Auditing and Mitigating Safety Risks in Large Language Models

Fatemehsadat Mireshghallah

May 2023
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
Large Language Models: The Good and the Bad …

- Large language models are very good at generating text

ChatGPT passes exams from law and business schools
Large Language Models: The Good and the Bad …

- Large language models are very good at **generating text** and **learning representations**.
Large Language Models: The Good and the Bad …

- Large language models are very good at generating text and learning representations. However:
  - They are extremely large models: **high capacity for memorization**
  - They are trained on huge, unvetted, scraped data: potential for **harmful/hateful/private content**
Large Models are Leaky

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.

xkcd.com/2169/
Large Models are Leaky: Data Extraction

Carlini et al. Extracting Training Data from Large Language Models. USENIX SEC 2021.
Large Models are Leaky: Data Extraction

- Github CoPilot

Title:

Hi everyone, my name is Anish Athalye and I'm a PhD student at Stanford University.
Large Models are Leaky: Data Extraction

- Github CoPilot

<table>
<thead>
<tr>
<th>Title:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi everyone, my name is Anish Athalye and I'm a PhD student at Stanford University.</td>
</tr>
</tbody>
</table>

https://www.anish.io

Anish Athalye
I am a PhD student at MIT in the PDOS group. I'm interested in formal verification, systems, security, and machine learning.

GitHub: @anishathalye  Blog: anishathalye.com
In This Talk …

- We try to answer these questions:

  Q1: How can we **audit and quantify privacy risks** of language models in terms of what information they have memorized?

  - We introduce tools and frameworks for quantifying privacy leakage
  - We use these tools to identify memorization patterns in pre-training and fine-tuning of LMs
In This Talk …

- We try to answer these questions:

  Q1: How can we **audit and quantify privacy risks** of language models in terms of what information they have memorized?

  Q2: How can we **limit the privacy risks of large language models** by limiting leakage and memorization during training?
In This Talk …

- We try to answer these questions:

Q1: How can we **audit and quantify privacy risks** of language models in terms of what information they have memorized?

Q2: How can we **limit the privacy risks of large language models** by limiting leakage and memorization during training?

  - We introduce methods for limiting memorization and leakage of training data.
In This Talk …

- We try to answer these questions:

  Q1: How can we **audit and quantify privacy risks** of language models in terms of what information they have memorized?

  Q2: How can we **limit the privacy risks of large language models** by limiting leakage and memorization during training?

    - We introduce methods for limiting memorization and leakage of training data
    - We discuss what it really means for a language model to be privacy-preserving
Talk outline

● **Question 1: Privacy Auditing**
  • [EMNLP2022a] Quantifying Privacy Risks of Masked Language Models Using MIA\(_s\)
  • [EMNLP2022b] Memorization in NLP Fine-tuning Methods

● **Question 2: Privacy Protection and Risk Mitigation**
  • [NeurIPS2022] Differentially private model compression
  • [FAccT2022] What does it mean for language models to preserve privacy?

● **Summary and Conclusion**
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?

Mr. Smith has lung Cancer.

Target sample (x)

Logits from x

Logits from training data
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
Quantifying Leakage in Language Models: Prior Work

- Pre-trained **autoregressive** (causal) models: $p(\text{has} | \text{Mr}, \text{Smith}; \theta)$
Quantifying Leakage in Language Models: Prior Work

● Pre-trained autoregressive (causal) models:
  • Extraction Attack on GPT-2 [Carlini et al. 2021]:
    ■ Generate 500k samples from the model
    ■ Sift through them using an MIA to find actual training samples: over 60% precision
Quantifying Leakage in Language Models: Prior Work

- Pre-trained **autoregressive** (causal) models:
  - Extraction Attack on GPT-2 [Carlini et al. 2021]:
    - Generate 500k samples from the model
    - Sift through them using an MIA to find actual training samples: over 60% precision

- Pre-trained **masked language models** $p(\text{has}|s\backslash\text{has}; \theta)$:

```
Mr  Smith  has  acute  asthma
```
How Do We Measure Leakage in Masked Language Models?

- **Extraction attacks** [Lehman et al. 2021]: Fill in the blank and sampling attacks, very low success rate.

Mr. Smith has ____

Mr. Smith has diabetes.
How Do We Measure Leakage in Masked Language Models?

- Extraction attacks [Lehman et al. 2021]: Fill in the blank and sampling attacks, very low success rate.
- **Membership Inference Attacks** (MIAs) [jagannatha2021]: Loss-based attack
  - Set a **hard threshold on loss** for inferring membership, based on a held-out test set
  - near random performance, lots of **false negatives**
How Do We Measure Leakage in Masked Language Models?

- Extraction attacks [Lehman et al. 2021]: Fill in the blank and sampling attacks, very low success rate.
- Membership Inference Attacks (MIAs) [jagannatha2021]: Loss-based attack, near random performance

Does this mean MLMs memorize less than auto-regressive models? Or do we just need stronger attacks?
How Do We Measure Leakage in Masked Language Models?

- Extraction attacks [Lehman et al. 2021]: Fill in the blank and sampling attacks, very low success rate.
- Membership Inference Attacks (MIAs) [jagannatha2021]: Loss-based attack, near random performance
- Stronger MIAs: Reference-based attacks (MIA) [Sankararaman2009, Ye2021]
How Do We Measure Leakage in Masked Language Models?

- Extraction attacks [Lehman et al. 2021]: Fill in the blank and sampling attacks, very low success rate.
- Membership Inference Attacks (MIAs) [jagannatha2021]: Loss-based attack, near random performance

- Stronger MIAs: Reference-based attacks (MIA) [Sankararaman2009, Ye2021]
  - A static, absolute threshold does not control for the intrinsic complexity of each utterance
How Do We Measure Leakage in Masked Language Models?

- Extraction attacks [Lehman et al. 2021]: Fill in the blank and sampling attacks, very low success rate.
- Membership Inference Attacks (MIAs) [jagannatha2021]: Loss-based attack, near random performance

- Stronger MIAs: Reference-based attacks (MIA) [Sankararaman2009, Ye2021]
  - 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 for MLMs

We propose a reference-based attack for MLMs:

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

- 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 for MLMs

We propose a reference-based attack for MLMs:

- 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>
We propose a reference-based attack for MLMS:

- 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>
Reference-based attack for MLMs

We propose a reference-based attack for MLMs:
- 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>
Reference-based attack for MLMs

- So instead of loss, we use likelihood ratio with respect to a reference model:

  \[
  LR(s) = \frac{p(s; \theta_R)}{p(s; \theta)}
  \]
Reference-based attack for MLMs

- So instead of loss, we use **likelihood ratio** with respect to a **reference** model:

  - likelihood ratio: \[
  LR(s) = \frac{p(s; \theta_R)}{p(s; \theta)}
  \]

- By **thresholding the LR**, we infer membership: \( LR(s) < t \rightarrow s \in D \)
Reference-based attack for MLMs

- So instead of loss, we use **likelihood ratio** with respect to a **reference** model:

  - likelihood ratio: \( LR(s) = \frac{p(s; \theta_R)}{p(s; \theta)} \)

- By **thresholding the LR**, we infer membership: \( \text{if } LR(s) < t \rightarrow s \in D \)

- Would need likelihood for each sample! However, **MLMs don’t naturally give probability distributions over full utterances!!**
Reference-based attack for MLMs

- So instead of loss, we use **likelihood ratio** with respect to a **reference** model:

  - likelihood ratio:
    
    \[
    LR(s) = \frac{p(s; \theta_R)}{p(s; \theta)}
    \]

  - By **thresholding the LR**, we infer membership: 
    \[
    \text{if } LR(s) < t \rightarrow s \in D
    \]

  - Would need likelihood for each sample! However, **MLMs don’t naturally give probability distributions over full utterances!!**

  - We need \( p(s; \theta) \), but the model gives us \( p(x_i|s\setminus x_i) \).
Likelihood Ratio for MLMs

- We view pre-trained MLMs as *energy-based probability distributions* on sequences:

\[
p(s; \theta) = \frac{e^{-E(s; \theta)}}{Z_\theta}
\]
Likelihood Ratio for MLMs

- We view pre-trained MLMs as energy-based probability distributions on sequences:

\[
p(s; \theta) = \frac{e^{-E(s; \theta)}}{Z_\theta}
\]

\(E(s; \theta)\) refers to the scalar energy of a sequence \(s\). \(Z\) is the normalization constant.
Likelihood Ratio for MLMs

- We view pre-trained MLMs as energy-based probability distributions on sequences:

  \[ p(s; \theta) = \frac{e^{-E(s; \theta)}}{Z_{\theta}} \]

  \( E(s; \theta) \) refers to the scalar energy of a sequence \( s \). \( Z \) is the normalization constant.

- Given this:

  \[ LR(s) = \log \left( \frac{p(s; \theta_R)}{p(s; \theta)} \right) = \log \left( \frac{e^{-E(s; \theta_R)}}{Z_{\theta_R}} \right) - \log \left( \frac{e^{-E(s; \theta)}}{Z_{\theta}} \right) \]

  \[ = -E(s; \theta_R) - \log(Z_{\theta_R}) + E(s; \theta) + \log(Z_{\theta}) \]

  \[ = E(s; \theta) - E(s; \theta_R) + \text{constant} \]
Likelihood Ratio for MLMs

- We view pre-trained MLMs as **energy-based probability distributions** on sequences:

\[
p(s; \theta) = \frac{e^{-E(s; \theta)}}{Z_\theta}
\]

\(E(s; \theta)\) refers to the scalar energy of a sequence \(s\). \(Z\) is the normalization constant.

- Given this:

\[
LR(s) = \log \left( \frac{p(s; \theta_R)}{p(s; \theta)} \right) = \log \left( \frac{e^{-E(s; \theta_R)}}{Z_{\theta_R}} \right) - \log \left( \frac{e^{-E(s; \theta)}}{Z_\theta} \right)
\]

\[
= -E(s; \theta_R) - \log(Z_{\theta_R}) + E(s; \theta) + \log(Z_\theta)
\]

\[
= E(s; \theta) - E(s; \theta_R) + \text{constant}
\]

- \(E(s; \theta)\) is calculated as **average of MLM conditionals** \(p(x_i|s \setminus x_i)\), over multiple masks.
Our Attack

General Data Distribution ($p$)
Our Attack

General Data Distribution ($p$) → Training Data ($D \sim p$) → Target Model $M_\theta$
Our Attack

Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks, F Mireshghallah, K Goyal, A Uniyal, T Berg-Kirkpatrick, R Shokri2022
Our Attack

General Data Distribution \((\mathcal{P})\)

Training Data \((\mathcal{D} \sim \mathcal{P})\)

Target Model \(M_\theta\)

Target Sample \((s)\)

Reference Model \(M_{\theta_R}\)

Mr. Smith has lung Cancer.

Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks, F Mireshghallah, K Goyal, A Uniyal, T Berg-Kirkpatrick, R Shokri2022
Our Attack

General Data Distribution \((p)\)

Training Data \((D \sim p)\)

Target Model \(M_\theta\)

Target Sample \((s)\)

Mr. Smith has lung Cancer.

Reference Model \(M_{\theta_R}\)

\[ LR(s) = \log \frac{P(s; \theta_R)}{P(s; \theta)} < t \]

Likelihood Ratio Test

Member

Non-member

Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks, F Miresghallah, K Goyal, A Uniyal, T Berg-Kirkpatrick, R Shokri 2022
Example: loss-based attack
Example: loss-based attack
Example: Reference-based attack
Experimental Setup

**Task and Model**
- Masked (causal) language modeling
- ClinicalBERT models

**Datasets**
- MIMIC III medical notes
- I2b2 medical notes

**Metrics**
- Area under ROC curve
- Recall for FPR@10%
Our likelihood ratio-based attack has an AUC of 0.90, vs the 0.66 of the loss-based attack.
Would other reference models work too?

<table>
<thead>
<tr>
<th></th>
<th>PubMed BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUC</strong></td>
<td></td>
</tr>
<tr>
<td>Baseline (loss-based)</td>
<td>0.66</td>
</tr>
<tr>
<td><strong>Ours (likelihood ratio)</strong></td>
<td>0.90</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td></td>
</tr>
<tr>
<td>Baseline (loss-based)</td>
<td>15.6</td>
</tr>
<tr>
<td><strong>Ours (likelihood ratio)</strong></td>
<td>79.2</td>
</tr>
</tbody>
</table>
Would other reference models work too?

<table>
<thead>
<tr>
<th></th>
<th>PubMed BERT</th>
<th>BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUC</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (loss-based)</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td><strong>Ours (likelihood ratio)</strong></td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (loss-based)</td>
<td>15.6</td>
<td>15.6</td>
</tr>
<tr>
<td><strong>Ours (likelihood ratio)</strong></td>
<td>79.2</td>
<td>71.5</td>
</tr>
</tbody>
</table>
So far …

- We show that unlike prior belief, pre-trained MLMs are also susceptible to membership inference attacks!

- Attacks need some adjustments based on model architecture and optimization

- However, we have only looked at pre-trained models so far! What are the memorization patterns of fine-tuning data?
Talk outline

● **Question 1: Privacy Auditing**
  • [EMNLP2022a] Quantifying Privacy Risks of Masked Language Models Using MIAs
  • [EMNLP2022b] Memorization in NLP Fine-tuning Methods

● **Question 2: Privacy Protection and Risk Mitigation**
  • [NeurIPS2022] Differentially private model compression
  • [FAccT2022] What does it mean for language models to preserve privacy?

● **Summary and Conclusion**
Memorization in Fine-tuning Large Language Models

- Large language models are often deployed with the pre-train and fine-tune paradigm:
  1. Pre-train on a huge (usually web-scraped) “public” corpus.
  2. Fine-tune on a smaller domain specific (usually private) dataset, for down-stream task.
Memorization in Fine-tuning Large Language Models

- Fine-tuning (domain adaptation) can be riskier in terms of privacy, as it is more often, on smaller domain specific datasets, such as emails, company messages, etc.

- Three main fine-tuning methods:
  1. Fine-tuning the model in full (all parameters)
Memorization in Fine-tuning Large Language Models

- Fine-tuning (domain adaptation) can be riskier in terms of privacy, as it is more often, on smaller domain specific datasets, such as emails, company messages, etc.

- Three main fine-tuning methods:
  1. Fine-tuning the model in full (all parameters)
  2. Fine-tuning the ‘head’: head is a dense classifier layer added on top of the transformer architecture to perform the given down-stream task.
Fine-tuning (domain adaptation) can be riskier in terms of privacy, as it is more often, on smaller domain specific datasets, such as emails, company messages, etc.

Three main fine-tuning methods:

1. Fine-tuning the model in full (all parameters)

2. Fine-tuning the 'head': head is a dense classifier layer added on top of the transformer architecture to perform the given down-stream task.

3. Fine-tuning Adapters
Experimental Setup

**Task and Model**
- Autoregressive (causal) language modeling
- Pre-trained GPT-2

**Datasets**
- Penn Tree Bank
- Wikipedia
- Enron email dataset

**Metrics**
- MIA Recall
- Exposure Metric
Memorization Phases

Training Progression (in epochs)

Validation PPL

MIA Recall

21 22 23 24 25 26
Memorization Phases

(1) Fitting + Memorization
Memorization Phases

(1) Fitting + Memorization

(2) Memorization Only

![Graph showing Memorization Phases](image)

- X-axis: Validation PPL
- Y-axis: MIA Recall

Legend:
- Blue dots: Fitting + Memorization
- Green dots: Memorization Only
There is high levels of memorization of the fine-tuning data.
Memorization Phases

Early stopping is necessary to avoid the ‘memorization only’ phase.
Memorization Trends

- **Head fine-tuning** has the **least desirable** utility-privacy trade-off, although it doesn’t have the greatest number of parameters (38 Million, vs 124 Million of full fine-tuning)
Memorization Trends

- **Head fine-tuning** has the least desirable utility-privacy trade-off, although it doesn’t have the greatest number of parameters (38 Million, vs 124 Million of full fine-tuning)
- **Adapter** fine-tuning and **full** fine tuning are on the **Pareto frontier**
So far …

- We show that unlike prior belief, **pre-trained MLMs** are also susceptible to **membership inference attacks**!

- **Domain adaptation data is heavily memorized** and susceptible to being inferred

- How can we **mitigate these privacy risks**, specifically for domain adaptation in **smaller models**?
Talk outline

- **Question 1: Privacy Auditing**
  - [EMNLP2022a] Quantifying Privacy Risks of Masked Language Models Using MIAs
  - [EMNLP2022b] Memorization in NLP Fine-tuning Methods

- **Question 2: Privacy Protection and Risk Mitigation**
  - [NeurIPS2022] Differentially private model compression
  - [FAccT2022] What does it mean for language models to preserve privacy?

- **Summary and Conclusion**
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

However …

- Large LMs have **high inference cost**!
  - It takes 202 seconds to run MNLI test set (20k samples) on a Tesla P100 on BERT
  - Users interact with **smaller models** on edge devices!
However …

- Large LMs have **high inference cost**!
  - It takes 202 seconds to run MNLI test set (20k samples) on a Tesla P100 on BERT
  - Users interact with **smaller models** on edge devices!

- Training small LMs with **DP-SGD** would yield **very poor utility**
  - Training models from scratch with DP-SGD yields undesirable privacy/utility trade-off due to noise
Let’s Compress Models!

- Compress models (via distillation, pruning, quantization, etc.) to decrease inference costs
Let’s Compress Models!

- How do we make sure that private data does not leak during compression?

What algorithms should one use to produce compressed private models and how do they impact private fine-tuning?
Private Compression: Knowledge Distillation

- We propose and analyze two frameworks:
  1. Differentially Private Knowledge Distillation (DP-KD)
Private Compression: Knowledge Distillation

- We propose and analyze two frameworks:
  1. Differentially Private Knowledge Distillation (DP-KD)

Knowledge Transfer (w/ DP-SGD: Clip and add noise)
Private Compression: Weight Pruning

- We propose and analyze two frameworks:

  1. Differentially Private Knowledge Distillation (DP-KD)
  2. Differentially Private Iterative Magnitude Pruning (DP-IMP):

    1. Start with a pre-trained model $M$
Private Compression: Weight Pruning

- We propose and analyze two frameworks:
  1. Differentially Private Knowledge Distillation (DP-KD)
  2. Differentially Private Iterative Magnitude Pruning (DP-IMP):
     1. Start with a pre-trained model $M$
     2. Fine-tune $M$ on private data for $N$ steps, using DP-SGD

![Diagram showing private data, DP-SGD, and a neural network model](image)
Private Compression: Weight Pruning

- We propose and analyze two frameworks:
  1. Differentially Private Knowledge Distillation (DP-KD)
  2. Differentially Private **Iterative Magnitude Pruning (DP-IMP):**
     1. Start with a pre-trained model $M$
     2. **Fine-tune** $M$ on private data for $N$ steps, using **DP-SGD**
     3. Prune $\alpha\%$ of the weights with the **lowest magnitude** from $M$
We propose and analyze two frameworks:

1. Differentially Private Knowledge Distillation (DP-KD)

2. Differentially Private Iterative Magnitude Pruning (DP-IMP):
   1. Start with a pre-trained model $M$
   2. Fine-tune $M$ on private data for $N$ steps, using DP-SGD
   3. Prune $\alpha\%$ of the weights with the lowest magnitude from $M$
   4. Repeat steps 2-3 till sparsity level $S\%$ is reached.
Experimental Setup

- **Tasks**
  - **Glue**:
    - Natural Language Inference (MNLI and QNLI)
    - Sentiment Analysis (SST2)
    - Quora Question Pairs (QQP, paraphrasing)

- **Datasets**
  - **MNLI**: 393k/20k
  - **QNLI**: 105k/5.4k
  - **SST2**: 67k/1.8k
  - **QQP**: 364k/391k
  - **E2E**: 42k/4.9k

- **Metrics**
  - **Classification Accuracy**
  - **Sparsity Level: 50%**
Summary of Findings

- DP Knowledge Distillation:

  1. Drop in accuracy: There is a considerable drop in the accuracy between the teacher and the student models.

```
<table>
<thead>
<tr>
<th>Model</th>
<th>MNLI Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine-tuned BERT (Teacher)</td>
<td></td>
</tr>
<tr>
<td>Fine-tuned Half-BERT (Baseline)</td>
<td>45</td>
</tr>
<tr>
<td>DP-KD Half-BERT (Ours, student)</td>
<td>30</td>
</tr>
</tbody>
</table>
```


Summary of Findings: Model Initializations

- **Model initialization** has a significant impact on final **accuracy**, especially when using **DP-SGD**. We try three initializations for the compressed model:

  1. Random initialization
Summary of Findings: Model Initializations

- **Model initialization** has a significant impact on final accuracy, especially when using DP-SGD. We try three initializations for the compressed model:
  
  1. Random initialization
  2. Initialization with layers of BERT
Summary of Findings: Model Initializations

- **Model initialization** has a significant impact on final **accuracy**, especially when using **DP-SGD**. We try three initializations for the compressed model:

  1. Random initialization
  2. Initialization with layers of BERT
  3. Initialization with DistilBERT
Summary of Findings

- **DP Knowledge Distillation:**

  2. Good initialization of students is crucial: Pre-trained DistilBERT achieves the best student performance.
Summary of Findings

- **DP Pruning:**

  DP pruning produces a compressed model that has better performance than fine-tuning and distillation.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>MNLI Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP-IMP Half-BERT (Ours, student)</td>
<td>74</td>
</tr>
<tr>
<td>DP-KD Half-BERT (Ours, student)</td>
<td>72</td>
</tr>
<tr>
<td>Fine-tuned Half-BERT (Baseline)</td>
<td>68</td>
</tr>
<tr>
<td>Fine-tuned BERT</td>
<td>76</td>
</tr>
</tbody>
</table>
Talk outline

● Question 1: Privacy Auditing
  • [EMNLP2022a] Quantifying Privacy Risks of Masked Language Models Using MIAs
  • [EMNLP2022b] Memorization in NLP Fine-tuning Methods

● Question 2: Privacy Protection and Risk Mitigation
  • [NeurIPS2022] Differentially private model compression
  • [FAccT2022] What does it mean for language models to preserve privacy?

● Summary and Conclusion
What does preserving privacy in language modeling require?

- To claim a language model is privacy preserving, it must only reveal private information (aka “secrets”) in the right contexts and to the right people. We have to define the following:

  - in what contexts a secret can be shared without violating privacy?
  - what information is contained in the secret?
  - which people know the secret (the "in-group")?
Challenges in Removing Secrets: Context

- Privacy is not a 0-1 thing, it’s a spectrum
  - A phone number could be private in one context, public in another
  - Subject, sender, recipient, information type all determine the context
Challenges in Removing Secrets: Context

Conversation A

Hi Alice how are things going?

Not great…

Did I already tell you I’m getting a divorce?

No I’m sorry to hear that!

What are you going to do about custody of the kids?
Challenges in Removing Secrets: Context

**Conversation A**

Hi Alice how are things going?

Not great…

Did I already tell you I’m getting a divorce?

No I’m sorry to hear that!

What are you going to do about custody of the kids?
Challenges in Removing Secrets: Context

Conversation A

Hi Alice how are things going?

Bob

Not great...

Did I already tell you I'm getting a divorce?

Alice

No I'm sorry to hear that!

What are you going to do about custody of the kids?

Bob

Hi Alice how are things going?

Bob

Not great.

Did I already tell you I'm getting a divorce?

Alice

No I'm sorry to hear that!

What are you going to do about custody of the kids?

Bob
Challenges in Removing Secrets: Context

Conversation A

Hi Alice how are things going?

Bob

Not great.

Did I already tell you I’m getting a divorce?

Alice

No I’m sorry to hear that!

What are you going to do about custody of the kids?

Bob

Hi Alice how are things going?

Bob

Not great.

Did I already tell you I’m getting a divorce?

Alice

No I’m sorry to hear that!

What are you going to do about custody of the kids?

Bob
Challenges in Removing Secrets: Context

Conversation B

Hey Bob how've you been??

Charlie

Pretty good wbu?

Bob

Did you hear Alice is getting divorced??
Secrets are hard to identify

Form and Meaning: There are many ways to communicate any piece of information.

Repeated information can still be private information.

Language evolves, and so does private information.
Existing Privacy Mechanisms for NLP

- **Scrubbing**:
  
  - Commonly used on medical data using NER methods
  
  - Sanitization is insufficient because private information is context dependent, not identifiable, and not discrete (static).
    
    - The first 2 digits are two two, and the remaining ones are three (223)
Differential Privacy

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 alternatives do we have?

- Publicly accessible data?
  - No, publicly accessible data is not public-intended: leaked messages, deleted texts, personal blogs

- Can users provide informed consent?
  - Mostly not. If such a consent mechanism were to exist, it would be challenging for users to reach an informed decision about the consequences of their actions.

- Private personalization
  - Maybe
Talk outline

● **Question 1: Privacy Auditing**
  - [EMNLP2022a] Quantifying Privacy Risks of Masked Language Models Using MIAs
  - [EMNLP2022b] Memorization in NLP Fine-tuning Methods

● **Question 2: Privacy Protection and Risk Mitigation**
  - [NeurIPS2022] Differentially private model compression

● **Summary and Conclusion**
Summary and Future Directions

• 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 – more efficient
    ■ Other data modalities
    ■ What is good vs. bad memorization? When do we want models to memorize?
    ■ How does memorization change when data representations change?
Summary and Future Directions

- We discussed and introduced privacy mitigation methods that limit the memorization of language models and rely on differential privacy
  - What are better model initializations?
  - Differential privacy provides worst-case guarantees and protects the membership of any given ‘record’, but what is a record in practice?
  - How do we interpret the provided DP guarantees in language models? How do we communicate them to users of these models?
  - How do we choose the right privacy parameter? What are the policies around that?
  - What alternative privacy definitions can we come up with? How can we reason about privacy?
Thank you!