2. Protecting a specific unit of data is *not the same as protecting privacy*
  3. The need for privacy does *not diminish with in-group size*
What are people’s expectations of privacy?

Privacy has been defined and discussed in many different fields, including computer security, law, law and psychology.

- People care about and value privacy, defined as respecting the appropriate norms of information flow for a given context.
What are people’s expectations of privacy?

Privacy has been defined and discussed in many different fields, including computer security, law, law and psychology.

<table>
<thead>
<tr>
<th>Security</th>
<th>Law</th>
</tr>
</thead>
<tbody>
<tr>
<td>• People care about and value privacy, defined as respecting the appropriate norms of information flow for a given context.</td>
<td>• To be effective, privacy law must focus on use, harm, and risk rather than on the nature of personal data.</td>
</tr>
</tbody>
</table>
What are people’s expectations of privacy?

Privacy has been defined and discussed in many different fields, including computer security, law, law and psychology.

**Security**
- People care about and value privacy, defined as *respecting the appropriate norms of information flow for a given context.*

**Law**
- To be effective, privacy law must focus on *use, harm, and risk* rather than on the *nature of personal data*.

**Psychology**
- Guarantees of privacy, that is, rules as to *who may and who may not observe or reveal information* about whom, must be *established* in any stable social system.
“Withdrawal into privacy is often a means of making life with an unbearable (or sporadically unbearable) person possible”

Importance of Context
Importance of Context

Will you share your SSN?
Importance of Context

Will you share your SSN?
Importance of Context
Importance of Context

Will you share your SSN?
Contextual integrity gives a framework to reason about norms that apply, in a given social context, to the flows of personal data.
ConfAlde: Benchmarking Contextual Privacy Reasoning in LLMs
ConfAlde: Benchmarking Contextual Privacy Reasoning in LLMs

Is this information type sensitive?

Only Information Type: SSN
ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs

Information

Is this information type sensitive?

Only Information Type: SSN

No Context

Tier 1

Tier 2

Tier 3

Tier 4
ConfAlde: Benchmarking Contextual Privacy Reasoning in LLMs

Is this information flow appropriate?

Tier 1: Only Information Type: SSN

Tier 2: Actor Use

Tier 3: No Context

Tier 4: Information
ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs

What information should flow, to whom?

- Tier 1: No Context
- Tier 2: Actor Use
- Tier 3: Theory of Mind
- Tier 4: Only Information Type: SSN
ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs

Which information should flow, and which should not?

Tier 1
- Only Information
- Type: SSN

Tier 2
- Actor
- Use

Tier 3
- Theory of Mind

Tier 4
- Real-world Applications

No Context
- Real-world Applications
- Theory of Mind
- Actor
- Use
- Only Information
- Type: SSN
• High levels of leakage in theory of mind based scenarios.

• Even CoT doesn't improve leakage, in fact it makes it slightly worse, underscoring the need for fundamental solutions!
Summary and Conclusion

- We probed and analyzed the privacy leakage of large language models through the lens of membership inference attacks
  - We only focused on membership inference attacks here, however, probing privacy leakage for deploying models in real-world cases needs to go beyond that:
    - Other types of attack: extraction, property inference
    - Other data modalities
Summary and Conclusion

- We discussed and introduced privacy mitigation methods that limit the memorization of language models and rely on differential privacy. We also discussed the limitations of such methods.
- We are using models differently now, so we need to protect them differently!
  - New privacy definitions that take into account interactiveness, access to datastores and inference-time concerns!
- Fundamental solutions: bake theory of mind and reasoning into decoding!
Thank you!

niloofar@cs.washington.edu