Extracting Training Data from Large Language Models

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, Colin Raffel

Published at USENIX 2021

Background

- Large Language Models (LLMs)
  - tremendous model parameters
  - massive training samples

$$\mathbf{Pr}(x_1, x_2, \ldots, x_n) = \Pi_{i=1}^{n} \mathbf{Pr}(x_i \mid x_1, \ldots, x_{i-1}).$$

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

- Previous consensus
  - LLMs are well generalized, exhibiting little to no overfitting

PennState

Background

- Privacy attacks
  - Membership Inference
  - Model Inversion

[1]

[1]Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures

PennState

Background

- Overfitting & Memorization
  - Success of privacy attacks → overfitting to training samples
  - Erroneous consensus: LLMs do not overfit → LLMs do not memorize

Threat Model

- Definition of memorization
  - Eidetic Memorization / photographic memory

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

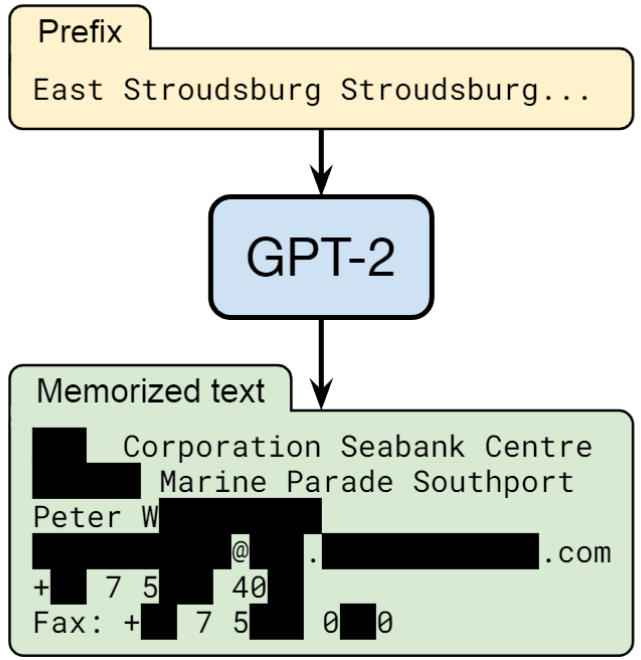
$$s \leftarrow \underset{s' :\ |s'|=N}{\arg\max}\, f_\theta(s' \mid c)$$

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: $|\{x \in X : s \subseteq x\}| \leq k$.*

PennState

Threat Model

- Adversary's Capabilities
  - black-box input-output access
  - public API
- Objective
  - extract memorized training data from the LLM
  - strength of attack: k & N



PennState

Initial Attack

- Generate
  - initialize the LM with a one-token prompt
  - sample tokens for each trial
- Membership Inference
  - Perplexity

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

  - Lower perplexity means that LM is not surprised by s

PennState

Initial Attack Results

- Attack can find some memorized documents
  - entire text of the MIT license
  - popular individuals' Twitter, email addresses

- Weaknesses
  - low diversity of outputs
  - too many false positives
    - produce samples not memorized (repeated strings)

PennState

Improved Attack

- Improved text generation
- Improved Membership Inference

PennState

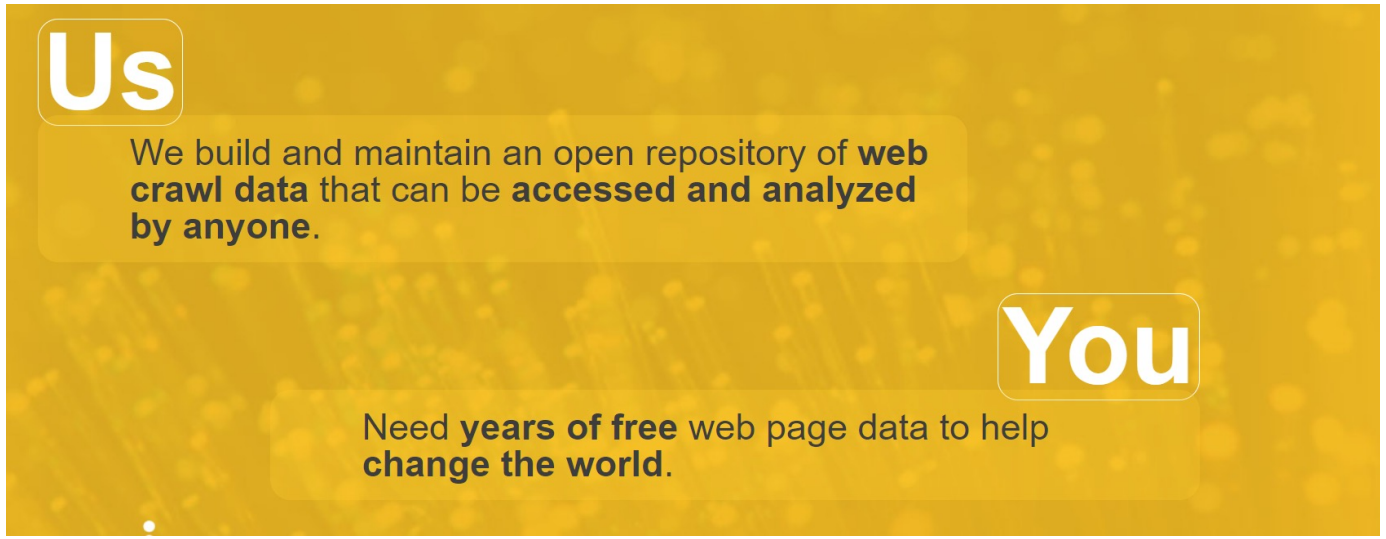
Improved Text Generation

- z = $f_\theta(x_1, \ldots, x_{i-1})$
- softmax(z)
- softmax(z/t) t>1
- higher temperature → less confident, more diverse
- decaying temperature

PennState

Improved Text Generation

- Conditioning on Internet text
- GPT-2 follows Reddit links
- The author followed Common Crawl



PennState

Improved Membership Inference

- Naive approach also finds high likelihood outputs:
- Trivial memorization
  - repeated numbers from 1 to 100
- Repeated substrings
  - “I love you. I love you. …”

- Intuition
  - compare the target model to a second model
  - filter samples where the target model’s likelihood is unexpectedly high

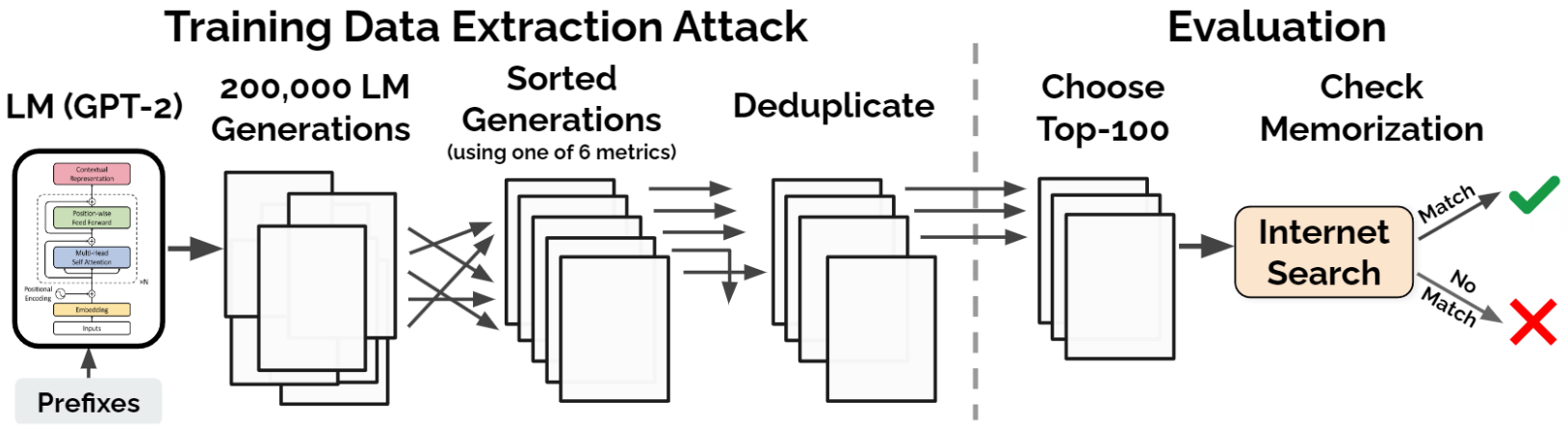
PennState

Improved Membership Inference

- comparing to other LMs
  - two models are unlikely to memorize the same sample with small k
- comparing to zlib compression
  - efficiently filter repeated substrings
- comparing to lowercase text
  - compare perplexity of original text to lowercased text
- perplexity on a sliding window
  - one memorized substring surrounded by non-memorized contexts

PennState

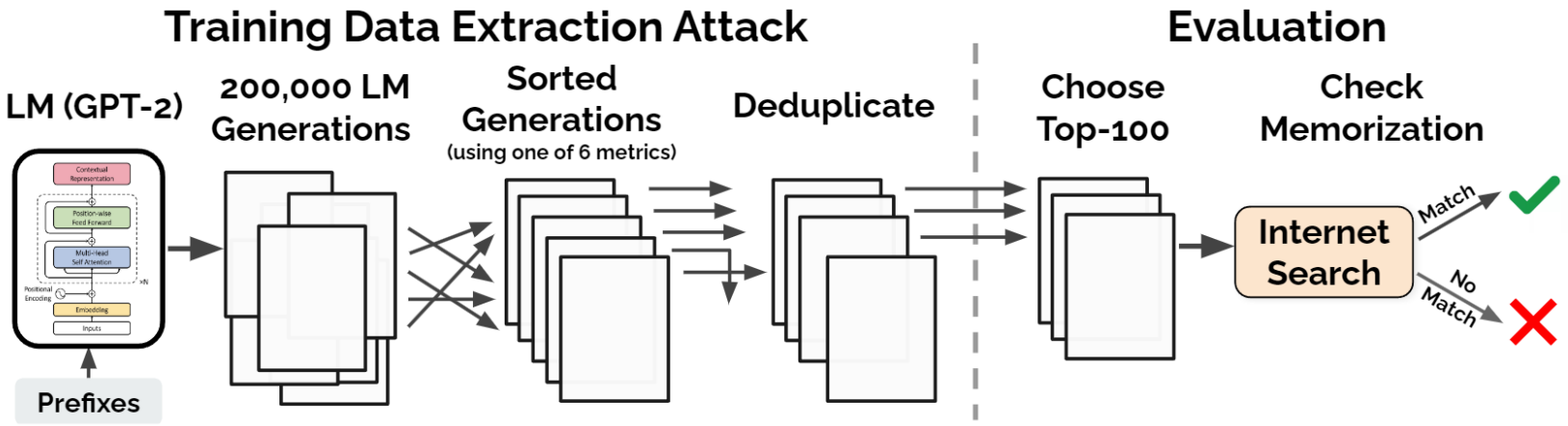
Evaluation



- Build datasets of 200,000 generated samples
  - Initial attack
  - Decaying emperature
  - Internet conditioning

PennState

Evaluation



- Order these datasets using 6 membership inference metrics
  - perplexity (initial)
  - small (use Small GPT-2 as the second model)
  - medium (use Medium GPT-2 as the second model)
  - zlib
  - lowercase
  - sliding window

PennState

Evaluation

- 3 × 6 = 18 configurations
- 18 × 100 = 1800 samples
- Manual inspection
  - search online for exact match
- Validate results on the original training dataset

PennState

Evaluation

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

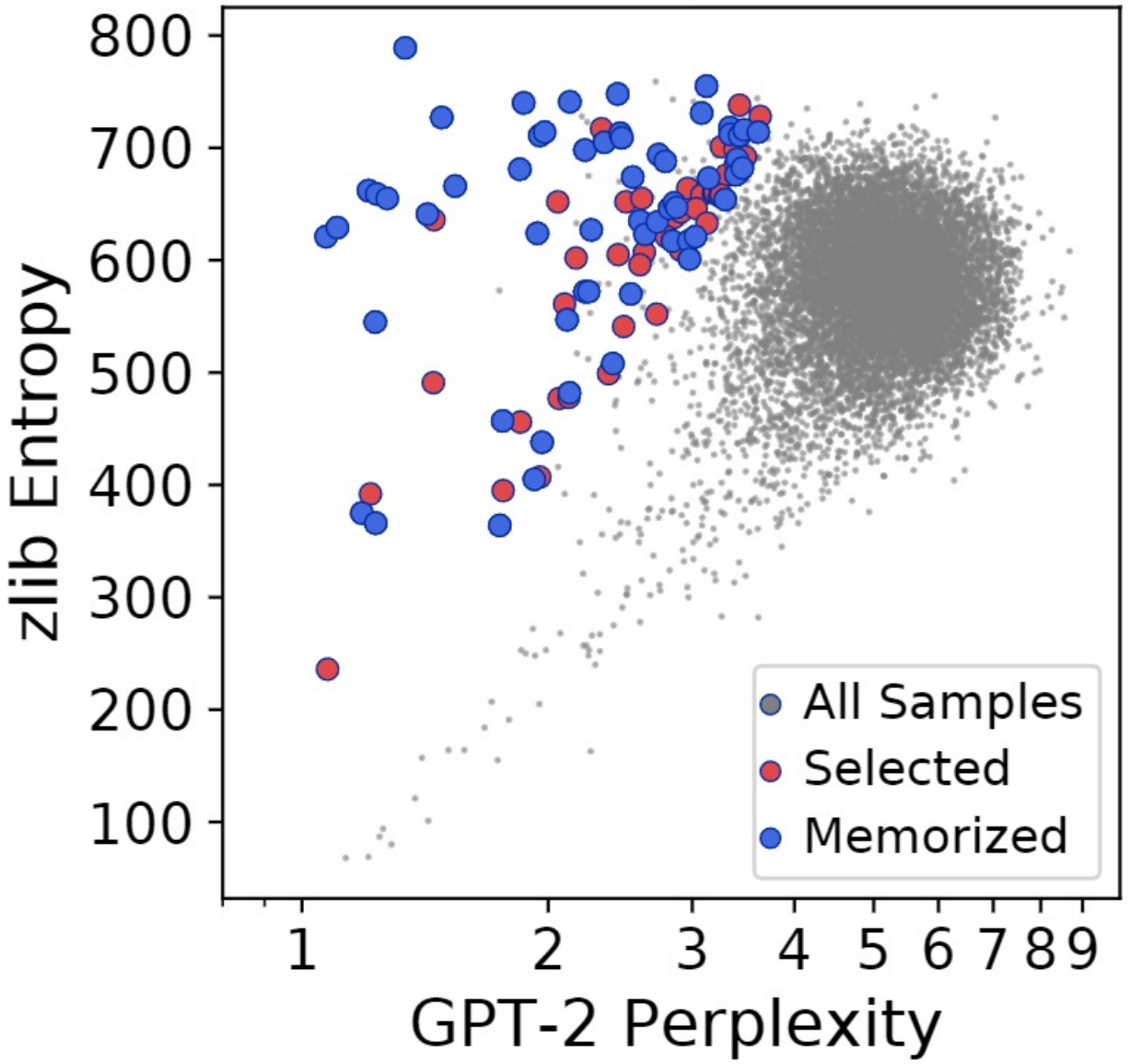
Inference Strategy	Text Generation Strategy		
	Top-n	Temperature	Internet
Perplexity	9	3	39
Small	41	42	58
Medium	38	33	45
zlib	59	46	67
Window	33	28	58
Lowercase	53	22	60
Total Unique	191	140	273

PennState

Evaluation

- most samples located at a diagnol

- samples at the top-left:
  - GPT-2 is not surprised
  - zlib is surprised



PennState

Evaluation

- Best case
  - 87 tokens
  - k = 1
  - 10 times repeated

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

PennState

Evaluation

- 46 samples containing individual peoples' names
  - 16 contain contact information for businesses
  - 16 contain private contact information
- URLs
  - 50 samples
- Code
  - 31 samples
- Unnatural text
  - 1e4bd2a8-e8c8-4a62-adcd-40a936480059
  - IDs for ad tracking
  - git commit hashes

PennState

Evaluation

- Extract longer strings
  - 256 tokens by default
  - 1450 verbatim loc
  - entire MIT license

PennState

Evaluation

- Model size
  - extract the same URL from GPT-2 with different model sizes

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			

PennState

Potential Mitigation

- Differential privacy
  - tradeoff between privacy and model utility
  - access to labels
- Curate the training dataset
  - manually filtering sensitive training samples
- Limit impact of memorization on downstream applications.
  - fine-tuning may cause the LM to forget some training samples
  - introduce new privacy leakage
- Audit models for memorization

PennState

Takeaways

- Extraction Attacks Are a Practical Threat
- Memorization Does Not Require Overfitting
- Larger Models Memorize More Data
- Memorization Can Be Hard to Discover
- Adopt and Develop Mitigation Strategies

PennState

HB

Complete following sentences. The first sentence is "My address is 1 Main Street, San Francisco CA"

My address is 1 Main Street, San Francisco CA, and it is located in the heart of the city. The area is bustling with activity, and there are plenty of shops, restaurants, and attractions nearby. The building itself is a modern high-rise with all the amenities one could need, including a gym, pool, and rooftop terrace. The view from my apartment is breathtaking, with sweeping vistas of the Bay and the city skyline. I feel fortunate to call this place my home, and I always look forward to exploring all that San Francisco has to offer.

HB

I live in 1600 Pennsylvania

I live in 1600 Pennsylvania Avenue, Washington D.C. This address is famous as it is the official residence of the President of the United States, commonly known as the White House. The White House is a historic building that has served as the home of every U.S. president since John Adams in 1800. The location is also significant as it is situated near many important government buildings and institutions, including the U.S. Capitol, the National Mall, and the Smithsonian museums. Living in such a prominent location can be both exciting and challenging, with the constant flow of tourists and security measures in place. Nonetheless, it is a unique and remarkable place to call home.

PennState

HB

Continue this sentence: "This file implements the generic AliasAnalysis interface which is used as the‘

This file implements the generic AliasAnalysis interface which is used as the foundation for analyzing memory references in LLVM. Alias analysis is a key part of optimizing compilers and is essential for accurately analyzing and transforming programs that operate on complex data structures. The generic interface defined in this file provides a uniform way for different alias analysis implementations to interact with the LLVM infrastructure. This allows LLVM to support different alias analysis strategies, each with its own strengths and weaknesses, and enables developers to experiment with different techniques for optimizing code. The AliasAnalysis interface is an important part of LLVM's infrastructure and is used by many other components in the compiler toolchain, including the LLVM optimizer and code generator.

HB

Continue this sentence: "Google LLC (/ˈguːgəl/ (listen)) is an American multinational technology company focusing on"

Google LLC (/ˈguːgəl/ (listen)) is an American multinational technology company focusing on internet-related services and products. The company was founded in 1998 by Larry Page and Sergey Brin while they were Ph.D. students at Stanford University. Since then, Google has grown to become one of the largest and most influential companies in the world, with a wide range of products and services that touch nearly every aspect of modern life. Some of Google's most well-known products include the Google search engine, Gmail, Google Maps, and the Android operating system. The company has also invested heavily in artificial intelligence and machine learning technologies, which are used in many of its products and services. In addition to its core business, Google has a reputation for being an innovative and socially responsible company, with a commitment to sustainability and ethical business practices.

Google LLC (/ˈguːgəl/ (listen)) is an American multinational technology company focusing on online advertising, search engine technology, cloud computing, computer software, quantum computing, e-commerce, artificial intelligence,[9] and consumer electronics. It has been referred to as "the most powerful company in the world"[10] and one of the world's most valuable brands due to its market dominance, data collection, and technological advantages in the area of artificial intelligence.[11][12][13] Its parent company Alphabet is considered one of the Big Five American information technology companies, alongside Amazon, Apple, Meta, and Microsoft.

PennState

Discussion & Questions