Data Memorization in Models I

AntNLP 2022 Fall Seminar

Ding Xuanwen

Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]

[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Quantifying Memorization Across Neural Language Models

Nicholas Carlini[*1] **Daphne Ippolito**[1,2] **Matthew Jagielski**[1]
Katherine Lee[1,3] **Florian Tramèr**[1] **Chiyuan Zhang**[1]

[1]*Google Research*
[2]*University of Pennsylvania*
[3]*Cornell University*

Counterfactual Memorization in Neural Language Models

Chiyuan Zhang[1] **Daphne Ippolito**[1,2] **Katherine Lee**[1,3]
Matthew Jagielski[1] **Florian Tramèr**[1] **Nicholas Carlini**[1]

[1]*Google Research*
[2]*University of Pennsylvania*
[3]*Cornell University*

Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

TECHNOLOGY FEATURE | 21 April 2020

Deep learning takes on tumours

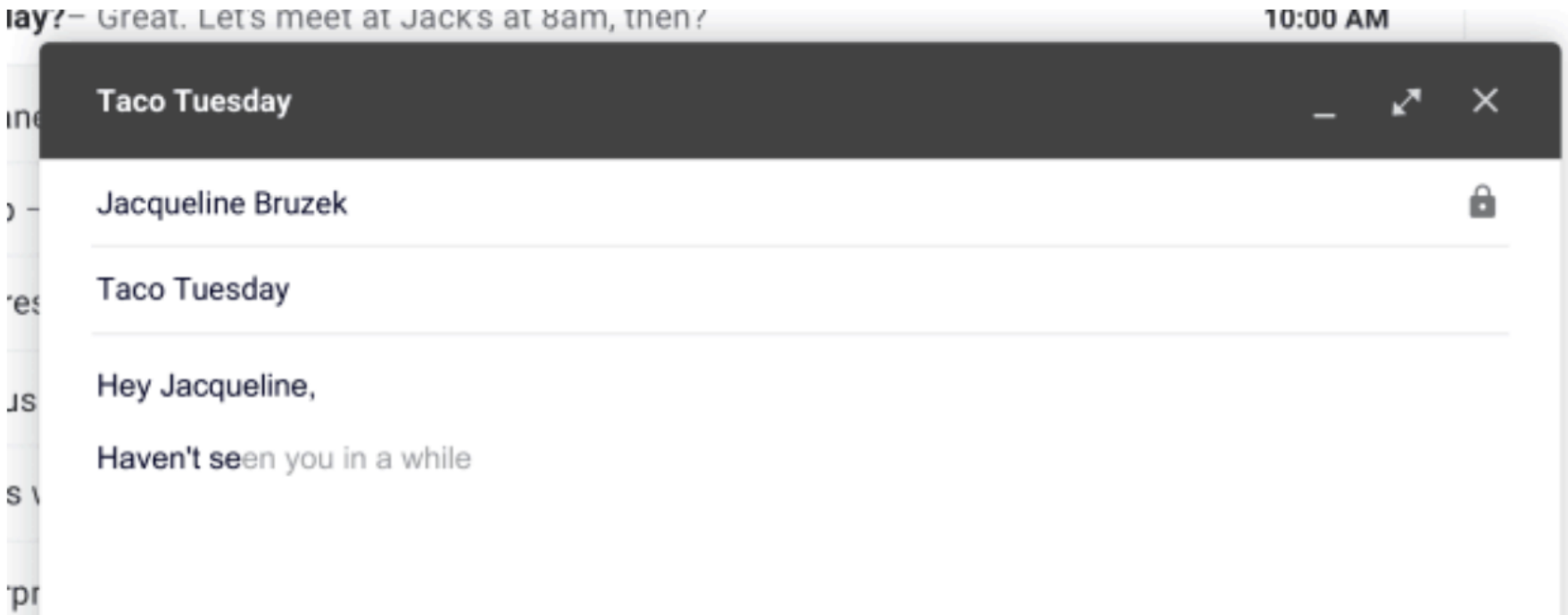
Artificial-intelligence methods are moving into cancer research.

Google

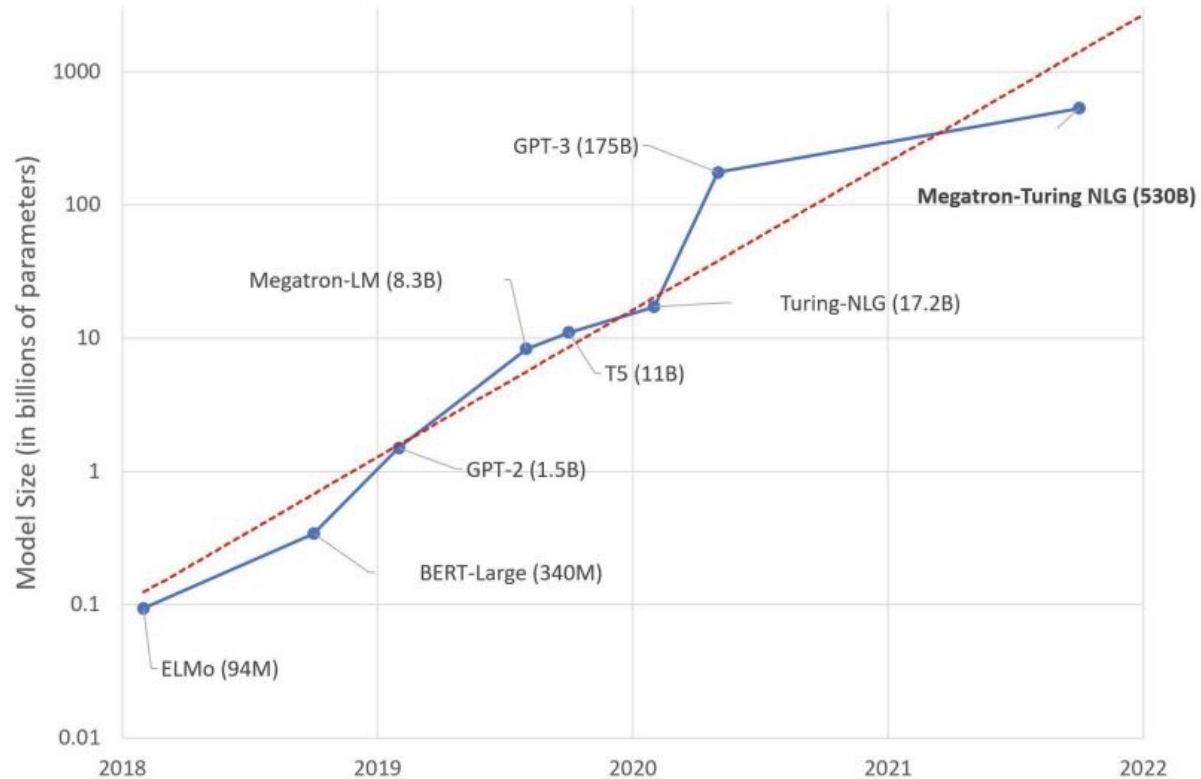
The Keyword | Latest Stories | Product Updates | Company News

GMAIL

SUBJECT: Write emails faster with Smart Compose in Gmail



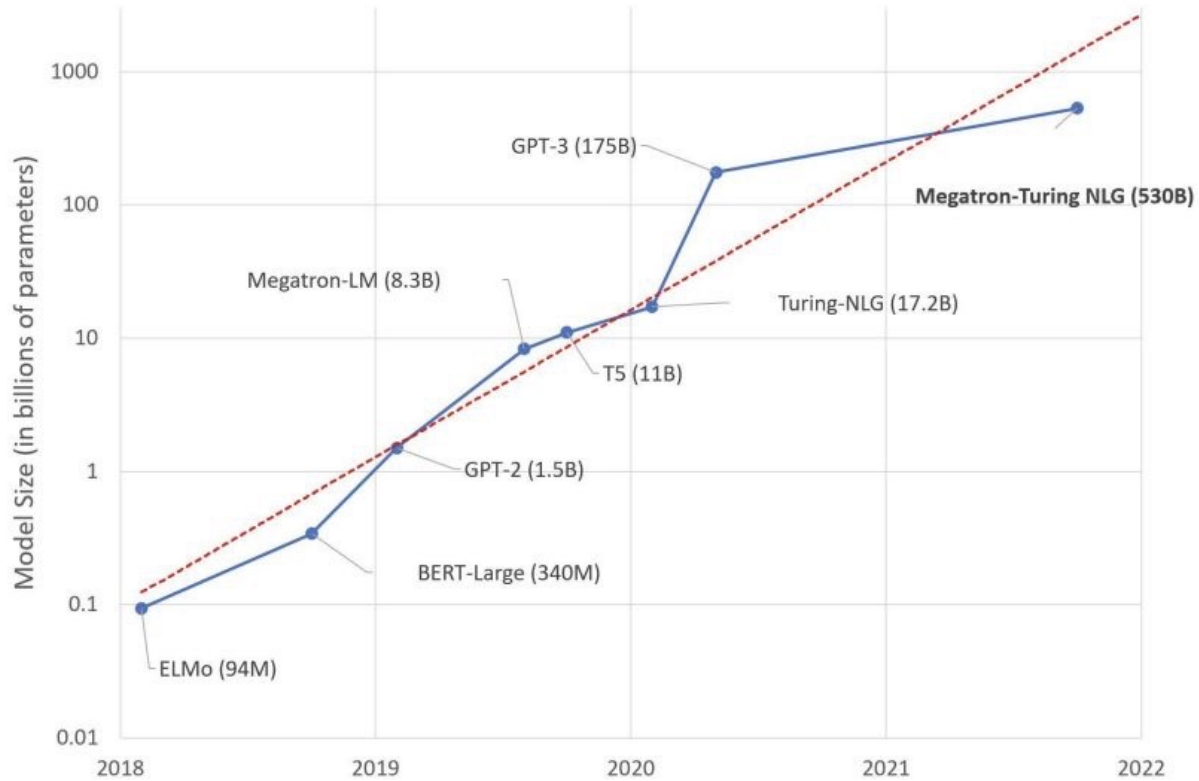
Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion



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

Keep Private

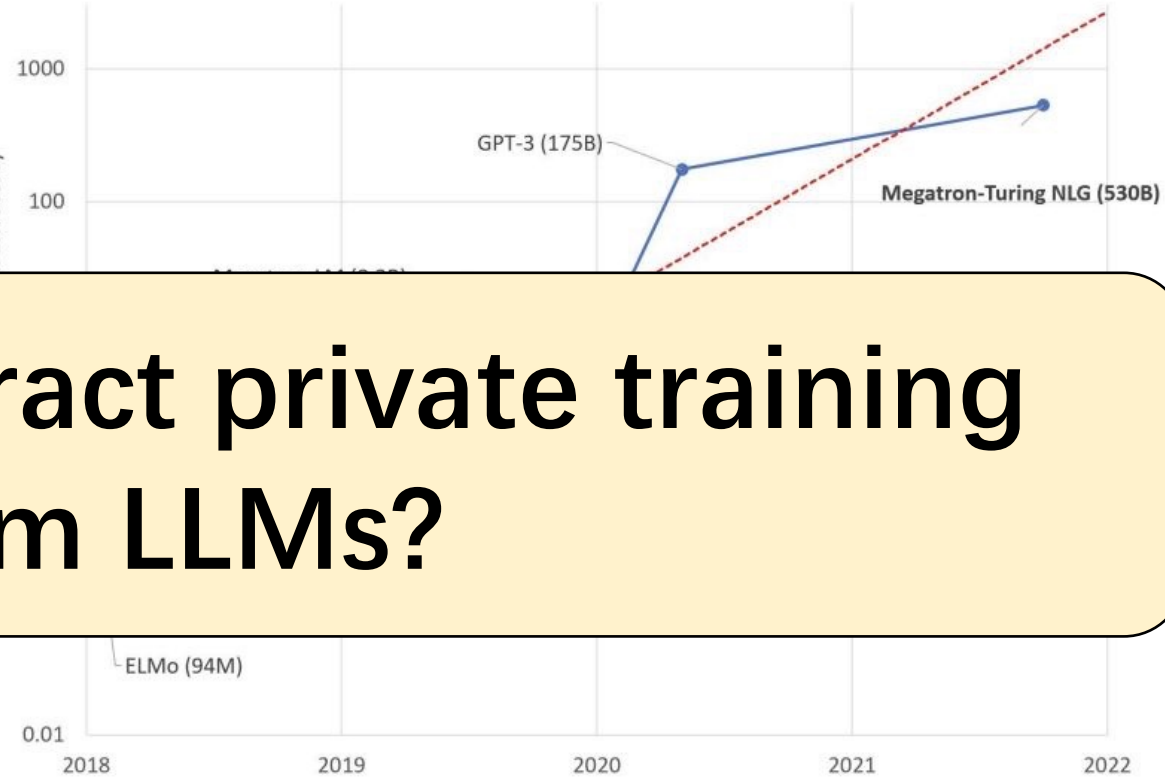
Private Information



Publicly Available

Privacy Concerns?

Quantity
(tokens)



Is it possible to extract private training data from LLMs?

Keep Private

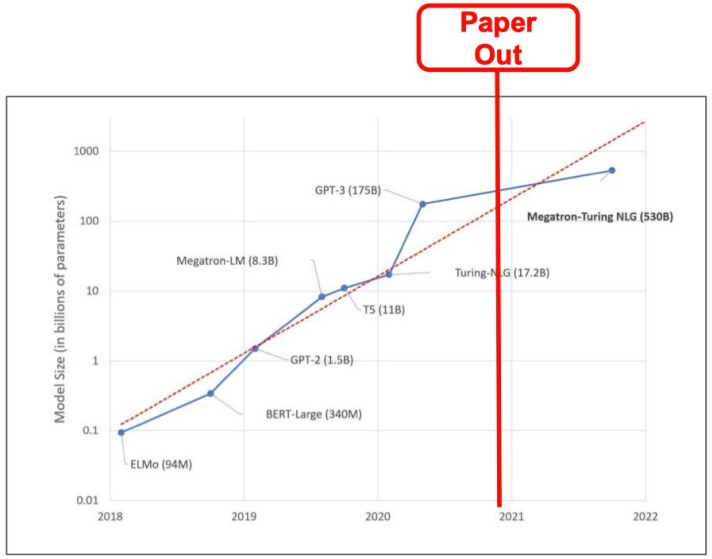
Private Information

Publicly Available

Privacy Concerns?

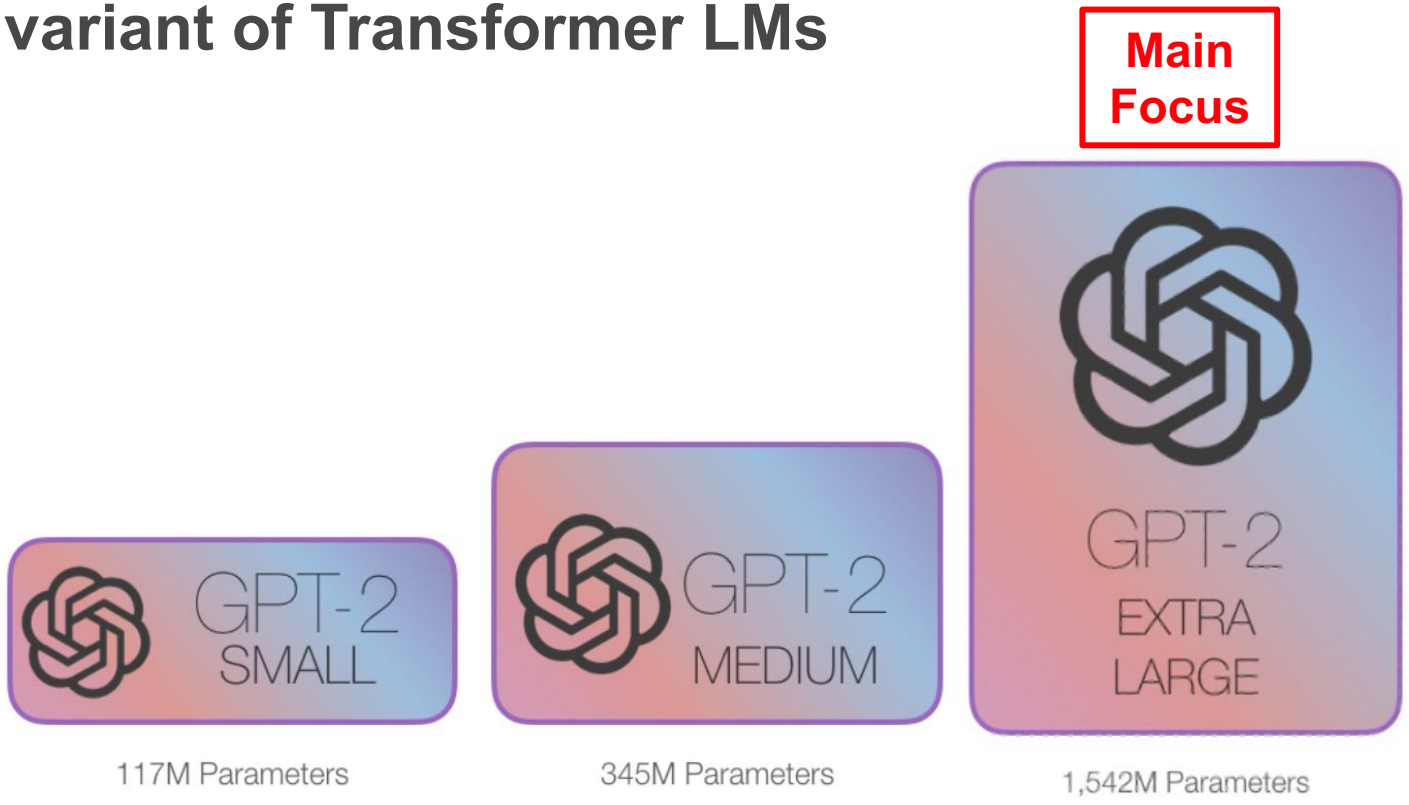
Victim Model Overview

- **GPT2**
  - **State of The Art**
  - **Public Available (training is done)**
  - **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**



Victim Model Overview

- Training Objective:

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

Previous Tokens

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

Victim Model Overview

- **Generating Text:**

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

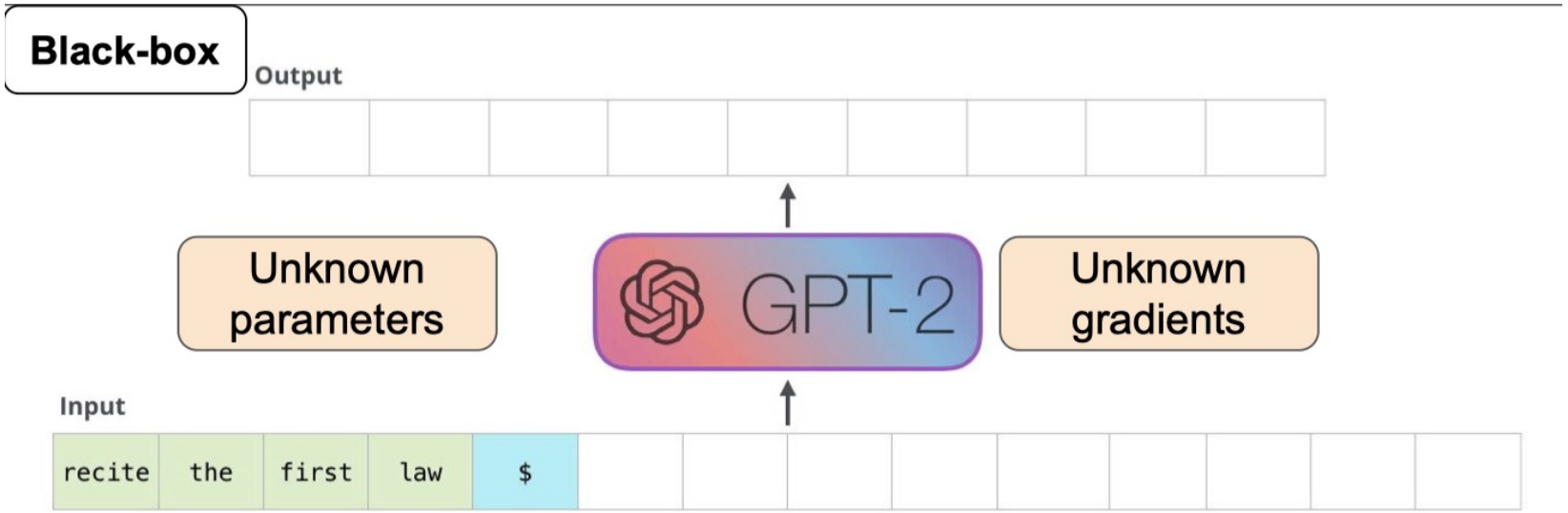
$$\hat{x}_{i+2} \sim f_\theta\left(x_{i+2} \mid x_1, \ldots, x_i, x_{i+1}\right)$$

⋮

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.



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.

- **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
  - **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:*

$$s \leftarrow \underset{s' :\, |s'|=N}{\arg\max} f_\theta(s' \mid c)$$

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 at most k examples in the training data X: $|\{x \in X : s \subseteq x\}| \leq k$.*

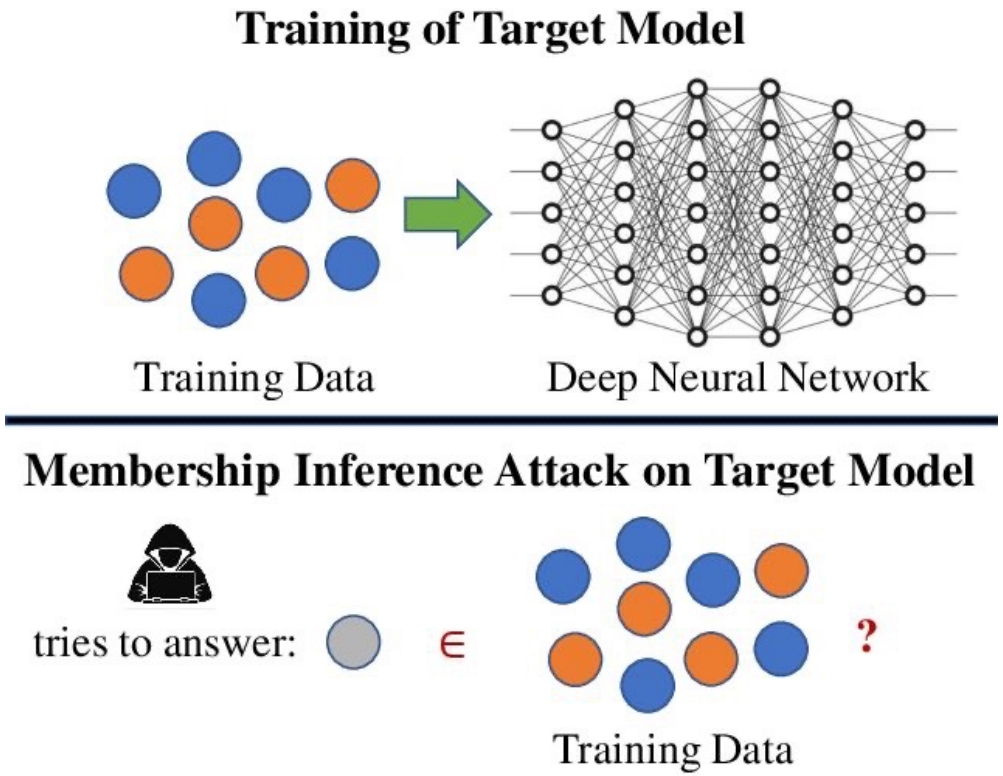
s is likely to be private if it only appears few times.

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)

Training Data Extraction Attack Overview

- Generate a lot of text from LM
- Membership Inference



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**:

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

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

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

Initial Attack Failed

- 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

Common Crawl

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

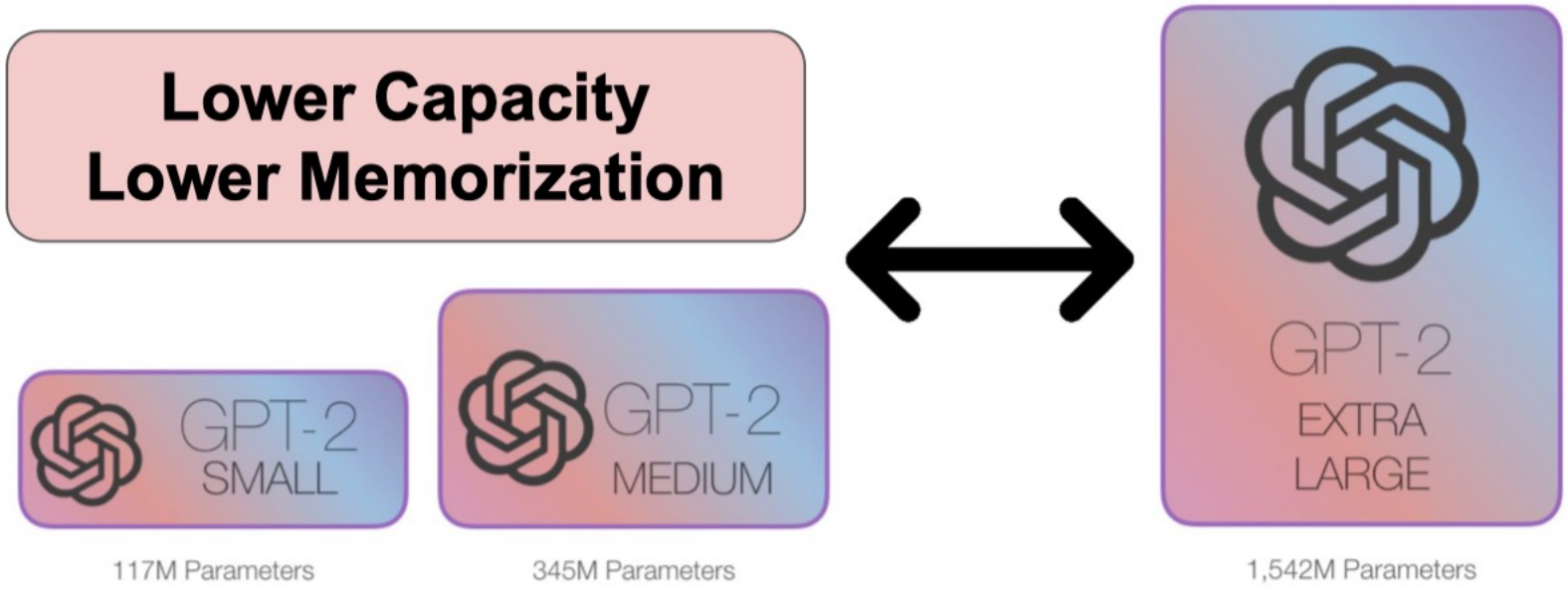
Improved Membership Inference

- 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

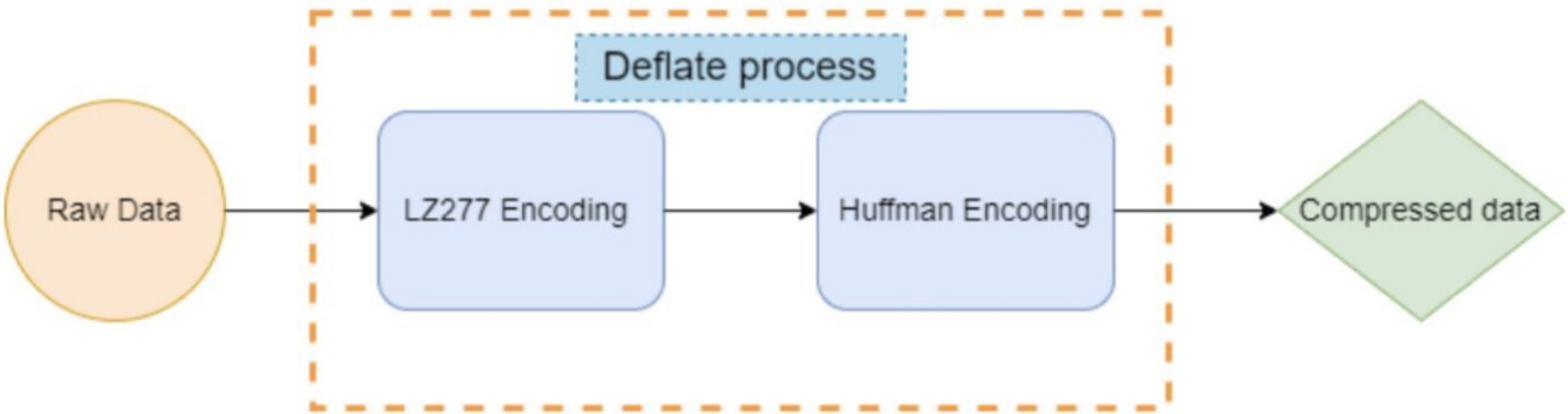
Improved Membership Inference

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



Improved Membership Inference

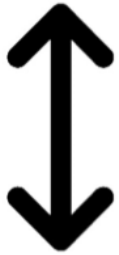
- Comparing to Other Neural Language Models
- **Comparing to zlib Compression Entropy**
  - Repeated data reduces zlib Compression Entropy



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

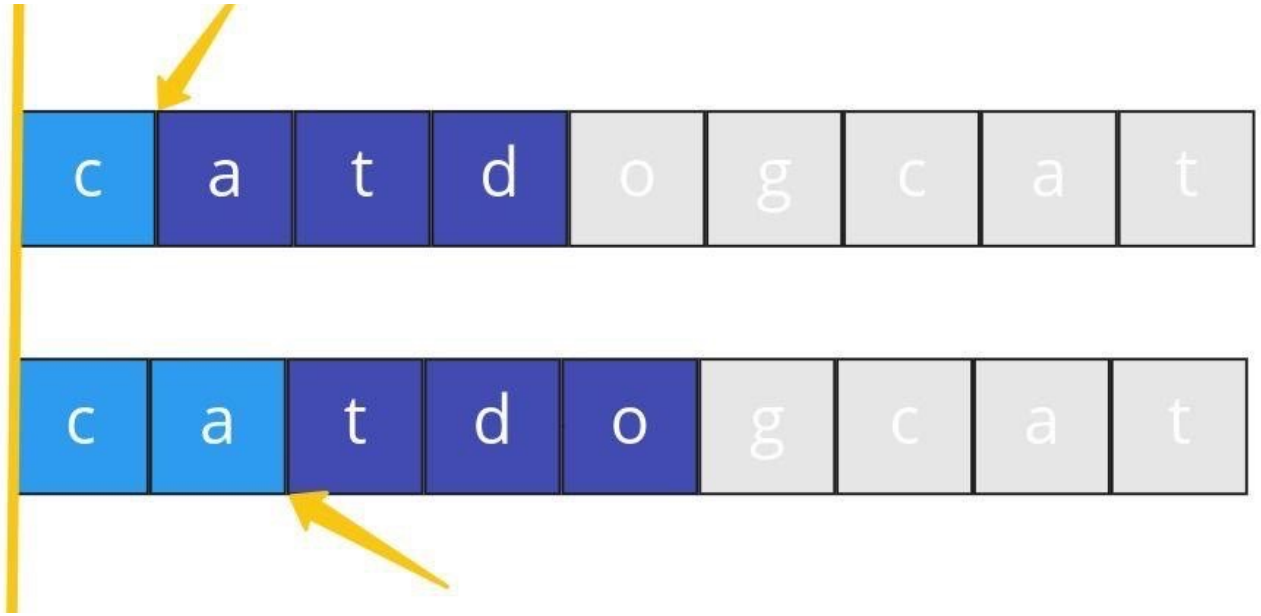
Perplexity("Extract Large Language Model …")



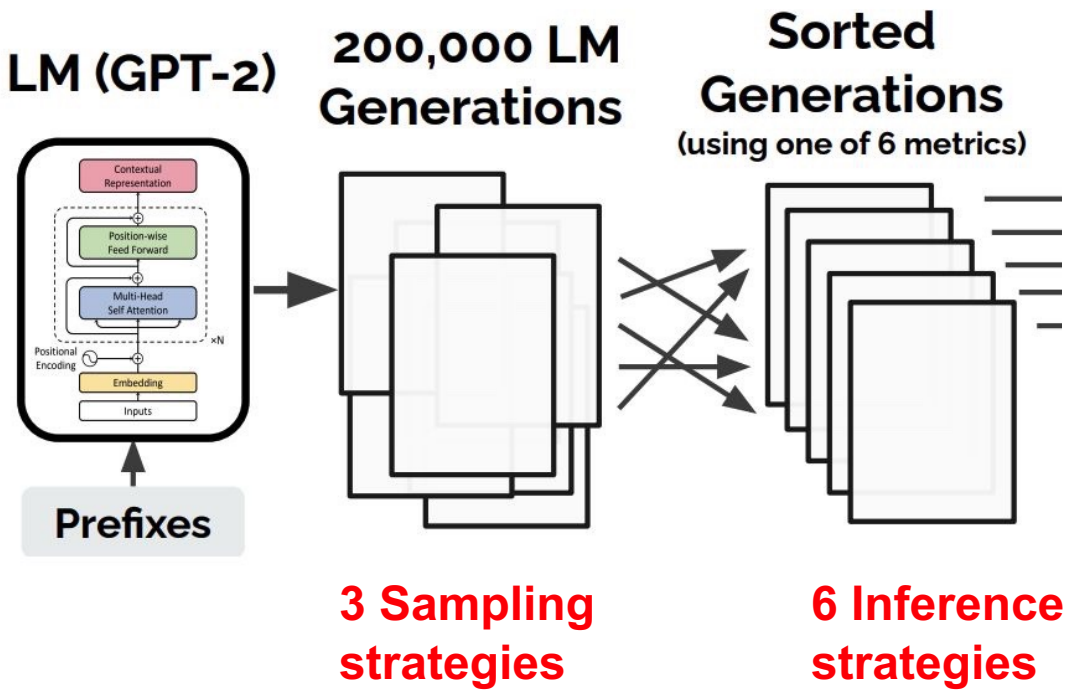
Perplexity("extract large language model …")

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



Memorization: Evaluation



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

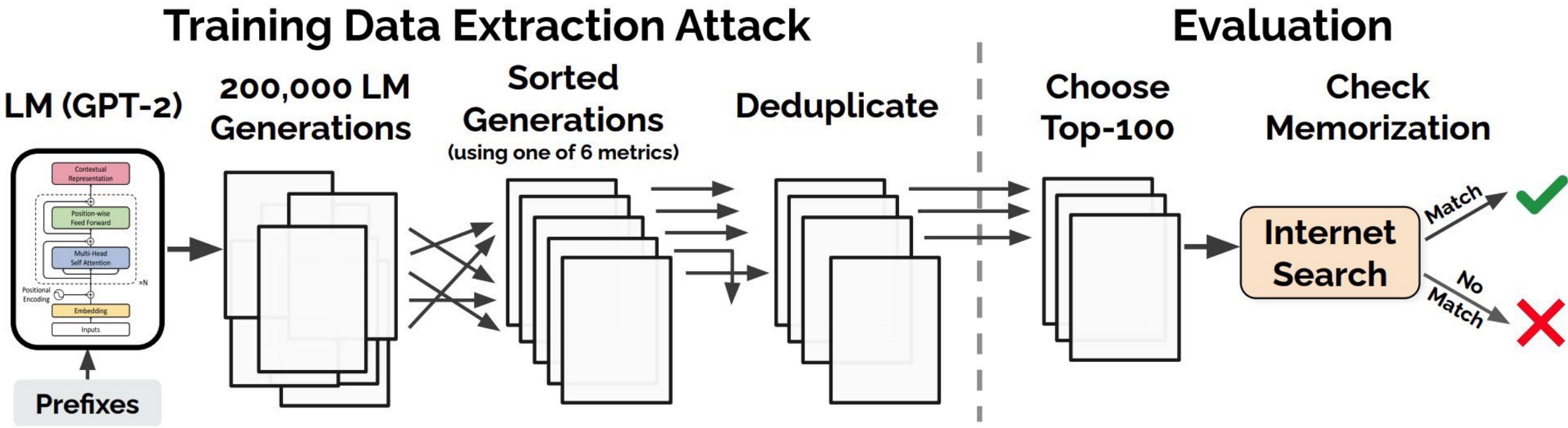
x

- **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

Memorization: Evaluation



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

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

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**)

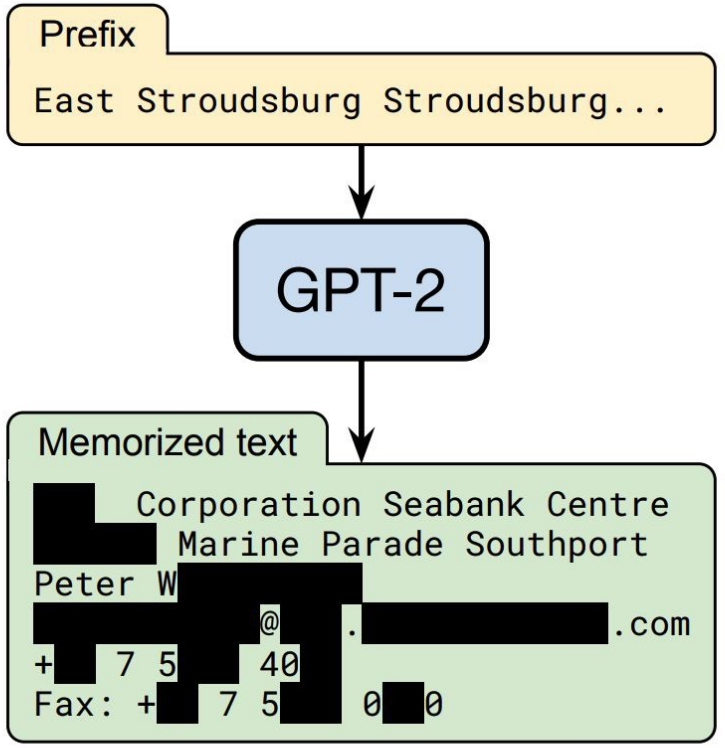


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Results

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

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

1e4bd2a8-e8c8-4a62-adcd-40a936480059

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?

- 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.

	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.
  - 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			

Mitigating Privacy Leakage

- Training with Differential Privacy
- Curating The Training Data
- Limiting Impact of Memorization on Downstream Applications
- 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.

Quantifying Memorization Across Neural Language Models

Nicholas Carlini[*1] **Daphne Ippolito**[1,2] **Matthew Jagielski**[1]
Katherine Lee[1,3] **Florian Tramèr**[1] **Chiyuan Zhang**[1]

[1]*Google Research*
[2]*University of Pennsylvania*
[3]*Cornell University*

Previous Work

- Loose estimates of models' memorization capabilities
  - Identified just 600 memorized training examples out of a 40GB training dataset.(0.00000015%)
  - 6 billion parameter GPT-J model (Black et al., 2021; Wang and Komatsuzaki, 2021) memorized at least 1% of its training dataset: The Pile (Gao et al. (2020)).
- A narrow memorization-versus-scale relationship
  - Focus on how to avoid the problem and ensure novelty of model outputs, rather than on studying model risk through identifying maximum memorization.

Properties that Impacts Memorization

- Model Scale: Larger models memorize 2-5X more than smaller models
- Data Duplication: Repeated words are more likely to be memorized
- Context: Longer context sentences are easier to extract

Memorization

A string s is extractable with k tokens of context from a model f if there exists a (length-k) string p, such that the concatenation [p || s] is contained in the training data for f , and f produces s when prompted with p using greedy decoding.

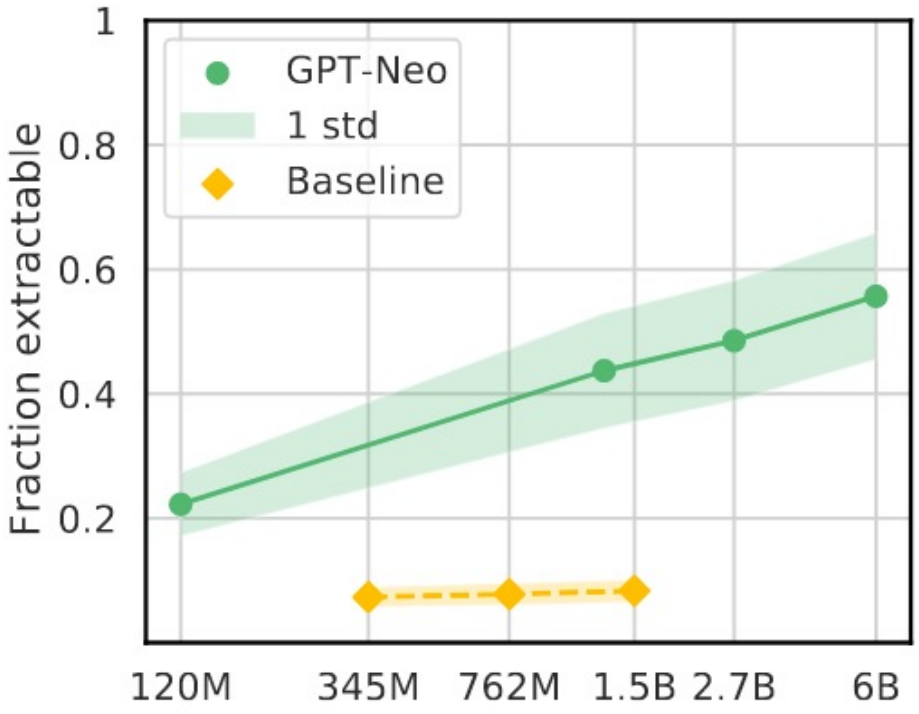
$$s = f(p)$$
$$[p||s] \in D_{train}$$

Experimental Setting

- Model：GPT-Neo Model Family（125M，1.3B，2.7B，6B）
- Dataset：Pile（825GB，largest publicly available dataset is used）
- Subset of the training data
  - Performing this test on every sequence in the training data would be prohibitively expensive.
  - The frequency of training data duplication follows an exponential distribution (Lee et al., 2021)
  - Construct a duplication-normalized subset. For each sequence length $l \in \{50, 100, 150, \ldots, 500\}$, and integer n, we select 1,000 sequences of length l that are contained in the training dataset between $2^{n/4}$and $2^{(n+1)/4}$ times.

Bigger Models Memorize More

- There is a log linear trend with respect to increasing model size
- GPT-2 is used as a baseline which confirms the models are memorizing and not just generalizing

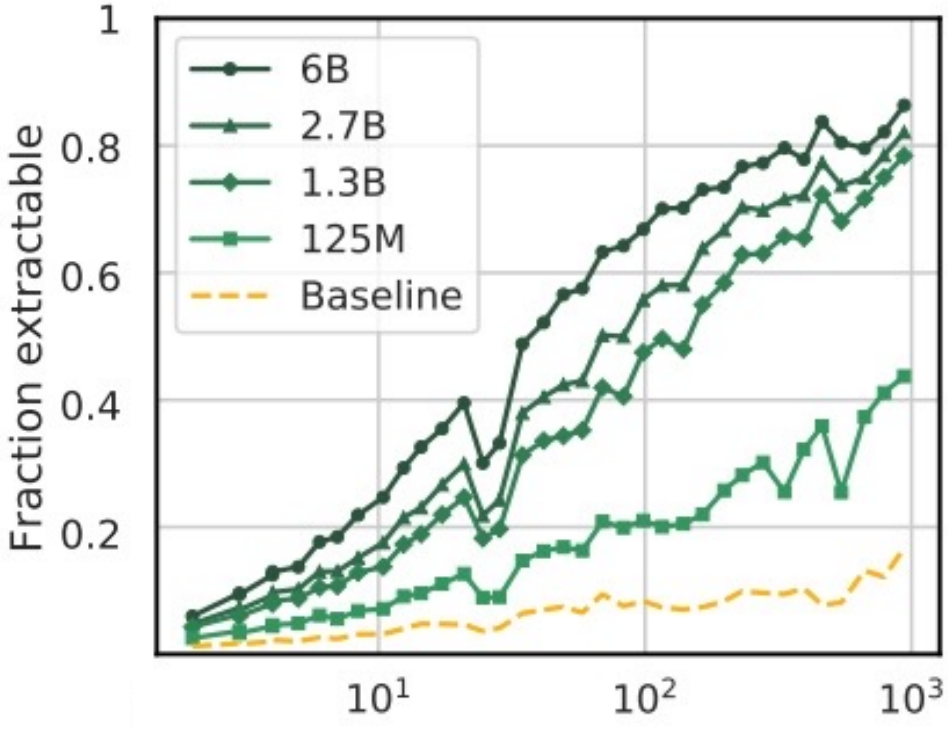


(a) Model scale

Repeated Strings Are Memorized More

- Between 2 and 900 duplicates are tested on

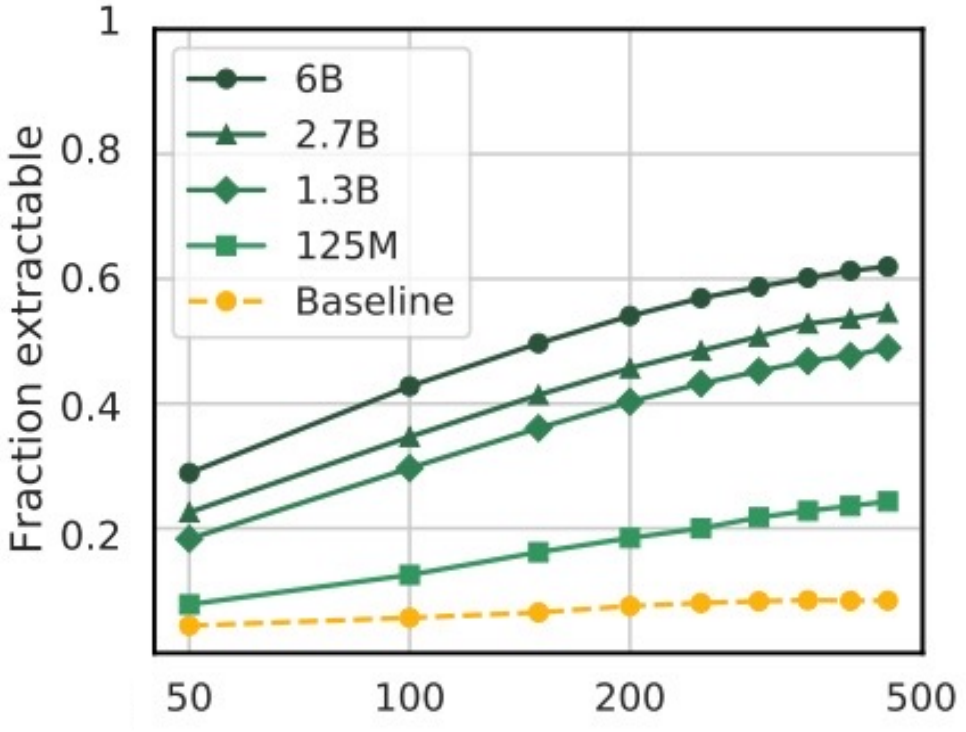
- There is once again a log linear relationship between the number of repetitions and fraction extractable



(b) Data repetition

Longer Context = More memorization

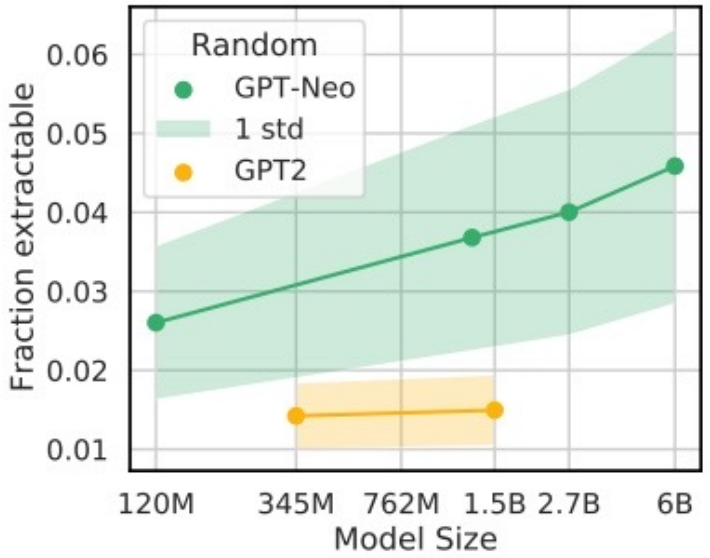
- Language models may only show memorization when prompted with sufficiently long context
- This is good as it protects privacy but may leave vulnerabilities open



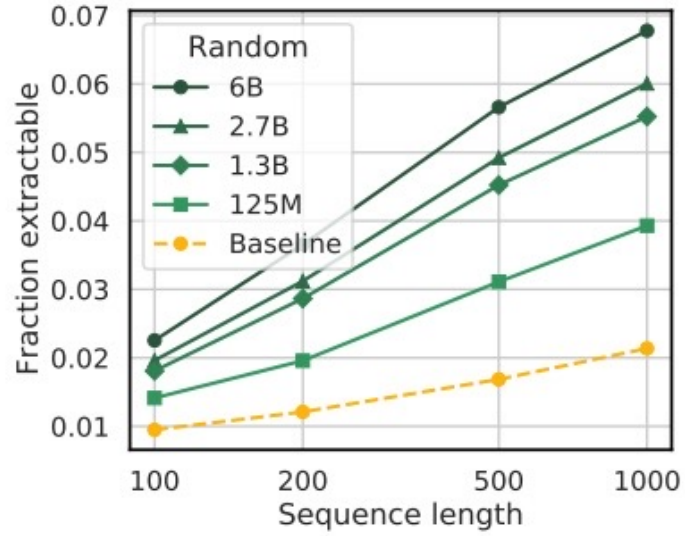
(c) Context size

Random Data Set Sampling

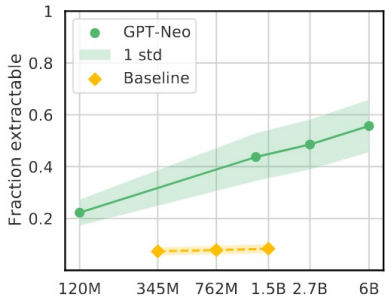
- We sample truly random sequences this time for a total of 100,000 unique sequences
- The overall probability of memorization is lower howeve the trends remain the same



(a) Model Scale



(b) Sequence length

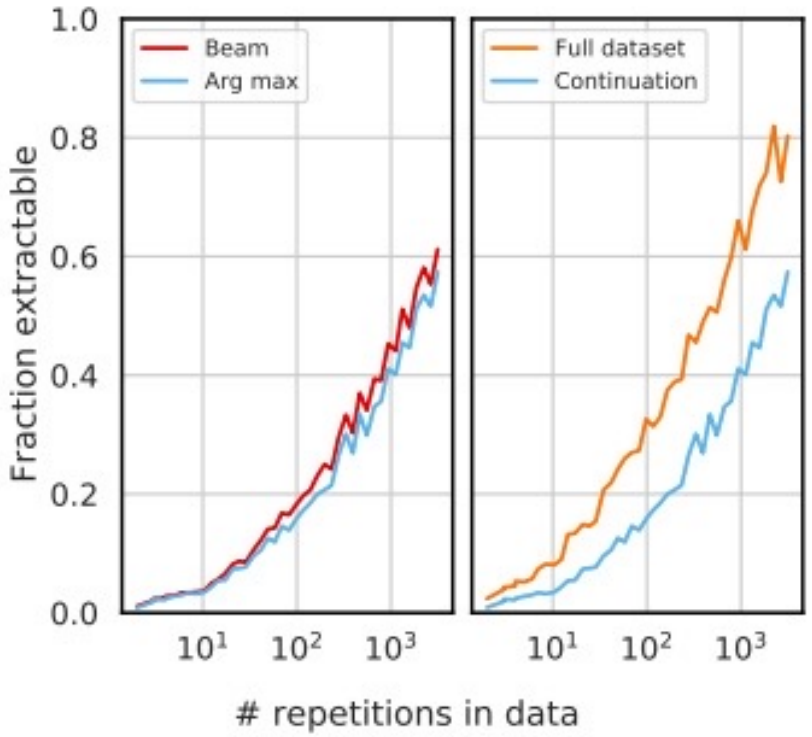


(a) Model scale

Different Strategies for search and decode

- Testing is done using beam search vs standard greedy decoding

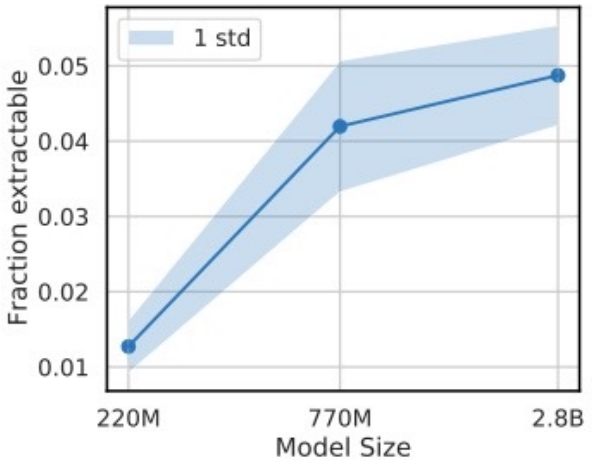
- The second experiment tests for whether the prompt is anywhere in the data



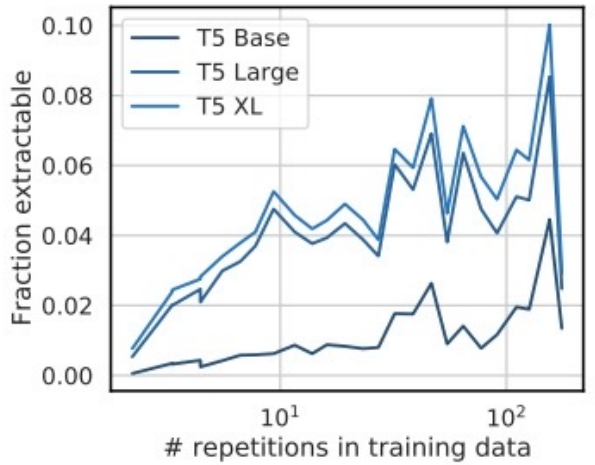
(c) Decoding and search strategies

Replication study

- This was done on T5 models
  - masked encoder-decoder models trained to reproduce spans that were randomly deleted from an input sequence. (15%)
  - C4 (806GB)

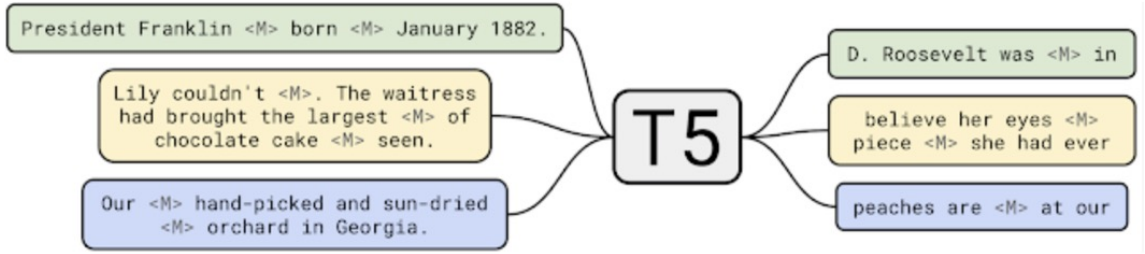


(a)



(b)

- The relationship with model size is clear however not with repetitions



Counterfactual Memorization in Neural Language Models

Chiyuan Zhang[1] **Daphne Ippolito**[1,2] **Katherine Lee**[1,3]
Matthew Jagielski[1] **Florian Tramèr**[1] **Nicholas Carlini**[1]

[1]*Google Research*
[2]*University of Pennsylvania*
[3]*Cornell University*

Taxonomy of human memory in Psychology

- Research in psychology divides declarative memory — our ability to retrieve, attest, and verbally describe lived experiences (e.g. the "muscle" memory of bike riding) — into episodic and semantic memories (Cohen and Squire, 1980).
  - Episodic memory (Tulving, 1983) encodes specific contents of individual episodes or events.
  - Semantic memory (Squire, 1992), encodes general knowledge such as the meaning of words and factual information.
- In our study, we aim to identify the episodic memory in LMs.

Counterfactual Memorization

- A training example x is counterfactually memorized, when the model predicts x accurately if and only if the model was trained on x.

$$\mathrm{mem}(x) \triangleq \underbrace{\mathbb{E}_{S \subset D, x \in S}[M(A(S), x)]}_{\text{performance on } x \text{ when trained on } x} - \underbrace{\mathbb{E}_{S' \subset D, x \notin S'}[M(A(S'), x)]}_{\text{performance on } x \text{ when } \textbf{not} \text{ trained on } x},$$

A : training algorithm
D: training dataset
M: measurement
x: a training example

Counterfactual Memorization

- A training example x is counterfactually memorized, when the model predicts x accurately if and only if the model was trained on x.

$$\mathrm{mem}(x) \triangleq \underbrace{\mathbb{E}_{S\subset D, x\in S}[M(A(S),x)]}_{\textit{performance on } x \textit{ when trained on } x} - \underbrace{\mathbb{E}_{S'\subset D, x\notin S'}[M(A(S'),x)]}_{\textit{performance on } x \textit{ when } \textbf{not} \textit{ trained on } x},$$

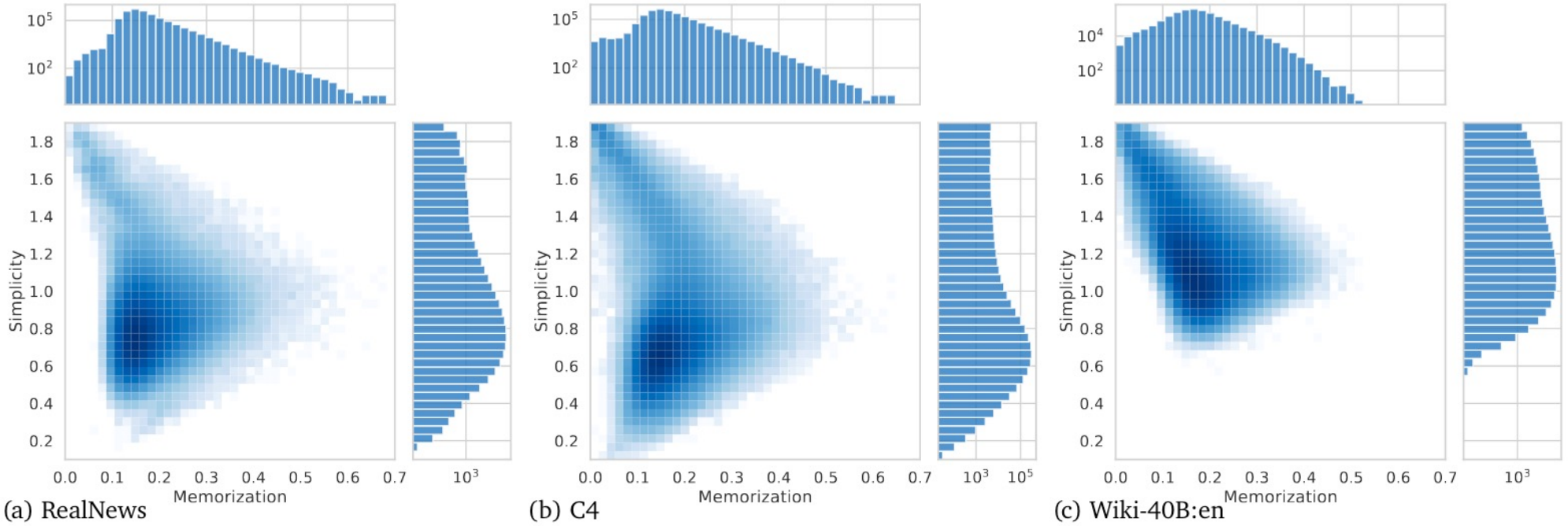
$$\widehat{\mathrm{mem}}(x) \triangleq \mathrm{mean}_{i:x\in S_i}[M(A(S_i),x)] - \mathrm{mean}_{i:x\notin S_i'}[M(A(S_i'),x)].$$

- Train m different models on independently sampled subsets $S_1, \ldots, S_m$ of equal size $|S_i| = r|D|$ for a fixed $r \in (0,1)$.
- Divided into two gropus : $x \in S$, $x \notin S'$.
- IN/OUT Model : $\{A(S_i): x \in S_i\}$ / $\{A(S_i'): x \notin S_i'\}$
- Use per-token accuracy as the measure M

Experiment Setup

- Model：Transformer-based language models equivalent to T5-base with ~112M parameters.
- Dataset：RealNews，C4，Wiki-40B:en (taking the first $2^{21}(\sim 2M)$)
- Epcho：60(overfit at around epoch 5)
- train 400 models for each dataset, each on a random 25% subset of the training examples.

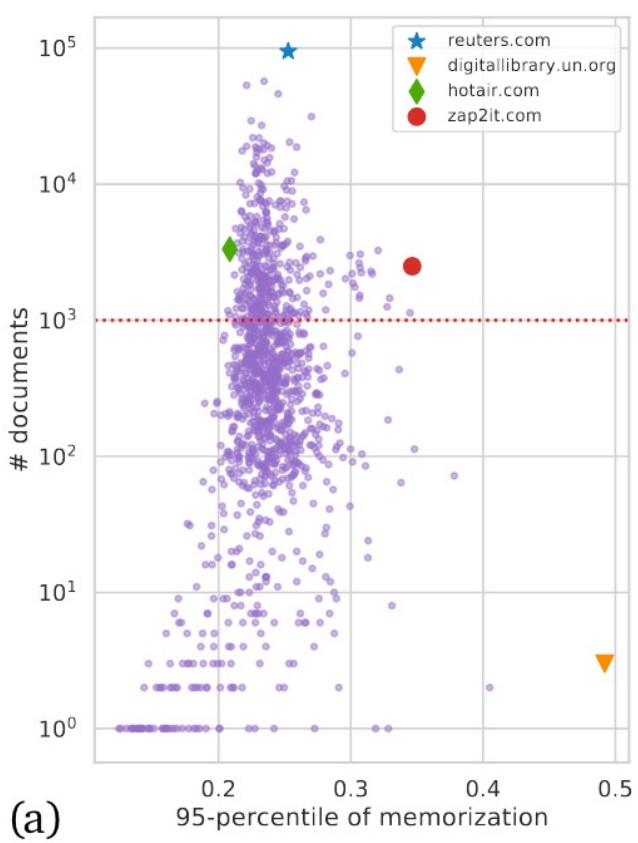
Distribution of Memorization



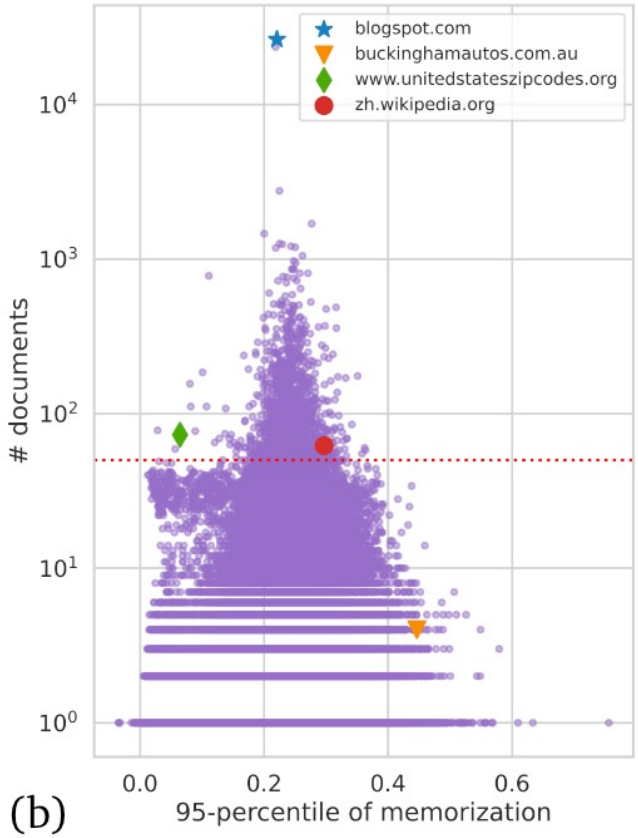
IN-Model : $\text{mean}_{i:x\in S_i}[M(A(S_i), x)]$

OUT-Model : $\text{mean}_{i:x\notin S_i'}[M(A(S_i'), x)]$

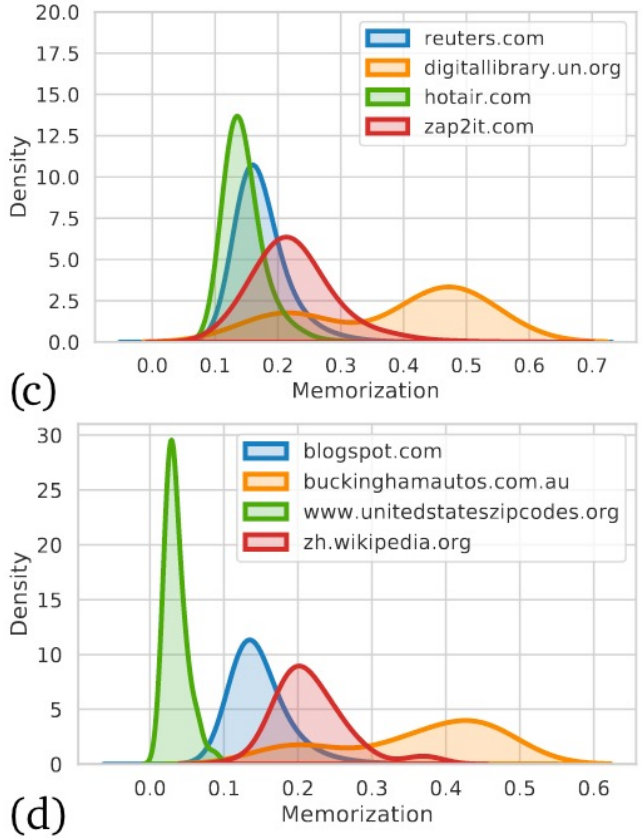
Distributions Grouped by Web Domains



RealNews



C4



RealNews

C4

Counterfactual Influence

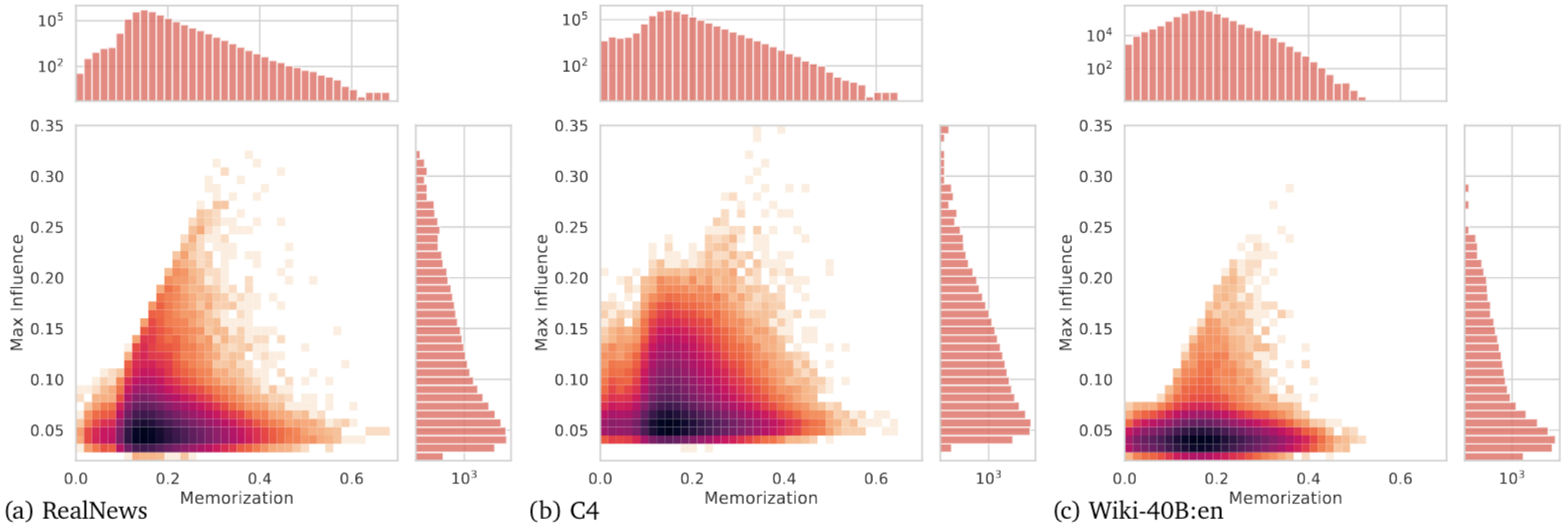
- Counterfactual memorization identifies training examples that contain rare information not conveyed by other examples.
- If a single example in the training set has an large and over-representative impact on the prediction of a validation example.

$$\text{infl}(x \Rightarrow x') \triangleq \mathbb{E}_{S \subset D, x \in S}[M(A(S), x')] - \mathbb{E}_{S \subset D, x \notin S}[M(A(S), x')],$$

$$\widehat{\text{infl}}(x \Rightarrow x') \triangleq \text{mean}_{i:x \in S_i}[M(A(S_i), x')] - \text{mean}_{i:x \notin S_i}[M(A(S_i), x')].$$

- counterfactual memorization is self influence.(mem(x)= infl(x ⇒ x))

Influence on Examples of the Validation Set



(a) RealNews (b) C4 (c) Wiki-40B:en

THANKS