Act I: The LLM Takeover?
What are Language Models?

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

The students opened their _______.

Options:
- 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

![xkcd.com/1838/](https://xkcd.com/1838/)
Large language models are very good at generating text...
Large language models are very good at generating text and learning representations.
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.
Carlini et al. Extracting Training Data from Large Language Models. USENIX SEC 2021.
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 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
- [EMNLP 2022a] Quantifying Privacy Risks of Masked Language Models Using MIAs
- [EMNLP 2022b] Memorization in NLP Fine-tuning Methods
- [FAccT 2022] What does it mean for language models to preserve privacy?

**Question 2:** How can we limit the risks of LLMs?

- [ACL 2023] Privacy-Preserving Domain Adaptation of Semantic Parsers
- [NeurIPS 2022] Differentially private model compression
- [NAACL 2021] 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).
Can an adversary infer whether a particular data point \( x \) is part of the training set?

Target sample (\( x \))

Mr. Smith has lung Cancer.
Measuring Leakage: Membership Inference Attacks

- Can an adversary infer whether a particular data point \( x \) is part of the training set?

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Member</th>
<th>Non-member</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target sample ( x )</td>
<td>Mr. Smith has lung Cancer.</td>
<td></td>
</tr>
</tbody>
</table>
Measuring Leakage: Membership Inference Attacks

- Can an adversary infer whether a particular data point \( x \) is part of the training set?

Logits from \( x \)

Logits from training data

- Mr. Smith has lung cancer.
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.
Membership Inference Attacks (MIAs):

- Loss-based attack
- Stronger MIAs: Reference-based attacks [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
We propose a reference-based attack:

- Complex training points: points that have higher loss

Mireshghallah et al. Quantifying Privacy Risks of Masked Language Models Using MIAs. EMNLP 2022
We propose a reference-based attack:

- **Complex training points:** points that have higher loss.

Mr. Smith has type 2 diabetes. Mr. Smith has fever.

Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.

Target Model Loss

Mireshghallah et al. Quantifying Privacy Risks of Masked Language Models Using MIAs. EMNLP 2022
We propose a reference-based attack:

- Complex training points: points that have higher loss.

Mr. Smith has type 2 diabetes.

Mr. Smith has fever.

Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.

Target Model Loss

Mireshghallah et al. Quantifying Privacy Risks of Masked Language Models Using MIAs. EMNLP 2022
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.

Mr. Smith has type 2 diabetes.
Mr. Smith has fever.
Mr. Smith is taking 5 mgs of Haloperidol 2 times a day.
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.

---

- **Mr. Smith**: Has type 2 diabetes.
- **Mr. Smith**: Has a fever.

---

- **Training data point 1**: 10
- **Target Model Loss 1**: 4
- **Reference Model Loss 1**: 2

---

- **Training data point 2**: 3
- **Target Model Loss 2**: 7
- **Reference Model Loss 2**: 3

---

- **Training data point 3**: 4
- **Target Model Loss 3**: 10
- **Reference Model Loss 3**: 4

---

*Mireshghallah et al. Quantifying Privacy Risks of Masked Language Models Using MIAs. EMNLP 2022*
Example: Reference-based attack

Graphs showing distributions for Members and Non-members:
- **Loss** distribution: Members and Non-members have different peak patterns and spread.
- **Likelihood Ratio** distribution: Members and Non-members show similar patterns, with a peak at Likelihood Ratio 0 and a tail extending to the right.

Legend:
- **Members**
- **Non-members**
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.

Target Model Likelihood

<table>
<thead>
<tr>
<th>Training point</th>
<th>Non-training point</th>
</tr>
</thead>
</table>

![Graph showing local optimality and likelihoods](image)
Securities fall to end Wall Street’s worst year after 2008, S&P 500 finishes 2022 down almost 20%

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: fine-tuned GPT2 on a dataset with the same distribution as target model
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?

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

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

<table>
<thead>
<tr>
<th></th>
<th>F 0.1</th>
<th>F 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.91</td>
<td>0.16</td>
</tr>
<tr>
<td>C</td>
<td>0.95</td>
<td>0.15</td>
</tr>
<tr>
<td>Neighborhoud (Ours)</td>
<td>3.76</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>1.73</td>
<td>0.29</td>
</tr>
</tbody>
</table>
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.

Candidate passage $x$:
“Joe Biden recently made a move to the White House that included bringing along his pet German Shepherd…”

DetectGPT

1. Perturb (reword with T5)
2. Score
3. Compare

From GPT-3

Yes

No

$x$ from other source
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 $\varepsilon$-DP, if for all databases $D$ and $D'$ that differ in data pertaining to one user, and for every possible output value $Y$:

$$\frac{\Pr[A(D) = Y]}{\Pr[A(D') = Y]} \leq e^\varepsilon.$$
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 Differential Privacy SGD (DP-SGD) during fine-tuning.

Differential Privacy SGD (DP-SGD) is the gold standard of private training.
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)
  - The gold standard of private training
  - Protects the membership of every single sample in the training data.

Don’t repeat this!!
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.
  - We need to be able to look at synthesized data, have it annotated and added to the training data to improve the semantic parser.

What is the weather like in Seattle Today?
- WeatherQueryApi
- Yield
  - AtPlace
    - Seattle
  - DateTime
    - Today

Existing annotated utterances
- Improved semantic parser
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?
Intuitive Baseline: We model $p(x)$, where $x$ is a private utterance.
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, $p_\theta y x$.


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 retrain the parser with this additional annotated data.
Does This Really Work?

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

- Differential privacy requires a unified definition for secret boundaries, which is very hard if not impossible to achieve for language data.
- Protecting a specific unit of data is not the same as protecting privacy.
- 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.

Security

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

Law

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.

Psychology
What are people's expectations of privacy?

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

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

Security

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

Law

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.

Psychology
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.

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

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.

Barry Schwartz, 1968, *The Social Psychology of Privacy*
Account Locked!

Birth date
DD/MM/YYYY

Social security number
SSN (9 digits)

ATM or Debit Card PIN
Card PIN

Continue
Context is key!

Will you share your SSN?

Importance of Context
Account Locked!

Will you share your SSN?
Great News! We Can Enter Your W-2 for You

Instead of filling up to 20 boxes yourself, let us import your W-2 into your return. You'll save time and finish your taxes faster.

All fields are required.

- **SSN (e.g. 123456789)**
- **User ID (username:EIN, i.e. abc123:23-1352630)**
- **Password (Box 1 Amount on your W-2: i.e. 2500.03)**

We keep your information completely secure. Learn more about our security.

Once imported, please verify all of the information matches your original 2017 W-2. If you have questions regarding your W-2, please contact payroll@drexel.edu. All W-2 data and credentials are maintained on Drexel University’s servers.

Enter your SSN (123456789), your UserID:EIN (lower case abc123:23-1352630, abc123:23-1352000 or abc123:47-3658191), and your password, the value in W-2 Box 1, with no commas, 2 decimals (i.e. 2500.03).

More Instructions
Will you share your SSN?
Contextual integrity gives a framework to reason about **norms** that apply, in a given social context, **to the flows personal data**.
Only Information Type: SSN
ConfAIde: Multi-tier benchmark

Only Information Type: SSN

Is this information type sensitive?
ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs

Is this information type sensitive?

Only Information Type: SSN

No Context
ConfAIde: Multi-tier benchmarking context-aware privacy reasoning in LLMs

Tier 1
- No Context

Tier 2
- Actor Use

Tier 3
- Only Information Type: SSN

Tier 4

Is this information flow appropriate?
ConfAIde: Multi-tier benchmark

Tier 1:
- Only Information
- Type: SSN

Tier 2:
- Actor
- Use

Tier 3:
- Theory of Mind

Tier 4:
- No Context

What information should flow, to whom?
ConfAIde: Multi-tier benchmark

Type: SSN

Tier 1

Tier 2

Tier 3

Tier 4

Real-world Applications

Theory of Mind

Actor Use

No Context

Only Information Type: SSN

Which information should flow, and which should not?
Surely, CoT will help?

- 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!

ConfAIde: Benchmarking Contextual Privacy Reasoning in LLMs

![Tier 4: Meeting Summary Secret Leakage](image)
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
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!