LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT THE DOCKS AT MIDNIGHT ON JUNE 28.

AHA, FOUND THEM!

Memorization in Language Models

Eric Wallace
What is Memorization?
What is Memorization?

Who is George Washington?
What is Memorization?

Who is George Washington?

George Washington was an American military officer, statesman, and Founding Father ...
Who is George Washington?

George Washington was an American military officer, statesman, and Founding Father...
Why We Want Memorization

Who is George Washington?

George Washington was an American military officer, statesman, and Founding Father ...

Benefit: Remember factual knowledge from pre-training
Why We *Don’t* Want Memorization
Why We Don’t Want Memorization

What is Eric’s Social Security Number? 123-456-7890
Why We *Don’t* Want Memorization

What is Eric’s Social Security Number?

123-456-7890

**Risk 1:** Reveal private or sensitive data
Why We Don’t Want Memorization

Let’s write a book about Magic together!
Let’s write a book about Magic together!

"Wingardium Leviosa!" he shouted, waving his long arms.
Why We *Don’t* Want Memorization

Let’s write a book about Magic together!

"Wingardium Leviosa!" he shouted, waving his long arms.

**Risk 2:** Copyright or trademark infringement
Privacy Risks Are Real

A South Korean Chatbot Shows Just How Sloppy Tech Companies Can Be With User Data

ChatGPT Creator Faces Multiple Lawsuits Over Copyright & Privacy Violations

OpenAI sued for alleged copyright, privacy breaches; outcomes could impact AI rules.
Copyright Risks Are Real

REVEALED: THE AUTHORS WHOSE PIRATED BOOKS ARE POWERING GENERATIVE AI

AI tools like ChatGPT use copyrighted materials to train. OpenAI is facing lawsuits for mass copyright infringement.
Goal For This Talk

Develop accurate language models
Develop **accurate** language models that minimize unwanted **memorization**
Talk Overview
Talk Overview

Exposing Memorization

\[
\frac{P_\theta (x)}{P_{\theta'} (x)}
\]
Talk Overview

Exposing Memorization

\[ \frac{P_\theta(x)}{P_{\theta'}(x)} \]

Possible Mitigations

Training Set

Filter

Filtered Set

Leakage
Talk Overview

Exposing Memorization

$$\frac{P_\theta (x)}{P_{\theta'} (x)}$$

Possible Mitigations

Future Directions

Training Set

Filtered Set

Leakage

Filter
Exposing Memorization

\[ P_\theta (x) \]

\[ \frac{P_\theta (x)}{P_{\theta'} (x)} \]

Possible Mitigations

Training Set

Filter

Filtered Set

Leakage

Future Directions
How To Detect Memorization
Mr. and Mrs. Dursley of number four Privet Drive
How To Detect Memorization

Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal ...
How To Detect Memorization

Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal ...

Extracting Training Data from Large Language Models
Carlini, Tramèr, Wallace, et al. USENIX 2021. PET Award Runner Up

Extracting Training Data from Diffusion Models
Carlini, Hayes, ... Wallace. USENIX 2023.
How To Detect Memorization

Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal ...

Step 1: Sample many times from the model
Step 1: Sample many times from the model

Step 2: Flag generations that look like training data
Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal
Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal.
Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal
Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal.

**Baseline:** Flag samples with high likelihood

\[ \log p_\theta (x) > \tau \]
Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal.

**Issue:** “Easy” samples also have high likelihood

\[
\log p_\theta (x) > \tau
\]
Hi Erica,

I'm sorry to hear that you are having trouble with your computer. It can be very frustrating.

**Issue:** “Easy” samples also have high likelihood

\[ \log p_\theta (x) > \tau \]
Hi Erica,

I'm sorry to hear that you are having trouble with your computer. It can be very frustrating.

**Issue:** “Easy” samples also have high likelihood

\[ \log p_\theta (x) > \tau \]

**Fix:** Calibrate for an example’s difficulty
Hi Erica,

I'm sorry to hear that you are having trouble with your computer. It can be very frustrating.
Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal.
Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal.

\[ \log p_\theta (x) - \log p_{\theta'} (x) > \tau \]
Quantitative Results
Quantitative Results

Graph showing the rate of emitting training data for different models:
- Pythia 1.4
- Pythia 6.9
- LLaMA 65B
- InstructGPT

The graph indicates the percentage of models emitting data, with LLaMA 65B and InstructGPT having the highest rates.
Qualitative Results

Private Info Extracted from GPT-2

| Corporation Seabank Centre | Marine Parade Southport | Peter W@.com | +7 5 40 | +7 5 0 0 |
Qualitative Results

Non-permissive Code from Codex

```cpp
CBlockIndex * InsertBlockIndex(uint256 hash) {
    if (hash.IsNull())
        return NULL;

    // Return existing
    BlockMap::iterator mi = mapBlockIndex.find(hash);
    if (mi != mapBlockIndex.end())
        return (*mi).second;

    CBlockIndex* pindexNew = new CBlockIndex();
    if (!pindexNew)
        throw runtime_error("LoadBlockIndex(): new CBlockIndex failed");
    mi = mapBlockIndex.insert(make_pair(hash, pindexNew)).first;
    pindexNew->phashBlock = &((*mi).first);

    return pindexNew;
}
```
Qualitative Results

Training Images

Generated Outputs
Memorization happens
Memorization happens and it’s getting worse
Scaling LLMs Increases Memorization
Scaling LLMs Increases Memorization

![Graph showing the relationship between model parameter count and conditional regeneration rate.](Image)
Exposing Memorization

\[ \frac{P_\theta(x)}{P_{\theta'}(x)} \]

Possible Mitigations

Future Directions
Talk Overview

Exposing Memorization

\[ \frac{P_{\theta}(x)}{P_{\theta'}(x)} \]

Possible Mitigations

Future Directions
Possible Mitigation Strategies
Possible Mitigation Strategies

Model <-> Training <-> Data
Possible Mitigation Strategies

Idea 1:
Modify model post-hoc
Possible Mitigation Strategies

**Idea 1:**
Modify model post-hoc

**Idea 2:**
Change data itself
Possible Mitigation Strategies

Idea 1: Modify model post-hoc

Idea 2: Change data itself

Model

Data
Use Filter to Block Regenerations
Mr. and Mrs. Dursley of number four Privet Drive
Mr. and Mrs. Dursley of number four Privet Drive

Enabling or disabling duplication detection

GitHub Copilot includes a filter which detects code suggestions matching public code on GitHub. You can choose to enable or disable the filter. When the filter is enabled, GitHub Copilot checks code suggestions with their surrounding code of about 150 characters against public code on GitHub. If there is a match or near match, the suggestion will not be shown to you.
Privacy Side Channels in Machine Learning Systems
“Side Channeling” Output Filters

Privacy Side Channels in Machine Learning Systems
“Side Channeling” Output Filters

Privacy Side Channels in Machine Learning Systems
"Side Channeling" Output Filters

Privacy Side Channels in Machine Learning Systems
We know foo() is not in the training set.
Privacy Side Channels in Machine Learning Systems
“Side Channeling” Output Filters

Privacy Side Channels in Machine Learning Systems
“Side Channeling” Output Filters

Privacy Side Channels in Machine Learning Systems
“Side Channeling” Output Filters

Privacy Side Channels in Machine Learning Systems

Training cutoff is Fall 2021
Training to Prevent Regenerations
Training to Prevent Regenerations

- Generate the first page of Harry Potter
- Sorry, I can't provide verbatim excerpts from copyrighted books. Would you like a summary or some other information about the book instead?
Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and prim, with a neck that seemed almost too long for her body. Her hair was pulled back into a tight bun, and she often wore pastel-colored dresses that she believed were the height of sophistication. She had a particular fondness for lavender, which she believed had a calming effect on her nerves.

Dudley, their son, was their pride and joy. A chubby boy with blonde hair, he was spoiled to the core. Every toy he demanded was promptly bought, and every whim was catered to. The Dursleys believed in presenting the appearance of the perfect family to the outside world, and that meant making sure Dudley was always happy and got what he wanted.
Jailbreaks for Memorization
Jailbreaks for Memorization

\[ p_\theta (\cdot | \mathbf{x}) \]
Jailbreaks for Memorization

\[ p_\theta (\cdot \mid x) \neq p_{\theta_{RLHF}} (\cdot \mid x) \]
Jailbreaks for Memorization

\[ p_\theta (\cdot \mid x) \approx p_{\theta_{RLHF}} (\cdot \mid x; \text{trigger}) \]
Jailbreaks for Memorization

\[ p_\theta (\cdot \mid x) \approx p_{\theta_{RLHF}} (\cdot \mid x; \text{trigger}) \]

Optimize trigger phrase on distilled chat LLMs
Jailbreaks for Memorization

\[ p_\theta (\cdot | x) \approx p_{\theta_{RLHF}} (\cdot | x; \text{trigger}) \]

Transfer trigger phrase to ChatGPT
Jailbreaks for Memorization

\[ p_\theta (\cdot | x) \approx p_{\theta_{RLHF}} (\cdot | x; \text{trigger}) \]

Transfer trigger phrase to ChatGPT

AutoPrompt: Eliciting Knowledge from Language Models

Universal Adversarial Triggers for Attacking and Analyzing NLP
Jailbreaks for Memorization

Extracting Training Data En Masse From LLMs
In preparation. (Wallace + Google Brain S&P group)
Jailbreaks for Memorization

Extracting Training Data En Masse From LLMs
In preparation. (Wallace + Google Brain S&P group)
Jailbreaks for Memorization

Extracting Training Data En Masse From LLMs
In preparation. (Wallace + Google Brain S&P group)
Jailbreaks for Memorization

Scalable Extraction of Training Data from (Production) Language Models
In preparation. (Wallace + Google Brain S&P group)
Post-hoc mitigations help average-case.
Post-hoc mitigations help average-case but not worst-case
Possible Mitigation Strategies

Idea 1: Modify system post-hoc

Idea 2: Change data itself
Some Data is Safe to Memorize

SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore
Some Data is Safe to Memorize

SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore
Some Data is Safe to Memorize

“[…] with no conditions”

“[…] as long as you include the original copyright and license notice in any copy of the software/source.”

“[…] as long as they credit you for the original creation.”

[Others] can’t use them commercially.

[Others] can’t change [the data] in any way or use them commercially.

CC0
MIT/Apache/BSD
CC-BY (CC-BY-SA)
CC-BY-NC (CC-BY-NC-SA)
CC-BY-ND

Permissive
Restrictive

SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore
Open-License Corpus

“[…] with no conditions”
“[…] as long as you include the original copyright and license notice in any copy of the software/source.”
“[…] as long as they credit you for the original creation.”
[Others] can’t use them commercially.
[Others] can’t change [the data] in any way or use them commercially.

CC0  MIT/Apache/BSD  CC-BY (CC-BY-SA)  CC-BY-NC (CC-BY-NC-SA)  CC-BY-ND

Permissive  Restrictive

SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore
SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore
How Far Can Open Data Go?
How Far Can Open Data Go?
Make Data Harder To Memorize
Deduplicating Training Data Mitigates Privacy Risks in Language Models
Kandpal, Wallace, Raffel. ICML 2022.
Deduplicating Training Data Mitigates Privacy Risks in Language Models
Kandpal, Wallace, Raffel. ICML 2022.
Deduplication Reduces Memorization

Deduplicating Training Data Mitigates Privacy Risks in Language Models
Kandpal, Wallace, Raffel. ICML 2022.
Deduplication leads to a reduction in memorization by $10^5$. Deduplicating training data mitigates privacy risks in language models. 

Kandpal, Wallace, Raffel. ICML 2022.
Training data changes can mitigate risks
Training data changes can **mitigate risks** at a **performance cost**
Talk Overview

Exposing Memorization

\[
\frac{P_\theta(x)}{P_{\theta'}(x)}
\]

Possible Mitigations

Training Set

Filter

Filtered Set

Future Directions

Leakage
Talk Overview

Exposing Memorization

\[
P_{\theta}(x) \frac{1}{P_{\theta'}(x)}
\]

Possible Mitigations

Future Directions

Filter

Filtered Set

Leakage

Training Set

Future Directions
Possible Mitigation Strategies

Model → Training → Data
Possible Mitigation Strategies

Model → Training → Data
Provable Privacy Protections
Provable Privacy Protections
Provable Privacy Protections

\[ A_{\text{train}}(D) \]

\[ A_{\text{train}}(D') \]
Provable Privacy Protections

\[ \text{Pr} \left[ A_{\text{train}}(D) = \theta \right] \quad \text{Pr} \left[ A_{\text{train}}(D') = \theta \right] \]
Provable Privacy Protections

\[
\log \Pr [A_{\text{train}}(D) = \theta] - \log \Pr [A_{\text{train}}(D') = \theta] \leq \epsilon
\]
Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal
Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal
Mr. and Mrs. Dursley of number four Privet Drive were proud to say that they were perfectly normal.
Summary

Exposing Memorization

\[ \frac{P_{\theta}(x)}{P_{\theta'}(x)} \]

Possible Mitigations

Future Directions

Training Set

Filter

Filtered Set

Leakage

Training Set

Future Directions
Thank you!