Privacy Concerns of Large Language Models

Xiangyu Qi & Tong Wu

Oct. 26

Outline

1. Introduction
2. Carlini et al,.2020
    a. Thread Model
    b. Extracting Training Data from LLMs
    c. Attack Evaluation
    d. Potential Mitigations (related to Kandpal et al.,2022)
3. Conclusion

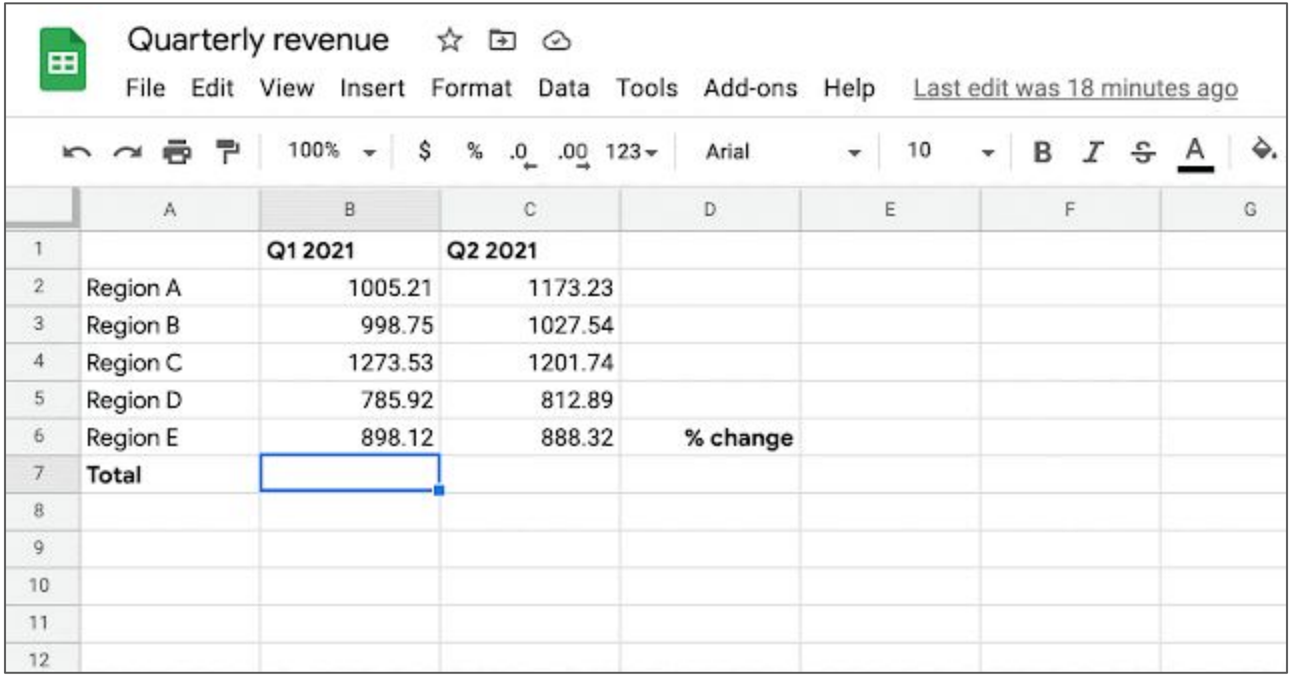
Deep Learning might be Trained on Sensitive Data

TECHNOLOGY FEATURE | 21 April 2020

Deep learning takes on tumours

Artificial-intelligence methods are moving into cancer research.

Deep Learning might be Trained on Sensitive Data



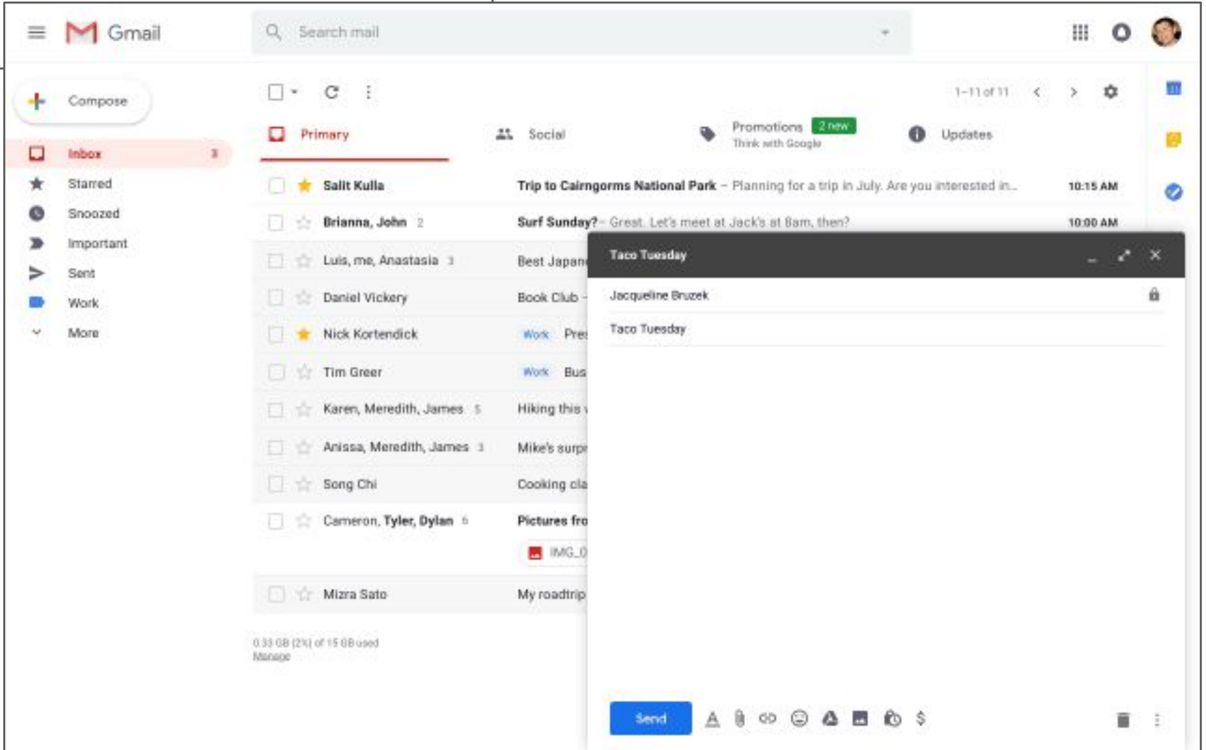
	A	B	C	D	E	F	G
1		**Q1 2021**	**Q2 2021**				
2	Region A	1005.21	1173.23				
3	Region B	998.75	1027.54				
4	Region C	1273.53	1201.74				
5	Region D	785.92	812.89				
6	Region E	898.12	888.32	**% change**			
7	**Total**						
8							
9							
10							
11							
12							

Image Source

Deep Learning might be Trained on Sensitive Data

GMAIL

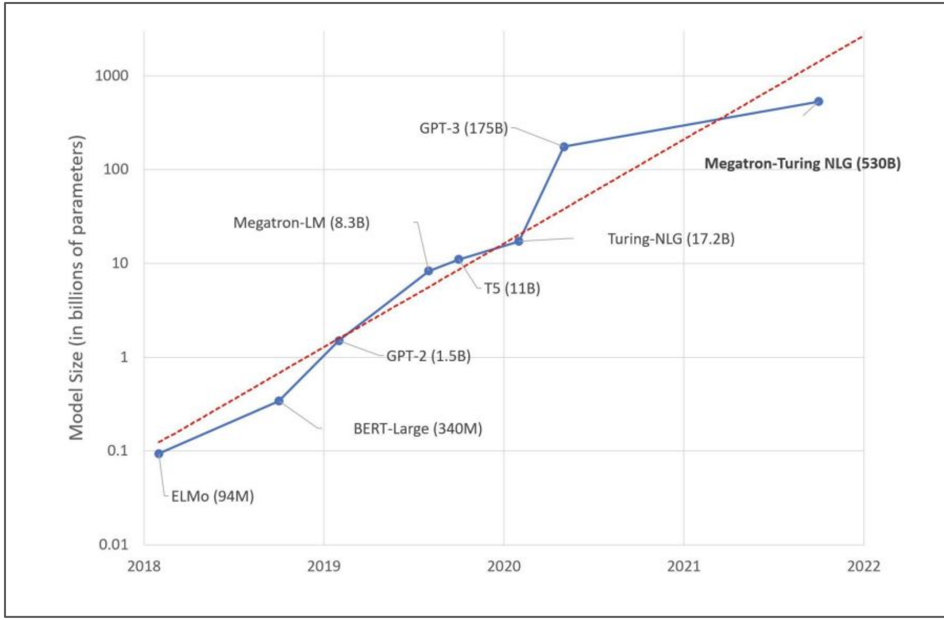
SUBJECT: Write emails faster with Smart Compose in Gmail



[Image Source](#)

LLMs increase fast

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion



LLMs Privacy Concerns

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

Private Information

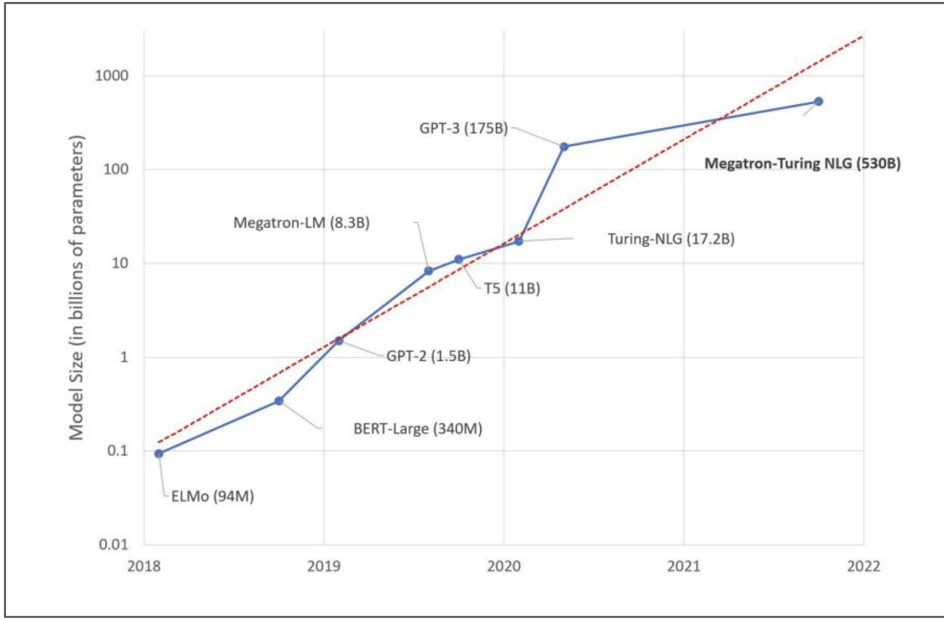


Image Source Image Source

LLMs Privacy Concerns

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

Keep Private

Private Information

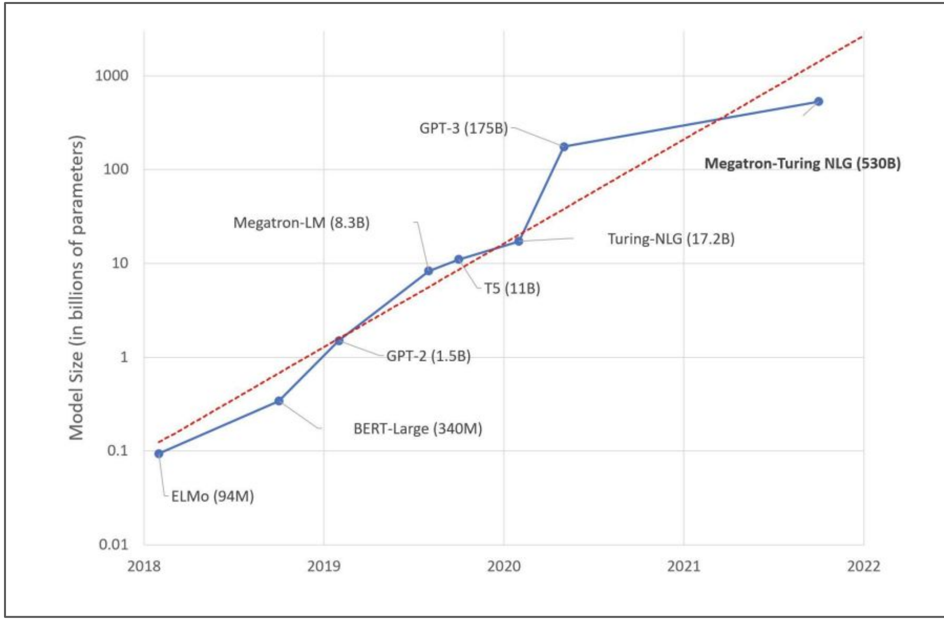


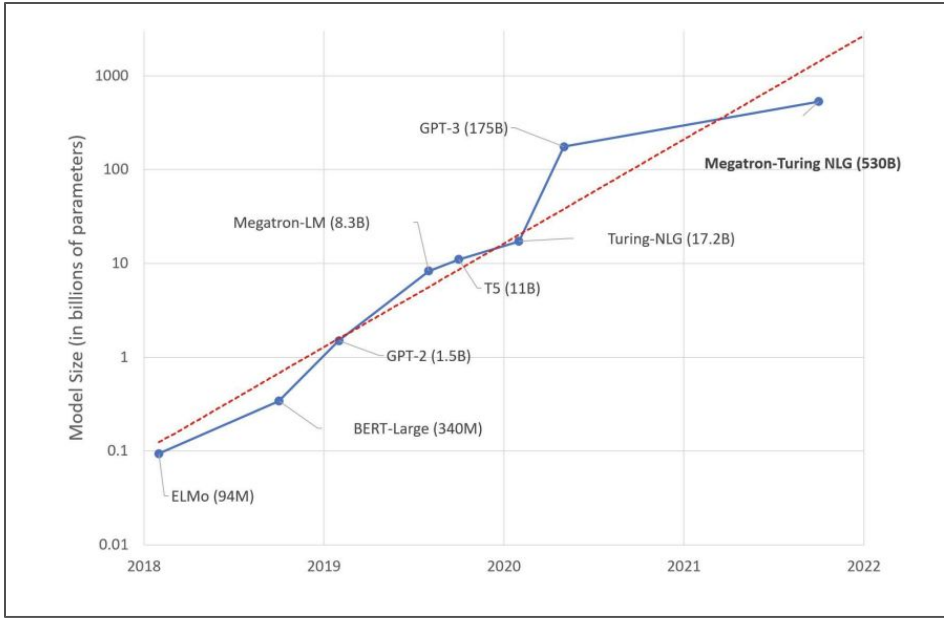
Image Source Image Source

LLMs Privacy Concerns

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

Keep Private

Private Information



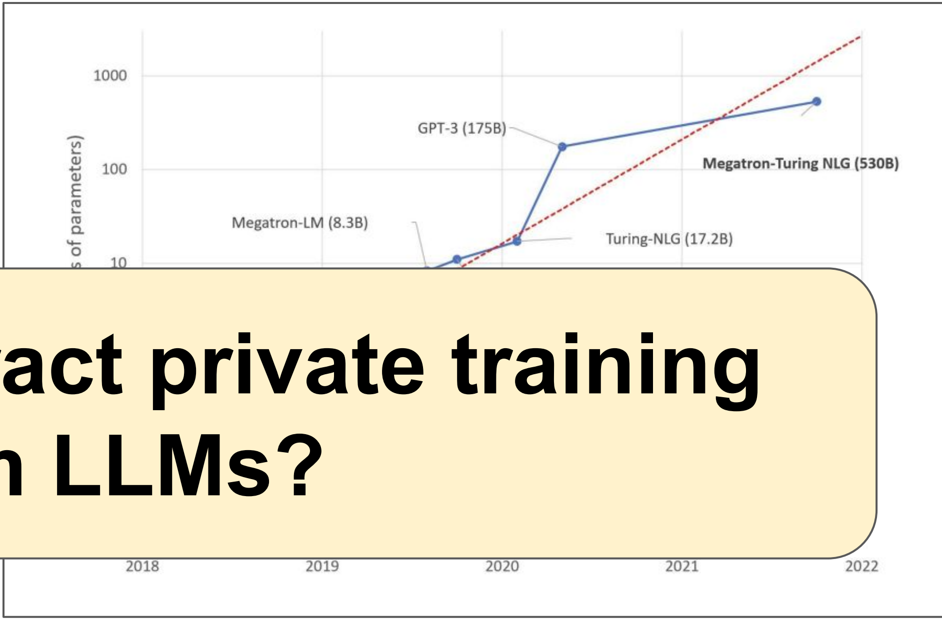
Publicly Available

Privacy Concerns?

Image Source Image Source

LLMs Privacy Concerns

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion



Is it possible to extract private training data from LLMs?

Information

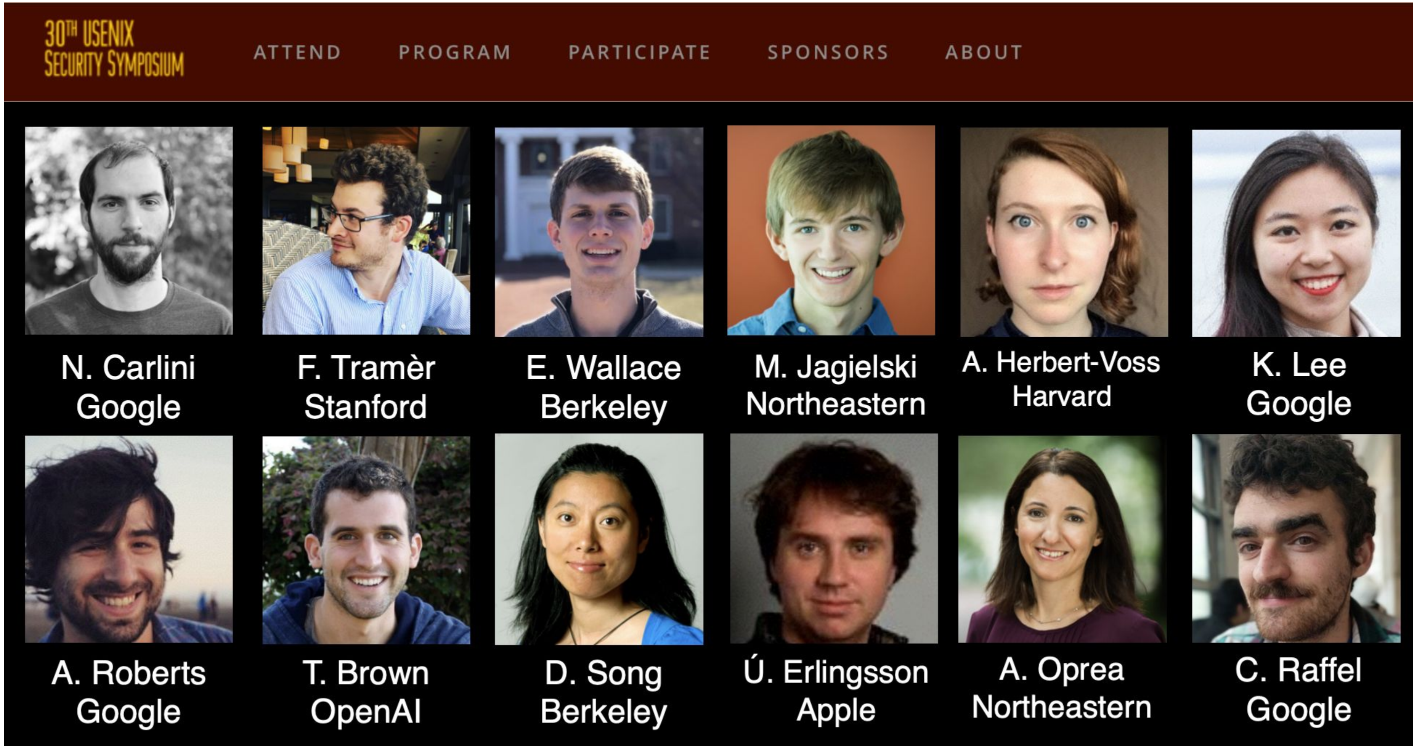
Public Available

Private Concerns?

[Image Source](#) [Image Source](#)

Extracting Training Data from Large Language Models

USENIX 2021



Some slides adapted from presentations of Carlini

Victim Model Overview

- **GPT-2**
  - **State of The Art**

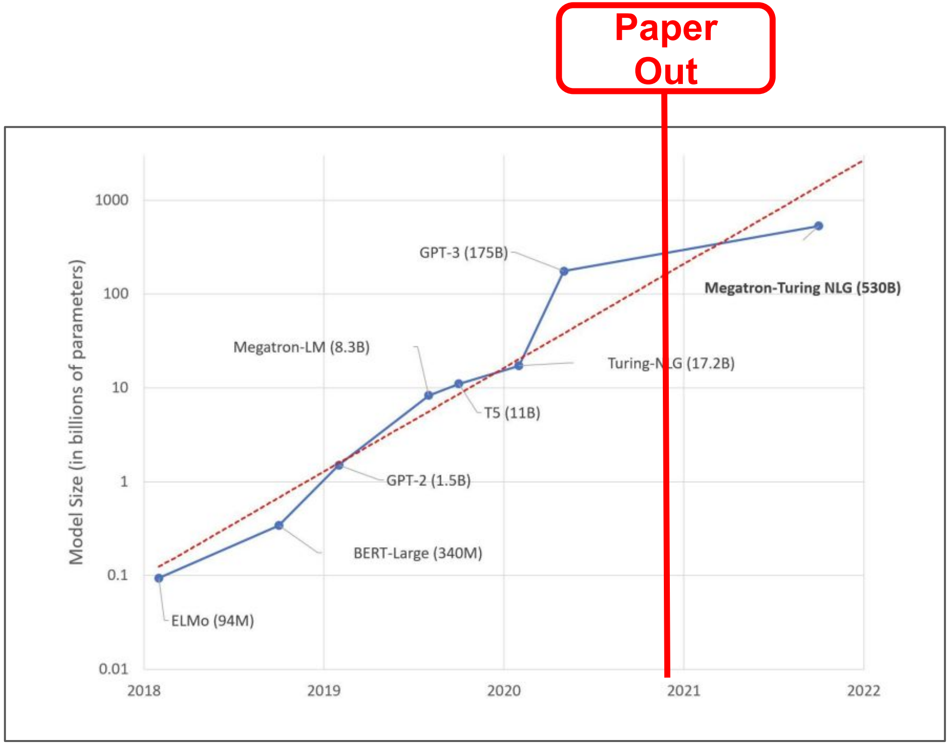


Image Source

Victim Model Overview

- **GPT-2**
  - State of The Art Model
  - **Public Available** (training is done)

GPT-2: 1.5B Release

November 5, 2019
4 minute read

As the final model release of GPT-2's staged release, we're releasing the largest version (1.5B parameters) of GPT-2 along with code and model weights to facilitate detection of outputs of GPT-2 models. While there have been larger language models released since August, we've continued with our original staged release plan in order to provide the community with a test case of a full staged release process. We hope that this test case will be useful to developers of future powerful models, and we're actively continuing the conversation with the AI community on responsible publication.

REPORT

GPT-2 MODEL

Image Source

Victim Model Overview

- **GPT-2**
  - State of The Art Model
  - Public Available
  - **Public (private) WebText data**
    - Scraped from the public Internet
    - 40 GB of text data from over 8M documents

Victim Model Overview

- **Models**:
  - **GPT-2 variant of Transformer LMs**

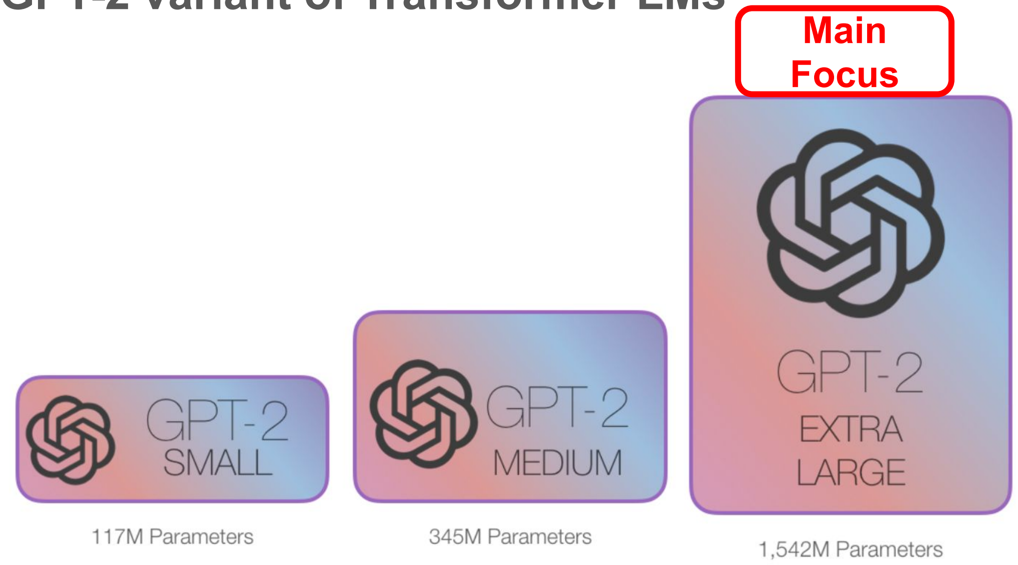


Image Source

Victim Model Overview

- **Training Objective:**

$$\mathcal{L}(\theta) = -\log \Pi_{i=1}^{n} f_\theta(\boxed{x_i} \,|\, \boxed{x_1, \ldots, x_{i-1}})$$

Previous Tokens

Victim Model Overview

- Training Objective:

$$\mathcal{L}(\theta) = -\log \Pi_{i=1}^{n} f_\theta(\boxed{x_i} \| \boxed{x_1, \ldots, x_{i-1}})$$

Previous Tokens

- **Optimal Solution:**
  - **Memorizing** the answer token given the previous tokens

Victim Model Overview

- **Generating Text:**

$$\boxed{\hat{x}_{i+1}} \sim f_\theta\left(x_{i+1} \mid x_1, \ldots, x_i\right)$$

$$\hat{x}_{i+2} \sim f_\theta\left(x_{i+2} \mid x_1, \ldots, x_i, \boxed{x_{i+1}}\right)$$

$$\vdots$$

Repeated process

Threat Model

- **Adversary's Capabilities:**
  - A **black-box** input-output access to a language model.
  - Adversary can
    - compute the probability of arbitrary sequences
    - obtain next-word predictions.

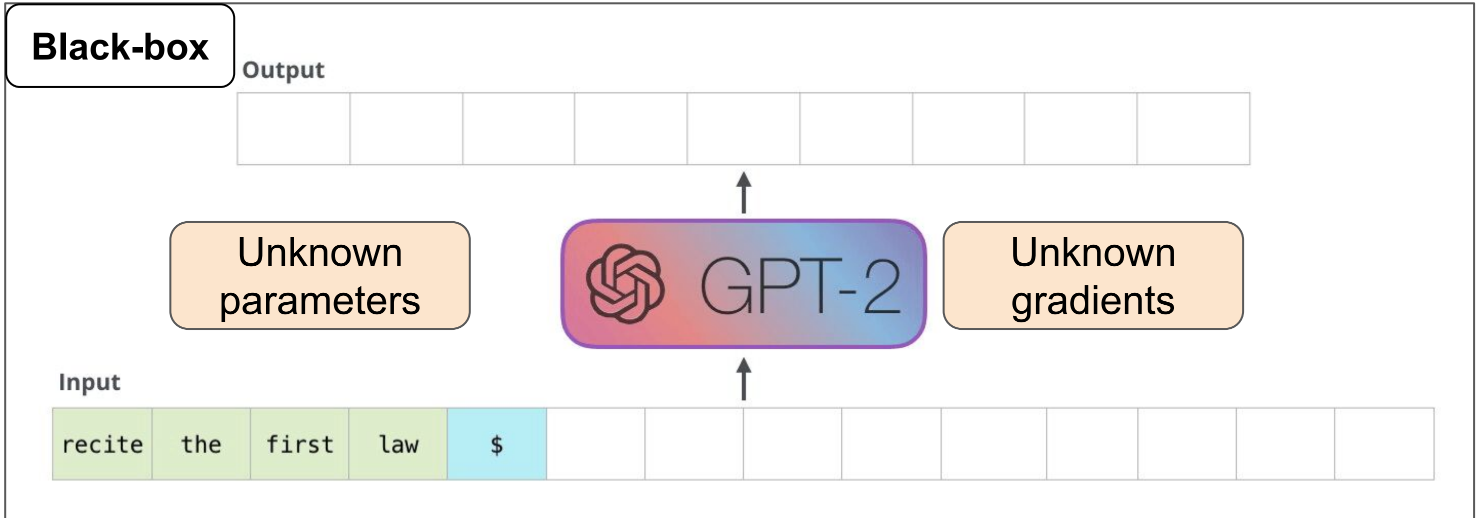


Image Source

Threat Model

- Adversary's Capabilities:
  - A black-box input-output access to a language model.
  - Adversary can
    - compute the probability of arbitrary sequences
    - obtain next-word predictions.

- **Adversary's Objective:**
  - Extract memorized training data from the model.

Measurement?

Measurement

- **Evaluating Memorization Using Manual Inspection**
    - Internet searches for sample, and check if the returning page is **exactly** the same.

Measurement

- Evaluating Memorization Using Manual Inspection
  - Internet searches for sample, and check if the returning page is **exactly** the same.

- **Validating Results on the Original Training Data**
  - Works with GPT-2 authors
  - Fuzzy match with training data

Threat Model

- Adversary's Capabilities:
  - A black-box input-output access to a language model.
  - Adversary can
    - compute the probability of arbitrary sequences
    - obtain next-word predictions.

- **Adversary's Objective:**
  - Extract memorized training data from the model.
  - The attack strength of is measured by how private a particular extracted example is.

Measurement?

Defining Language Model Memorization

- Memorization is essential in many ways (No privacy concerns).
- Beneficial Memorization:
    - **Memorizing** the correct **spellings** of words

Defining Language Model Memorization

- Memorization is essential in many ways (No privacy concerns).

- Beneficial Memorization:
    - Memorizing the correct spellings of words
    - **Memorizing the common knowledge**:
        - Prefix: “My address is 1 Main Street, San Francisco CA”,
        - Model generates “94107” which is a correct zip code for San Francisco, CA

Defining Language Model Memorization

Definition 1 (Model Knowledge Extraction) *A string s is* extractable[4] *from an LM f_θ if there exists a prefix c such that:*

$$s \leftarrow \boxed{\mathop{\arg\max}}_{s' : |s'|=N} f_\theta(s' \mid c)$$

An appropriate sampling strategy

String *s* can be generated from an LLM

k-Eidetic Memorization

Definition 2 (k-Eidetic Memorization) *A string s is k-eidetic memorized (for $k \geq 1$) by an LM f_θ if s is extractable from f_θ and s appears in $\boxed{\text{at most k examples}}$ in the training data X: $|\{x \in X : s \subseteq x\}| \leq k$.*

***s* is likely to be private if it only appears few times.**

k-Eidetic Memorization

- Memorizing the **correct spellings** of one particular word is not severe. (k is large)
- Memorizing the zip code of a particular city might be eidetic memorization (depends on k)
- Memorizing an **individual person's name and phone number** clearly (informally) violates privacy expectations (k is small)

Pre-Lecture Question

Q1. Describe what assumptions Carlini et al. make for their threat models and how they measure the success of their training-data extraction methods.

- Threat models
  - Adversary’s Capabilities: A black-box access to a LM.
  - Adversary’s Objective: Extract private memorized training data.
  - Adversary’s Target: GPT-2 and its variants
- Measurement of the extraction method:
  - Manual Inspection
  - Fuzzy match
  - Evaluated the private degree by k-Eidetic memorization

Training Data Extraction Attack Overview

- **Generate a lot of text from LM**
- **Membership Inference**

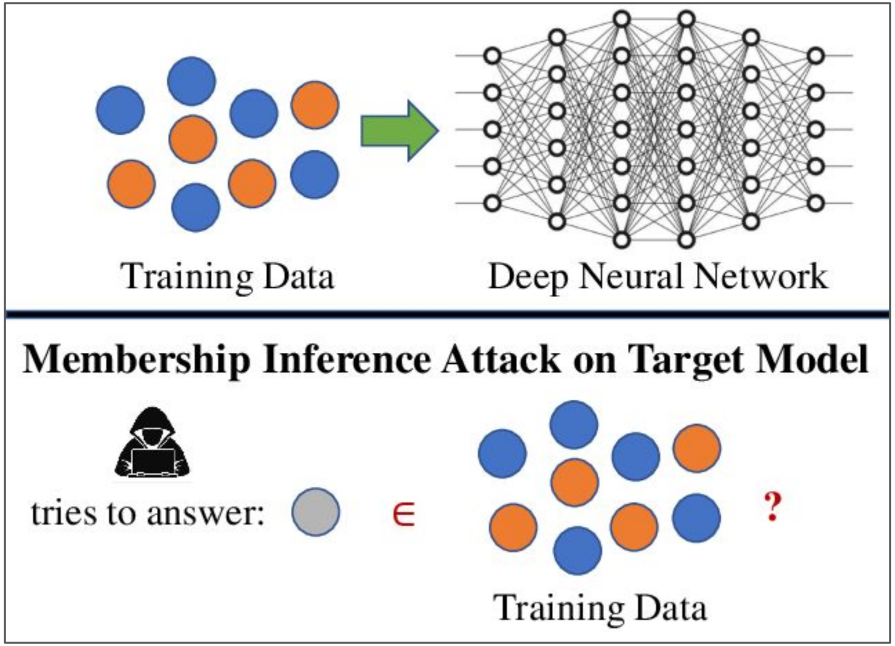


Image Source

Initial Training Data Extraction Attack

- **Initial Text Generation Scheme**
  - generate from **one-token prompt** by sampling with **likelihood**

Initial Training Data Extraction Attack

- Initial Text Generation Scheme
  - generate with **one-token prompt** by sampling with **likelihood**

- **Initial Membership Inference**
  - Predicting whether each sample was present in the training data by **perplexity**:

$$\mathcal{P} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log f_\theta(x_i|x_1,\ldots,x_{i-1})\right)$$

Low perplexity means the model assign high probability

Initial Training Data Extraction Attack

- **Initial Extraction Results**
  - Generate 200,000 samples, sort according to perplexity
  - **Interesting Findings** but (large k-eidetic memorization):

Image Source Image Source

Initial Training Data Extraction Attack

- **Initial Extraction Results**
  - Generate 200,000 samples, sort according to perplexity
  - **Interesting Findings** but (large k-eidetic memorization):

Initial Attack Failed

Image Source Image Source

Initial Attack Failed

- **Sampling scheme tends to produce a low diversity of outputs.**

Image Source

Initial Training Data Extraction Attack

- Sampling scheme tends to produce a low diversity of outputs.
- **Initial membership inference has large false positives**
  - High likelihood to **repetitive** sequences

I love you. I love you. I love you. I love you…

Improved Text Generation Schemes: Temperature

- Sampling with **a decaying temperature**
  - Temperature can cause the model **less confident** and **more diverse** for the output.
  - A decaying temperature then
    - gives a sufficient diverse set of prefixes
    - follows a high-confidence paths

$$\frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

T: Temperature

Improved Text Generation Schemes: Using Internet Text

- **Conditioning** on **Internet** Text
  - Exploring prefixes from text scraped from the Internet

[Image Source](#)

Improved Membership Inference

- Many uninteresting samples that are assigned spuriously high likelihood

Method: Filtering out these uninteresting (yet still high-likelihood samples) by comparing to a second LM

Improved Membership Inference

- **Comparing to Other Neural Language Models**
  - Train a **smaller** GPT-2 model on same training set.
  - Smaller models have less memorization.

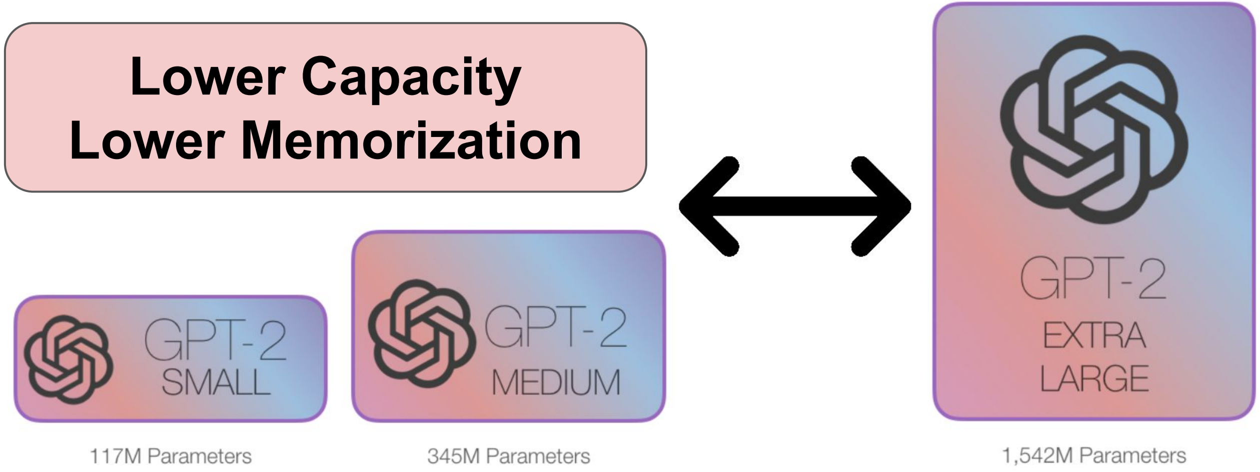


Image Source

Improved Membership Inference

- Comparing to Other Neural Language Models
- **Comparing to zlib Compression Entropy**
    - Repeated data reduces zlib Compression Entropy

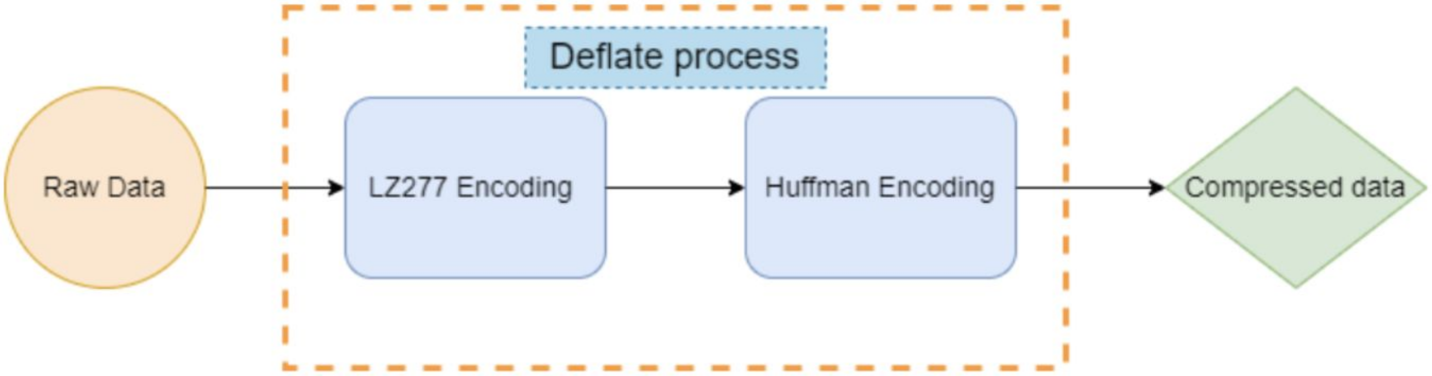


Image Source

Improved Membership Inference

- Comparing to Other Neural Language Models
- Comparing to zlib Compression Entropy
- **Comparing to Lowercased Text**
  - Comparing the **perplexity** before and after lowercasing all samples

Perplexity("Extract Large Language Model …")

↕

Perplexity("extract large language model …")

Improved Membership Inference

- Comparing to Other Neural Language Models
- Comparing to zlib Compression Entropy
- Comparing to Lowercased Text
- **Perplexity on a Sliding Window**
    - Memorized token surrounded by non-memorized tokens

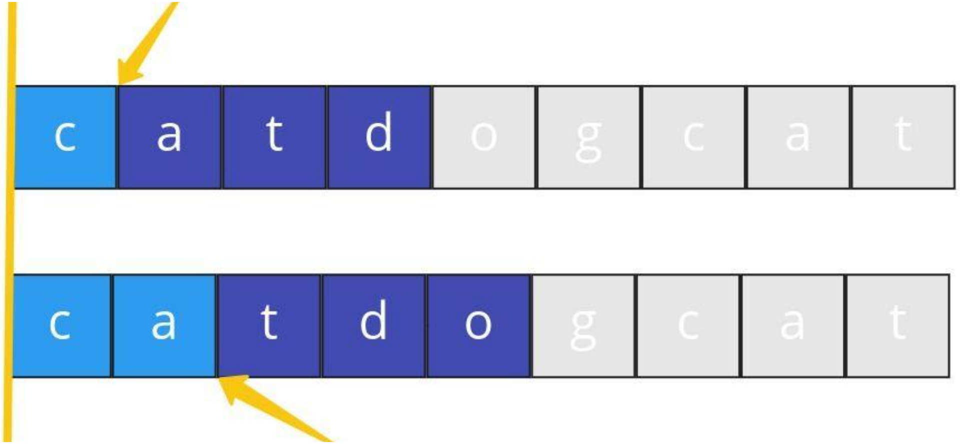


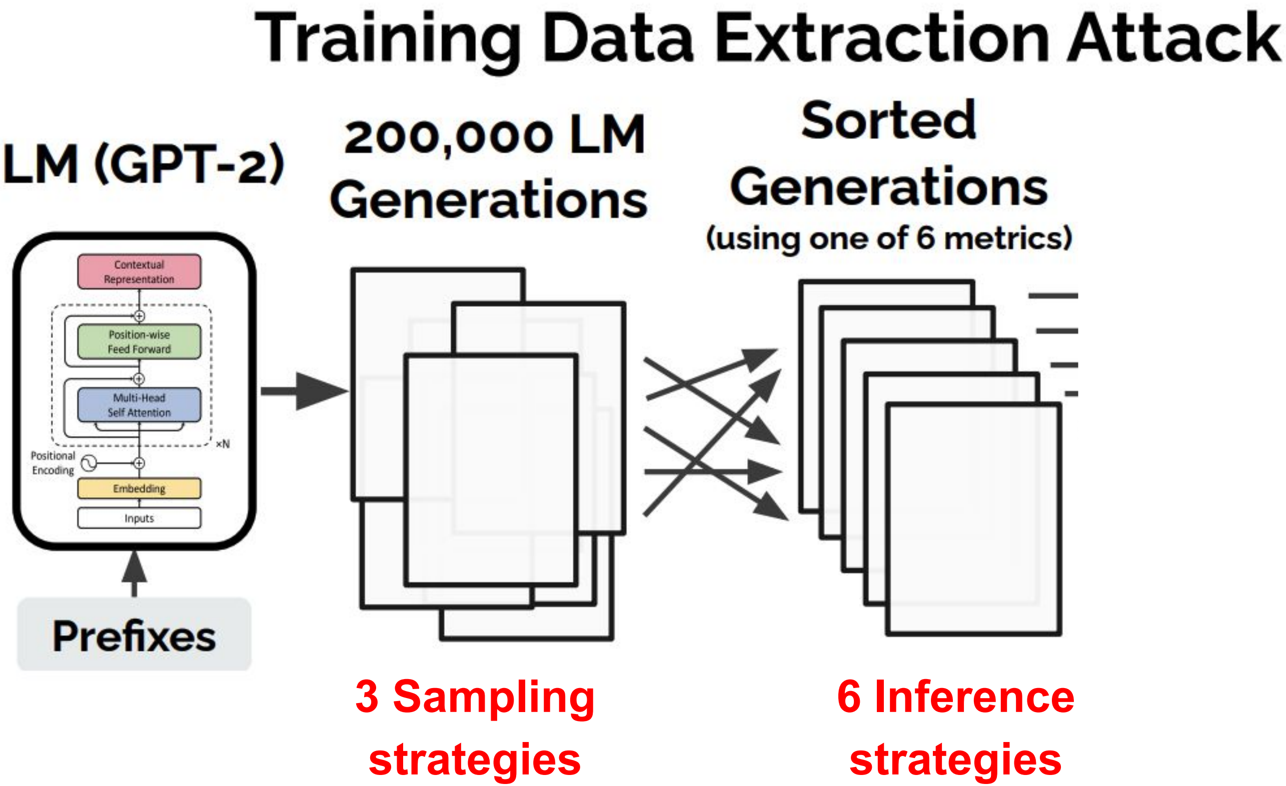
Image Source

Pre-Lecture Question

Q2. Carlini et al. presented their initial (and naive) attack results but they were not successful. What improvements did they make after the initial attempt?

- **Improved Text Generation Schemes:**
  - Sampling With A Decaying Temperature
  - Conditioning on Internet Text
- **Improved Membership Inference:**
  - Comparing to Other Neural Language Models
  - Comparing to zlib Compression Entropy
  - Comparing to Lowercased Text
  - Perplexity on a Sliding Window

Pipeline



Training Data Extraction Attack
LM (GPT-2)
Contextual Representation
Position-wise Feed Forward
Multi-Head Self Attention
×N
Positional Encoding
Embedding
Inputs
Prefixes
200,000 LM Generations
Sorted Generations
(using one of 6 metrics)
3 Sampling strategies
6 Inference strategies

Memorization: Evaluation

- **3 Sampling strategies**
  - Top-n
  - Temperature
  - Internet

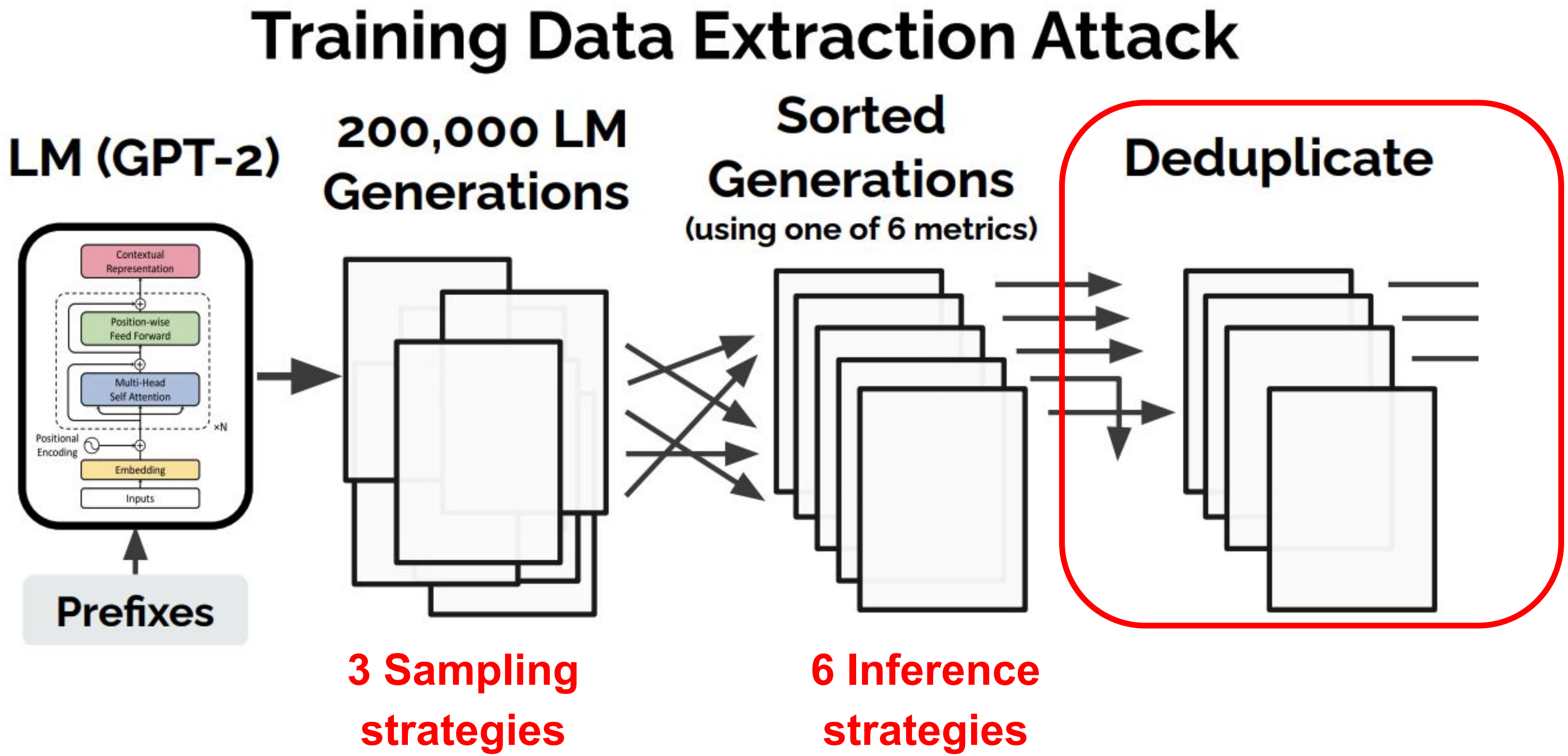
X

- **6 Inference strategies**
  - Perplexity
  - Small (second LM)
  - Medium (second LM)
  - zlib
  - Lowercase
  - Window

Memorization: Evaluation

- **Configurations**
  - Generating three datasets: 3 x 200,000 samples
  - For each dataset, applying 6 inference methods and select 100 samples from top-1000 samples.
  - 3 x 6 different configurations to extract training data
  - **Result: 1,800** total samples of potentially memorized content

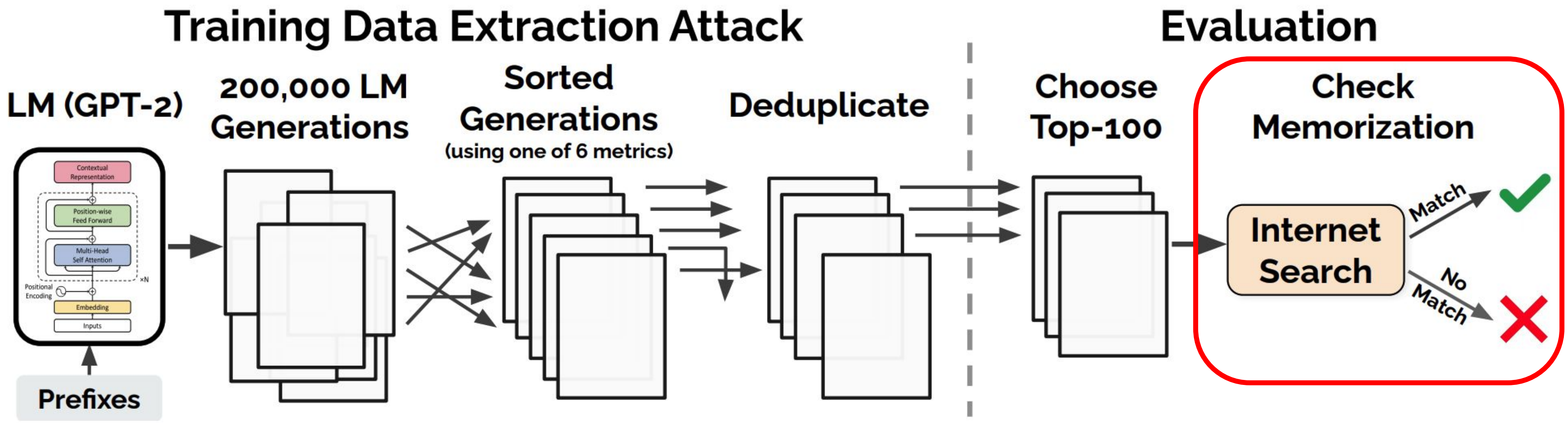
Pipeline


Training Data Extraction Attack
LM (GPT-2)
200,000 LM Generations
Sorted Generations
(using one of 6 metrics)
Deduplicate
Contextual Representation
Position-wise Feed Forward
Multi-Head Self Attention
×N
Positional Encoding
Embedding
Inputs
Prefixes
3 Sampling strategies
6 Inference strategies

Data Deduplication

- **Avoid** “double-counting” memorized content
- Trigram-multiset
  - “my name my name my name” has two trigrams (“my name my” and ”name my name”)
  - If two samples have similar trigram multisets, then they are duplicates

Pipeline



Training Data Extraction Attack
Evaluation
LM (GPT-2)
Contextual Representation
Position-wise Feed Forward
Multi-Head Self Attention
×N
Positional Encoding
Embedding
Inputs
Prefixes
200,000 LM Generations
Sorted Generations
(using one of 6 metrics)
Deduplicate
Choose Top-100
Check Memorization
Internet Search
Match
No Match

Results

Identify **604** unique memorized training examples from among the **1,800** possible candidates

Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

Results

Inference Strategy	Text Generation Strategy		
	Top-n	Temperature	Internet
Perplexity	9	3	39
Small	41	42	58
Medium	38	33	45
zlib	59	46	67
Window	33	28	58
Lowercase	53	22	60
Total Unique	191	140	273

Results

Inference Strategy	Text Generation Strategy		
	Top-n	Temperature	Internet
Perplexity	9	3	39
Small	41	42	58
Medium	38	33	45
zlib	59	46	67
Window	33	28	58
Lowercase	53	22	60
Total Unique	191	140	273

Examples of Memorized Content

- Personally Identifiable Information
  - **46** examples that contain individual peoples' name (omit samples related to news)
  - **32** examples that contain contact information (16 businesses contact, **16 private contact**)

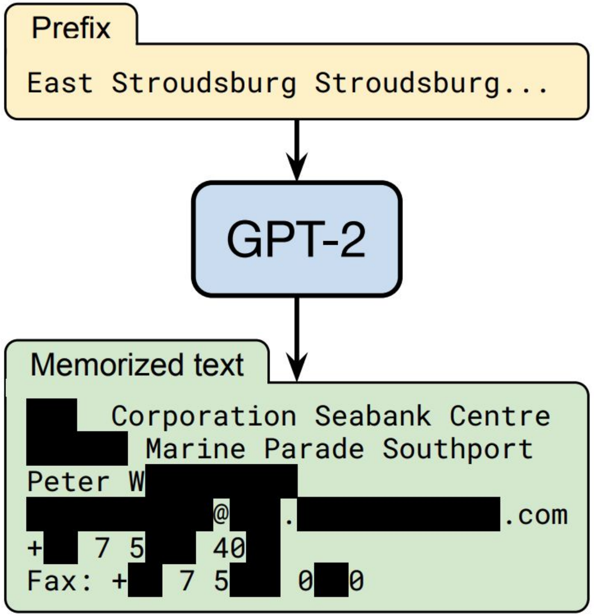


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Results

Category	Count	Description
US and international news	109	General news articles or headlines, mostly about US politics
Log files and error reports	79	Logs produced by software or hardware
License, terms of use, copyright notices	54	Software licenses or website terms of use, copyright for code, books, etc.
Lists of named items	54	Ordered lists, typically alphabetically, of games, books, countries, etc.
Forum or Wiki entry	53	User posts on online forums or entries in specific wikis
Valid URLs	50	A URL that resolves to a live page
Named individuals	46	Samples that contain names of real individuals. We limit this category to *non-news samples*. E.g., we do not count names of politicians or journalists within news articles
Promotional content	45	Descriptions of products, subscriptions, newsletters, etc.
High entropy	35	Random content with high entropy, e.g., UUIDs Base64 data, etc.

Category	Count	Description
Contact info	32	Physical addresses, email addresses, phone numbers, twitter handles, etc.
Code	31	Snippets of source code, including JavaScript
Configuration files	30	Structured configuration data, mainly for software products
Religious texts	25	Extracts from the Bible, the Quran, etc.
Pseudonyms	15	Valid usernames that do not appear to be tied to a physical name
Donald Trump tweets and quotes	12	Quotes and tweets from Donald Trump, often from news articles
Web forms	11	Lists of user menu items, Website instructions, navigation prompts (e.g., "please enter your email to continue")
Tech news	11	News related to technology
Lists of numbers	10	Lists of dates, number sequences, π, etc.
Sports news	9	News related to sports
Movie synopsis, cast	5	List of actors, writers, producers. Plot synopsis.
Pornography	5	Content of pornographic nature, often lists of adult film actors.

Examples of Memorized Content

- Unnatural Text
  - **21** examples of random number sequences with at least 50 bits of entropy
  - **9** examples of k = 1 eidetic memorized content

Memorized String	**Sequence Length**	**Occurrences in Data**	
		Docs	**Total**
`Y2...`[redacted]`...y5`	87	1	10
`7C...`[redacted]`...18`	40	1	22
`XM...`[redacted]`...WA`	54	1	36
`ab...`[redacted]`...2c`	64	1	49
`ff...`[redacted]`...af`	32	1	64
`C7...`[redacted]`...ow`	43	1	83
`0x...`[redacted]`...C0`	10	1	96
`76...`[redacted]`...84`	17	1	122
`a7...`[redacted]`...4b`	40	1	311

Table 3: **Examples of $k = 1$ eidetic memorized, high-entropy content that we extract** from the training data. Each is contained in *just one* document. In the best case, we extract a 87-characters-long sequence that is contained in the training dataset just 10 times in total, all in the same document.

Correlating Memorization with Model Size & Insertion Frequency

- Two Questions of Interest
  - How many times a string must appear for it to be memorized?
  - How does the model size impact the memorization?

Correlating Memorization with Model Size & Insertion Frequency

- Two Questions of Interest
  - How many times a string must appear for it to be memorized?
  - How does the model size impact the memorization?

- A Case Study: probe the memorization of GPT-2 on reddit urls.

Correlating Memorization with Model Size & Insertion Frequency

- Two Questions of Interest
  - How many times a string must appear for it to be memorized?
  - How does the model size impact the memorization?

- A Case Study: probe the memorization of GPT-2 on reddit urls.
  - Prompt GPT-2 with the prefix :

```
{"color":"fuchsia","link":"https://www.
reddit.com/r/The_Donald/comments/
```

Correlating Memorization with Model Size & Insertion Frequency

- Two Questions of Interest
  - How many times a string must appear for it to be memorized?
  - How does the model size impact the memorization?

- A Case Study: probe the memorization of GPT-2 on reddit urls.
  - Prompt GPT-2 with the prefix :

```
{"color":"fuchsia","link":"https://www.
reddit.com/r/The_Donald/comments/
```

  - Use top-n sampling to generate 10,000 possible extensions, and test whether any URLs in the training document were generated.

Correlating Memorization with Model Size & Insertion Frequency

A Case Study: probe the memorization of GPT-2 on reddit urls

- Setup
  - Test on GPT-2 models with different sizes — XL (1.5B), M (345M), S (117M)

	Occurrences		Memorized?		
URL (trimmed)	**Docs**	**Total**	**XL**	**M**	**S**
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

Correlating Memorization with Model Size & Insertion Frequency

A Case Study: probe the memorization of GPT-2 on reddit urls

- Setup
  - Test on GPT-2 models with different sizes — XL (1.5B), M (345M), S (117M)
  - Look into urls with different number of occurrences in the training dataset.

	Occurrences		Memorized?		
URL (trimmed)	**Docs**	**Total**	**XL**	**M**	**S**
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

Correlating Memorization with Model Size & Insertion Frequency

A Case Study: probe the memorization of GPT-2 on reddit urls

- Results
  - Larger models can memorize more.

URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

Correlating Memorization with Model Size & Insertion Frequency

A Case Study: probe the memorization of GPT-2 on reddit urls

- Results
  - Larger models can memorize more.
  - Models tend to memorize texts with higher number of occurrences.

	Occurrences		Memorized?		
URL (trimmed)	Docs	Total	XL	M	S
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

Correlating Memorization with Model Size & Insertion Frequency

A Case Study: probe the memorization of GPT-2 on reddit urls

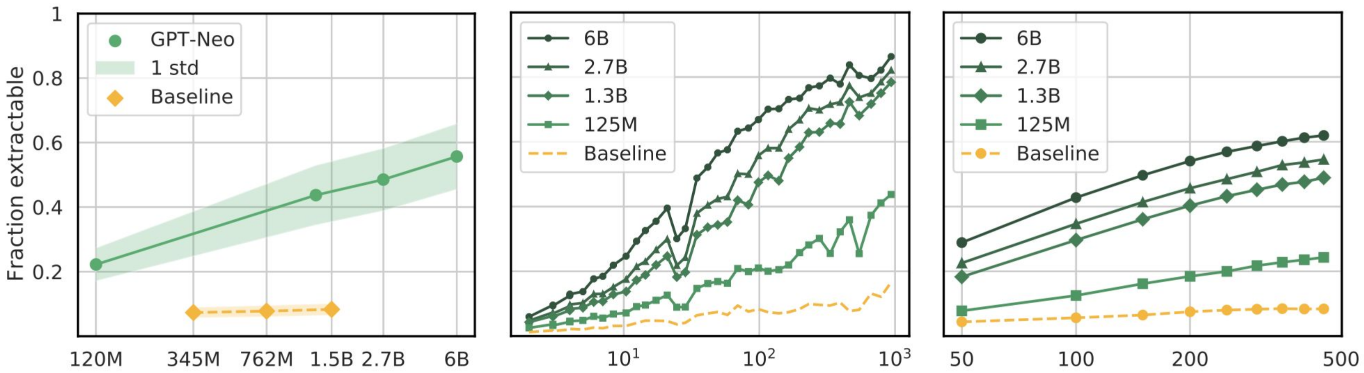
- Limitations: only identify a narrow relationship — i.e. qualitatively study the ability to memorize < 30 URLs…

	Occurrences		Memorized?		
URL (trimmed)	**Docs**	**Total**	**XL**	**M**	**S**
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

More Quantitative Studies on The Factors That Impact Memorization

Quantifying memorization across neural language models, Carlini et al.

Protocol: (1) Directly use prefixes of the original training examples as prompts; (2) verifying whether the model has the ability to complete the rest of the example verbatim.

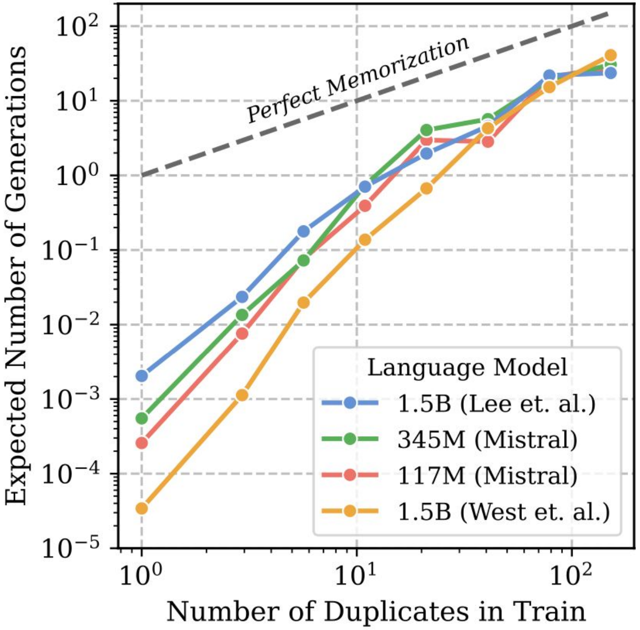


(a) Model scale (b) Data repetition (c) Context size

More Quantitative Studies on The Factors That Impact Memorization

Deduplicating Training Data Mitigates Privacy Risks in Language Models, Kandpal et al.

Protocol: do unconditional generations and report the expected number of generations w.r.t number of duplicates (occurrences) of training sequences.



Mitigating Privacy Leakage

- **Training with Differential Privacy**
  - The key idea of differential privacy: with a differentially private training algorithm, the existence or absence of any single training sample/entry will not result in a “significantly” different model.

Mitigating Privacy Leakage

- **Training with Differential Privacy**
    - The key idea of differential privacy: with a differentially private training algorithm, the existence or absence of any single training sample/entry will not result in a “significantly” different model.

      => Intuitively, models generated by a differentially private training algorithm should not “significantly” memorize any single training sample/entry.

Mitigating Privacy Leakage

- **Training with Differential Privacy**
    - The key idea of differential privacy: with a differentially private training algorithm, the existence or absence of any single training sample/entry will not result in a “significantly” different model.
    - Widely used algorithm: differentially private stochastic gradient descent (DP-SGD), which adds noise to gradients during training.

Mitigating Privacy Leakage

- **Training with Differential Privacy**
  - The key idea of differential privacy: with a differentially private training algorithm, the existence or absence of any single training sample/entry will not result in a “significantly” different model.
  - Widely used algorithm: differentially private stochastic gradient descent (DP-SGD), which adds noise to gradients during training.
  - Differential privacy probably won’t save the day!
    (1) tradeoffs between privacy and utility
    (2) do not prevent memorization of information that occurs across a large number of records
    ….

Mitigating Privacy Leakage

- **Curating The Training Data**
  - Carefully source the training data.
    E.g. avoid websites that are known to host sensitive content

Mitigating Privacy Leakage

- **Curating The Training Data**
  - Carefully source the training data.
  - Limit the amount of sensitive content that are present in the training data.
    E.g. identify and filter personal information or content with restrictive terms of use.

Mitigating Privacy Leakage

- **Curating The Training Data**
  - Carefully source the training data.
  - Limit the amount of sensitive content that are present in the training data.
  - Deduplicate Training Data.

    > (Kandpal et al., 2022) : after deduplicating training data in sequences level, Carlini's attacks are much less effective.

		Normal Model	Deduped Model
Training Data Generated	Count	1,427,212	68,090
	Percent	0.14	0.007
Mem. Inference AUROC	zlib	0.76	0.67
	Ref Model	0.88	0.87
	Lowercase	0.86	0.68

Mitigating Privacy Leakage

- **Limiting Impact of Memorization on Downstream Applications**
  - A Future Direction: how memorization is inherited by fine-tuned models?
- **Audit Models to Empirically Determine The Privacy Level**

…………..

Lessons

- Extraction attacks are a practical threat.
- Memorization does not require overfitting.
- Large models memorize more data & texts that have higher number of occurrences are more likely to be memorized.

Future Work

- Better prefix selection strategies might identify more memorized data.
- Adopt and develop mitigation strategies for building more private large language models.

Pre-Lecture Question

Q3. Under the same threat model, can you think of any stronger attack methods? What if the adversary also has access to the model weights (and even the gradient information)?