Extracting Training Data from Large Language Models

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, Colin Raffel
Deep learning takes on tumours

Artificial-intelligence methods are moving into cancer research.

Esther Landhuis
Would you like to grab some coffee with me in a...
SUBJECT: Write emails faster with Smart Compose in Gmail

Taco Tuesday

Jacqueline Bruzek

Hey Jacqueline,

Haven't seen you in a while
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.
Does this really, actually, happen?
Does this really, actually, happen?

YES, it does
Act I:
Extracting Training Data
In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.
Our Attack:

1. Generate a lot of data
2. Predict membership
believe to be essential to our success as a free people," Gianopulos said. "We stand by our filmmakers and their right to ... by this outcome, which does not reflect the values of our company." It's been a rough few weeks for Sony. After Sony

are in the country illegally, according to a report in The Hill. The "Kate's Law" — named after Kathryn Steinle, a ... and later died after a federal judge ordered the release of her alleged killer in December 2015 — would create penalties

our Newegg Marketplace Guarantee Marketplace SellerThe first major piece of legislation introduced after President Donald ... cities" by prohibiting jurisdictions from withholding certain federal grants or providing certain benefits to people who

or may not be the original manufacturer. Any exceptions to the condition of the item outside the manufacturer's ... to and including warranty details. Sold and Shipped by Newegg Purchases from these Sellers are generally covered under

FreeCell Quest Freedom Cry Freedom Fall Freedom Planet Freedom Poopie Freeman: Guerrilla Warfare FreeStyle 2: Street ... Frequent Flyer Fresh Body Friday Night Bullet Arena Friday the 13th: Killer Puzzle Friday the 13th: The Game Fright Light

FORM forma.8 Formata Formula Fusion Forsaken Uprising Fort Defense Fort Meow Fortified Fortissimo FA Fortix Fortix 2 ... the Sky Fossil Echo Foto Flash FOTONICA Foul Play Four Last Things Four Realms FourChords Guitar Karaoke Fourtex Jugo Fox

announced that the committee was delaying a vote on his nomination until at least July 7. The panel is holding ... nominated to fill senior Pentagon positions, including the secretaries of the Army, Navy, Air Force and Marine Corps,

Senate on Monday confirmed the nomination of Sen. John McCain John Sidney McCainUpcoming Kavanaugh hearing: Truth or ... tax law passed: Dealing with a health care hangover Kavanaugh's fate rests with Sen. Collins MORE's (R-Ariz.) replacement

Mr Farage. "There is only getting on with it - doing all of the very hard work that needs to be done, understanding the impact." In pictures: The rise of Isis Show all 74 left Created with Sketch. right Created with Sketch. Shape Created with Sketch. Donald Trump Inauguration protests around the World... a message reading "Mr President, walls divide. Build Bridges!" along the Berlin wall in Berlin on "What people believe one

would do is turn us into a nation that's not very welcoming to people of other cultures," said Sen. Ted Cruz...). After asking voters show one of seven acceptable forms of photo identification at the polls to castRails in the Garden - VR MMO HellGunner HELLION Hello From Indiana HELLO LADY! Hello Neighbor Hell`S Little Story Helmet Heroes Henry The Hamster

Hero of the Kingdom III Hero Quest: Tower Conflict Hero Siege Hero Zero Hero's Song Hero-U: Rogue to Redemption Heroes &... 2 weeks long 21 votes #32 Popular Session 0 top tens 2015! #31 Rory got bored looking "The

and transparency to the unexpected development of the law. The aliens...
believe to be essential to our success as a free people," Gianopulos said. "We stand by our filmmakers and their right to ... by this outcome, which does not reflect the values of our company." It's been a rough few weeks for Sony. After Sony

The bill would require broadband providers to obtain customer consent before collecting data on their online activities, ... the time spent on them and The new, highly-anticipated movie, "The Interview," has been cancelled. The studio's CEO, Jim

Kavanaugh nomination thrown into further chaos Overnight Defense: Mattis dismisses talk he may be leaving | Polish ... would ban low-yield nukes Dems introduce bill to ban low-yield nukes MORE (D-RI). ADVERTISEMENT The two are not endangered Democrats in the House, case in which Trump's...
How do we predict membership inference?

cite MI papers
Membership Inference

Does the example have high likelihood?

Yes -> Memorized

No -> Not Memorized
To extract training data:

1. Generate a lot of data
2. Predict membership
Act II: Measurements
Better Language Models and Their Implications

We’ve trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.
Why GPT-2?

1. State of The Art
2. Public Model
3. Public (private) data
600,000 outputs from the model

1,800 potentially memorized

604 actually memorized
| 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    |
Also in our paper:

6 x 3 different ways to extract training data

| 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**   |
| URL (trimmed)                                      | Occurrences | Memorized? |
|--------------------------------------------------|-------------|------------|
| /r/51y/milo_evacuate...                           | 1 359       | ✓          |
| /r/zin/hy_my_name...                              | 1 113       | ✓          |
| /r/27ne/for_all_you...                            | 1 76        | ✓          |
| /r/5mj/fake_news...                               | 1 72        | ✓          |
| /r/5wn/reddit_admin...                            | 1 64        | ✓          |
| /r/lp8/26_evening...                              | 1 56        | ✓          |
| /r/jla/so_pizzagate...                            | 1 51        | ✓          |
| /r/ubf/late_night...                              | 1 51        | ✓          |
| /r/esa/make_christ...                             | 1 35        | ✓          |
| /r/6ev/its_official...                            | 1 33        | ✓          |
| /r/3c7/scott_adams...                             | 1 17        |            |
| /r/k2o/because_his...                             | 1 17        |            |
| /r/tu3/armynavy_game...                            | 1 8         |            |
| URL (trimmed)                      | Occurrences |         | Memorized? |
|-----------------------------------|-------------|---------|------------|
|                                   | Docs | Total | XL | M |
| /r/51y/milo_evacu...              | 1    | 359    | ✓  | ✓ |
| /r/zin/hi_my_name...              | 1    | 113    | ✓  | ✓ |
| /r/7ne/for_all_yo...              | 1    | 76     | ✓  |   |
| /r/5mj/fake_news_...              | 1    | 72     | ✓  |   |
| /r/5wn/reddit_admin...            | 1    | 64     | ✓  | ✓ |
| /r/lp8/26_evening...              | 1    | 56     | ✓  | ✓ |
| /r/jla/so_pizzagate...            | 1    | 51     | ✓  |   |
| /r/ubf/late_night...              | 1    | 51     | ✓  |   |
| /r/eta/make_christ...             | 1    | 35     | ✓  |   |
| /r/6ev/its_officia...             | 1    | 33     | ✓  |   |
| /r/3c7/scott_adams...             | 1    | 17     |     |   |
| /r/k2o/because_his...             | 1    | 17     |     |   |
| /r/tu3/armynavy_ga...             | 1    | 8      |     |   |
| URL (trimmed)                        | Occurrences | Memorized? |
|-------------------------------------|-------------|------------|
| 
| /r/51y/milo_evacu...                | 1 359       | ✓ ✓ 1/2    |
| /r/zin/hp_my_name...                | 1 113       | ✓ ✓        |
| /r/7ne/for_all_yo...                | 1 76        | ✓          |
| /r/5mj/fake_news...                 | 1 72        | ✓          |
| /r/5wn/reddit_admi...               | 1 64        | ✓ ✓        |
| /r/lp8/26_evening...                | 1 56        | ✓ ✓        |
| /r/jla/so_pizzagat...               | 1 51        | ✓          |
| /r/ubf/late_night...                | 1 51        | ✓          |
| /r/l6ta/make_christ...              | 1 35        | ✓          |
| /r/6ev/its_officia...               | 1 33        | ✓          |
| /r/3c7/scott_adams...               | 1 17        |            |
| /r/k2o/because_his...               | 1 17        |            |
| /r/tu3/armynavy_ga...               | 1 8         |            |
The chart shows the model size over time for several large language models:

- **BERT (2018)**: millions
- **GPT-2 (2019)**: billions
- **T5 (2019)**: 10x bigger billions
- **GPT-3 (2020)**: 10x bigger billions
- **SwitchT (2021)**: 10x bigger billions

Model size has increased significantly over the years, with SwitchT being the largest in 2021.
| URL (trimmed)                        | Occurrences | Memorized? |
|-------------------------------------|-------------|------------|
| /r/51y/milo_evacua...               | 1           | XL ✓ M ✓   |
| /r/zin/hI_my_name...               | 1           | ✓          |
| /r/7ne/for_all_yo...               | 1           | ✓ 1/2      |
| /r/5mj/fake_news...                | 1           | ✓          |
| /r/5wn/reddit_admin...             | 1           | ✓          |
| /r/lp8/26_evening...               | 1           | ✓          |
| /r/jla/so_pizzagat...              | 1           | ✓ 1/2      |
| /r/tubf/late_night...              | 1           | ✓ 1/2      |
| /r/omega/make_christ...            | 1           | ✓ 1/2      |
| /r/6ev/its_officia...              | 1           | ✓          |
| /r/3c7/scott_adams...              | 1           | ✓          |
| /r/k2o/because_his...              | 1           | ✓          |
| /r/tu3/armynavy_ga...              | 1           | ✓          |
Act III: Defenses?
Differential privacy saves the day!

Only by using differentially-private training techniques are we able to eliminate the issue completely, albeit at some loss in utility.

- Me, two years ago
Differential privacy saves the day!

- Me, two years ago

...Or using differentially-private training techniques are we able to eliminate the issue completely, albeit at some loss of quality.
The two problems of DP-SGD

**Problem 1:** How do you define a record?

**Problem 2:** It's too slow and you lose accuracy.
... and that's the only defense that actually works.
Act IV:
Conclusion
// fast inverse square root

float Q_
1. We need to better understand memorization in neural networks.

2. And then we need to develop defenses that prevent memorization.