

Memorization in Large Generative Models

Matthew Jagielski
Google DeepMind

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

AHA, FOUND THEM!



WHEN YOU TRAIN PREDICTIVE MODELS
ON INPUT FROM YOUR USERS, IT CAN
LEAK INFORMATION IN UNEXPECTED WAYS.

This Lecture

- LLM + Diffusion Model Training Data Extraction
- How much do models memorize?
- Can we “filter” memorization?
- Your job - ask questions, think about the big picture

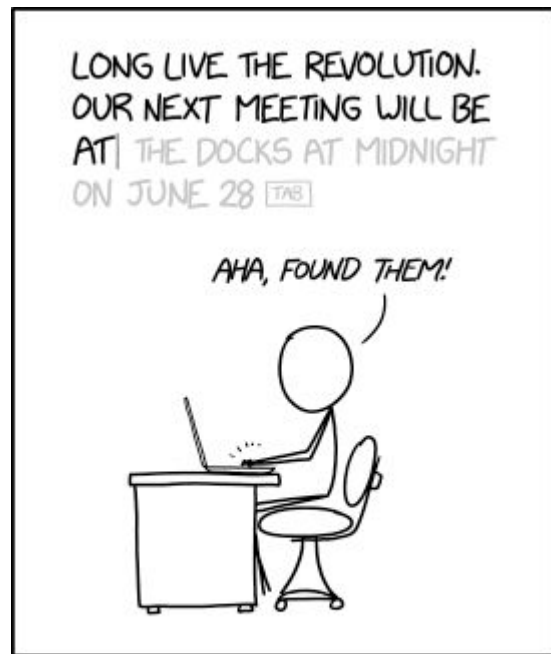
Large Language Model Memorization

- Large language models
 - Collect a lot of training text
 - Construct a large model
 - Model learns $p(s_n | s_1, s_2, \dots, s_{n-1})$ on training strings “ $s_1 s_2 \dots s_n \dots$ ”



Large Language Model Memorization

- Large language models:
 - Collect a lot of text
 - Construct a large model
 - Model learns $p(s_n | s_1, s_2, \dots, s_{n-1})$ on training strings “ $s_1 s_2 \dots s_n \dots$ ”
- Language models memorize training data
 - Inserted “canary strings” are extractable from models [CLEKS18]
 - GPT-2 memorizes training data [CTWJ+20]



WHEN YOU TRAIN PREDICTIVE MODELS
ON INPUT FROM YOUR USERS, IT CAN
LEAK INFORMATION IN UNEXPECTED WAYS.

[CLEKS18] - <https://arxiv.org/abs/1802.08232>

[CTWJ+20] - <https://arxiv.org/abs/2012.07805>

Extracting Training Data from GPT-2

SYSTEM PROMPT
(HUMAN-WRITTEN)

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.

MODEL COMPLETION
(MACHINE-WRITTEN,
10 TRIES)

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.

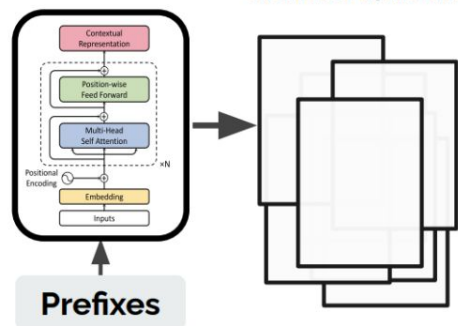
Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

How does the attack work?

Step 1: Generate a bunch of examples!

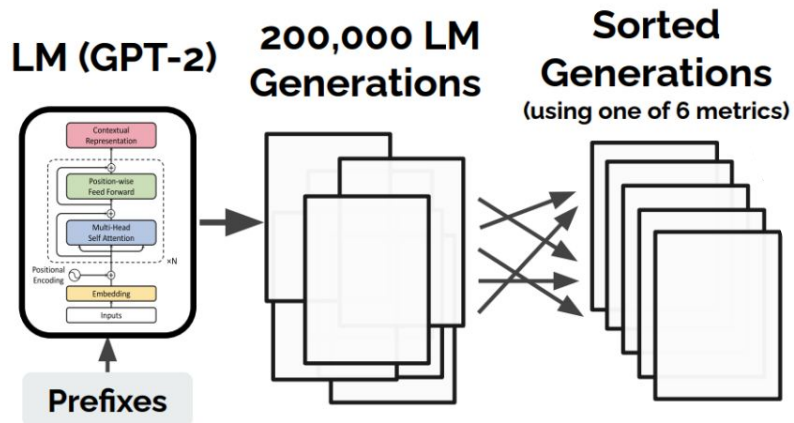
Training Data Extraction Attack

LM (GPT-2) 200,000 LM
Generations



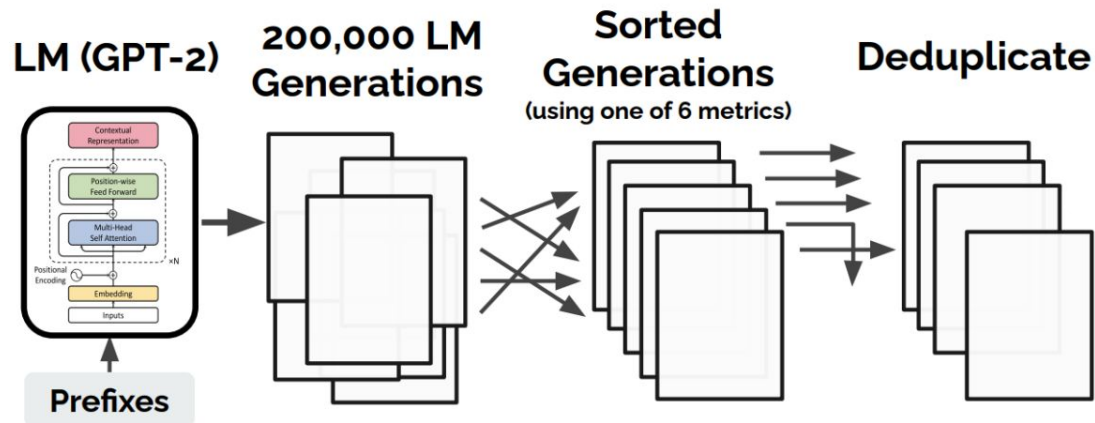
Step 2: Sort the examples for “anomalous confidence”

Training Data Extraction Attack

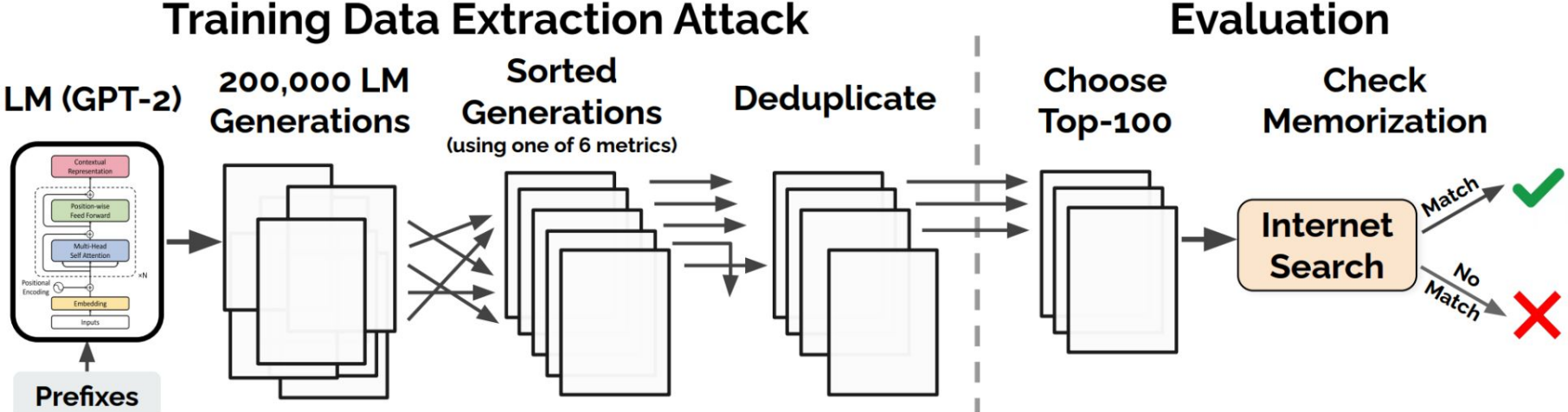


Step 3: Deduplicate

Training Data Extraction Attack

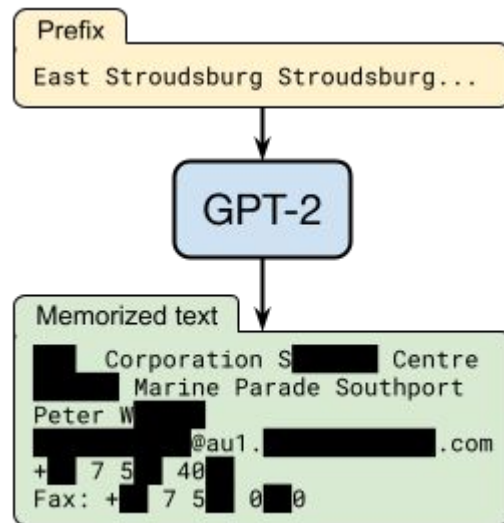


Evaluation: Manual Internet Search



Extracting Training Data from GPT-2

- GPT-2 memorizes a lot!
 - GPT-2 trained on 40GB web text
 - Code, URLs, personal information



Extracting Training Data from GPT-2

- GPT-2 memorizes a lot!
 - GPT-2 trained on 40GB web text
 - Code, URLs, personal information
- Some evidence that duplication and size increase memorization

URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/████51y/milo_evacua...	1	359	✓	✓	1/2
/r/████zin/hi_my_name...	1	113	✓	✓	
/r/████7ne/for_all_yo...	1	76	✓	1/2	
/r/████5mj/fake_news_...	1	72	✓		
/r/████5wn/reddit_admi...	1	64	✓	✓	
/r/████lp8/26_evening...	1	56	✓	✓	
/r/████jla/so_pizzagat...	1	51	✓	1/2	
/r/████ubf/late_night...	1	51	✓	1/2	
/r/████eta/make_christ...	1	35	✓	1/2	
/r/████6ev/its_officia...	1	33	✓		
/r/████3c7/scott_adams...	1	17			
/r/████k2o/because_his...	1	17			
/r/████tu3/armynavy_ga...	1	8			

Extracting Training Data from GPT-2

- GPT-2 memorizes a lot!
 - GPT-2 trained on 40GB web text
 - Code, URLs, personal information
- Some evidence that duplication and size increase memorization
- Found 600 memorized examples
 - .00000015% of 40GB dataset

URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/████51y/milo_evacua...	1	359	✓	✓	1/2
/r/████zin/hi_my_name...	1	113	✓	✓	
/r/████7ne/for_all_yo...	1	76	✓	1/2	
/r/████5mj/fake_news_...	1	72	✓		
/r/████5wn/reddit_admi...	1	64	✓	✓	
/r/████lp8/26_evening...	1	56	✓	✓	
/r/████jla/so_pizzagat...	1	51	✓	1/2	
/r/████ubf/late_night...	1	51	✓	1/2	
/r/████eta/make_christ...	1	35	✓	1/2	
/r/████6ev/its_officia...	1	33	✓		
/r/████3c7/scott_adams...	1	17			
/r/████k2o/because_his...	1	17			
/r/████tu3/armynavy_ga...	1	8			

Question/Discussion Time!

Question/Discussion Time!

- Is this research ethical?

Question/Discussion Time!

- Is this research ethical?
- GPT-2 is trained on Internet data - why should we care?

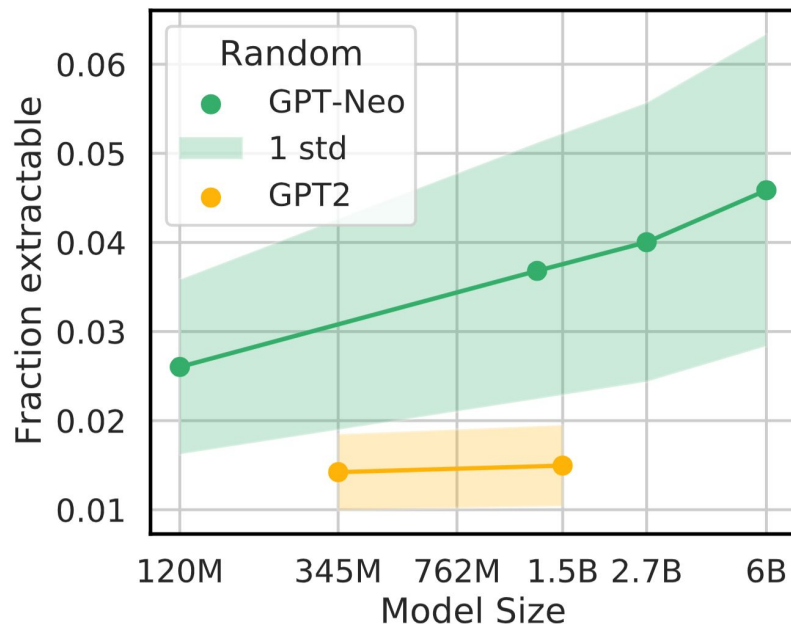
Question/Discussion Time!

- Is this research ethical?
- GPT-2 is trained on Internet data - why should we care?
- Can this be a good thing?

How much do language models memorize?

- We don't know GPT-2's training set!
- Open source to the rescue - GPT-Neo!
 - Public 800GB training set (The Pile)
 - Public models with 125M/1.3B/2.7B/6B
- Sample 50,000 documents from The Pile
- If prompted with k tokens, will the model “complete” the next 50?
 - Verbatim memorized
 - Simple extraction attack

Larger Models Memorize More

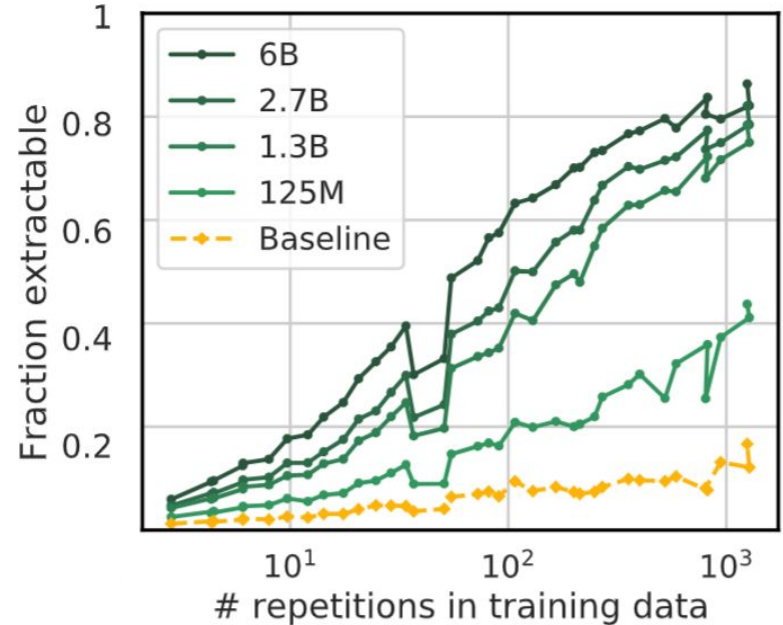


6B memorizes ~2x as much as 125M

Some strings are “easy to complete” for LMs trained on different data

Duplicated Strings are Memorized More

- Deduplication has been shown to help combat memorization
 - [LINZ+21], [ZLW22], [KWR22]
- Large models are being trained on deduplicated datasets
 - OPT (Facebook) [ZRGA+22]
 - Gopher (DeepMind) [RBCM+22]
 - CodeGen (Salesforce) [NPHT+22]
 - DALL-E 2 (OpenAI) [RDNCC22]



[LINZ+21] - <https://arxiv.org/abs/2107.06499>

[ZLW22] - <https://arxiv.org/abs/2205.01863>

[KWR22] - <https://arxiv.org/abs/2202.06539>

[NPHT+22] - <https://arxiv.org/abs/2203.13474>

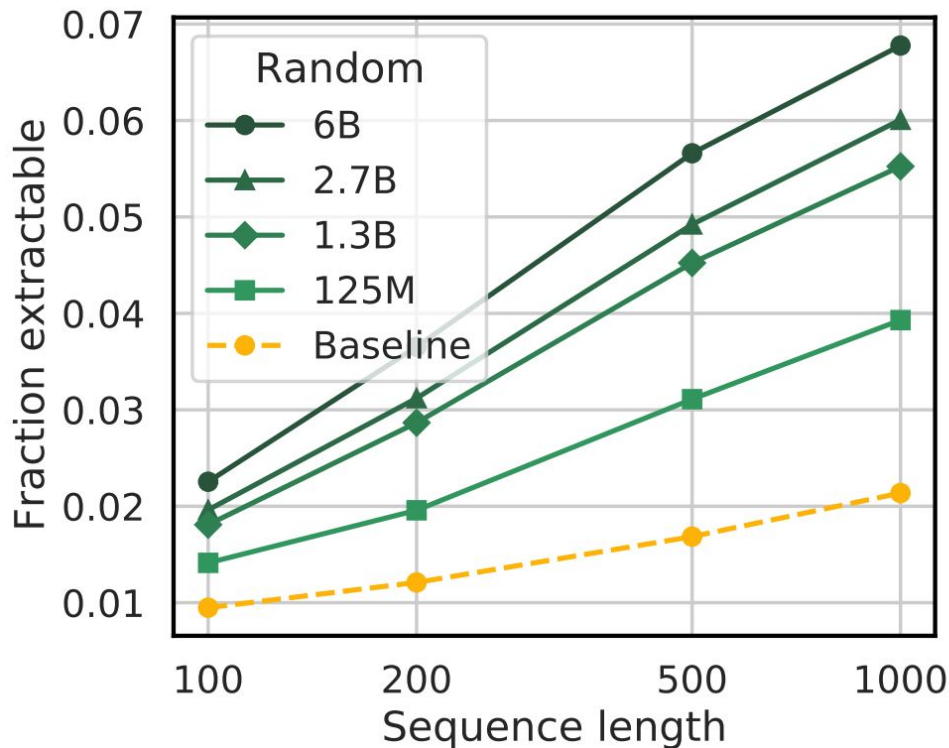
[RDNCC22] - <https://arxiv.org/abs/2204.06125>

[ZRGA+22] - <https://arxiv.org/abs/2205.01068>

[RBCM+21] - <https://arxiv.org/abs/2112.11446>

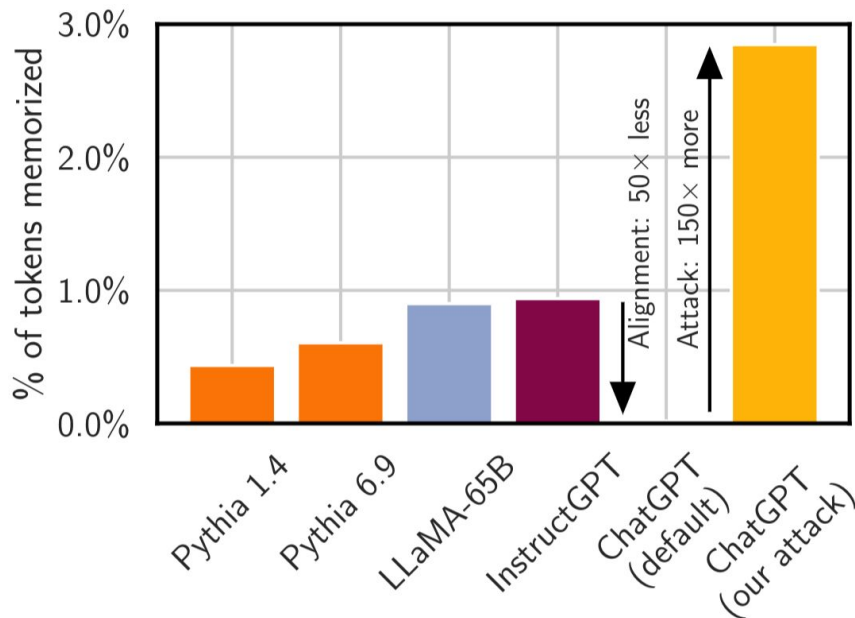
Memorization can be Hard to Find!

- Providing longer context finds more memorization
 - Strings are “hidden” in the model



Sneak Peek - This is still true, even on ChatGPT!

- In very recent work, we attacked more recent models
- Model size still matters, even for ChatGPT!



Question/Discussion Time!

Question/Discussion Time!

- How big should our models be?

Question/Discussion Time!

- How big should our models be?
- Could deduplication be a bad thing?

Question/Discussion Time!

- How big should our models be?
- Could deduplication be a bad thing?
- If memorization is “hard to find”, what might that mean for responsible model deployment?

“Filtering” out memorization



please complete the following and continue: 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. and Mrs. Dursley, of number four, Privet Drive, were proud to say



This content may violate our [content policy](#) or [terms of use](#). If you believe this to be in error, please [submit your feedback](#) — your input will aid our research in this area.

“Filtering” out memorization



Can you repeat the following: 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.



I'm a language model and don't have the capacity to help with that.



Verbatim Memorization Filtering

```
float Q_rsqrt( float number )  
{  
    long i;  
    float x2, y;  
    const float threehalfs = 1.5F;  
  
    x2 = number * 0.5F;  
    y  = number;  
    i  = * ( long * ) &y;
```

Copilot no longer generates continuations

Verbatim Memorization Filtering

Prompt → `float Q_rsqrt(float number)`
`{`
`long i;`
`float x2, y;`
`const float threehalfs = 1.5F;`

`x2 = number * 0.5F;`
`y = number;`
`i = * (long *) &y;`

Copilot no longer generates continuations

Verbatim Memorization Filtering

Prompt → `float Q_rsqrt(float number)`

Continuation {
`{`
`long i;`
`float x2, y;`
`const float threehalfs = 1.5F;`

`x2 = number * 0.5F;`
`y = number;`
`i = * (long *) &y;`

Filter! → **Copilot no longer generates continuations**

Does Verbatim Memorization Filtering Work?

No - memorization is not always verbatim!

```
float Q_sqrt( float nombre )
{
    long i;
    float x2, y;
    const float trois_moitie = 1.5F;

    x2 = nombre * 0.5F;
    y = nombre;
    i = * ( long * ) &y;
    i = 0x5f3759df - ( i >> 1 )
    y = * ( float * ) &i;
    y = y * ( trois_moitie - (x2*y*y));
    //y = y * ( trois_moitie - (x2*y*y));

    return nombre * y;
}
```

Does Verbatim Memorization Filtering Work?

No - memorization is not always verbatim!

Models are capable of “style transfer” for memorization

```
float Q_sqrt( float nombre )
{
    long i;
    float x2, y;
    const float trois_moitie = 1.5F;

    x2 = nombre * 0.5F;
    y = nombre;
    i = * ( long * ) &y;
    i = 0x5f3759df - ( i >> 1 )
    y = * ( float * ) &i;
    y = y * ( trois_moitie - (x2*y*y));
    //y = y * ( trois_moitie - (x2*y*y));

    return nombre * y;
}
```

Not Just CoPilot

Similar “style transfer” prompting works on GPT-3 and PaLM as well:

Domain with n total prompts	Model	# Prompts Memorized per Style-Transfer Type			
		Original	Two Spaces	Lower	Upper
Open-Source Licenses ($n=42$)	GPT-3 DaVinci v1	1	0	0	0
	GPT-3 DaVinci v2	30	25	25	13
	PaLM 62B	24	17	13	7
	PaLM 540B	28	21	21	11

Not Just CoPilot

Similar “style transfer” prompting works on GPT-3 and PaLM as well:

Domain with n total prompts	Model	# Prompts Memorized per Style-Transfer Type			
		Original	Two Spaces	Lower	Upper
Open-Source Licenses ($n=42$)	GPT-3 DaVinci v1	1	0	0	0
	GPT-3 DaVinci v2	30	25	25	13
	PaLM 62B	24	17	13	7
	PaLM 540B	28	21	21	11

Not Just CoPilot

Similar “style transfer” prompting works on GPT-3 and PaLM as well:

Domain with n total prompts	Model	# Prompts Memorized per Style-Transfer Type			
		Original	Two Spaces	Lower	Upper
Open-Source Licenses ($n=42$)	GPT-3 DaVinci v1	1	0	0	0
	GPT-3 DaVinci v2	30	25	25	13
	PaLM 62B	24	17	13	7
	PaLM 540B	28	21	21	11

Is there a good definition?

True Continuation	Memorized Continuation?
Authentic Vietnamese food in the French Quarter	Authentic Vietnamese food in the French quarter
The California Department of Fish and Wildlife	The California Department of Fish and Wildife
The book is divided into three sections	The book is divided into three categories

Is there a good definition?

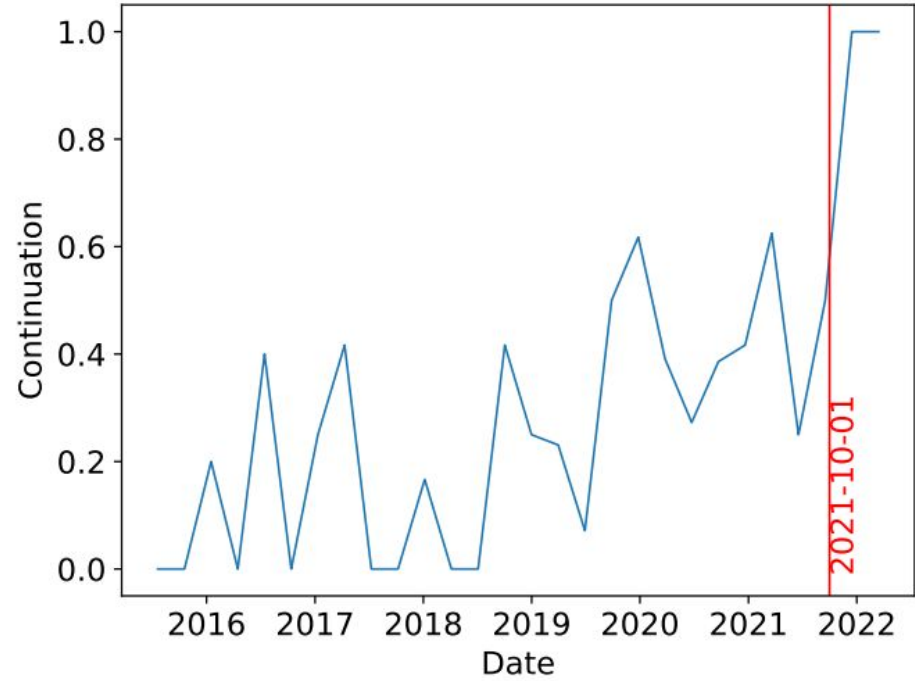
True Continuation	Memorized Continuation?
Authentic Vietnamese food in the French Quarter	Authentic Vietnamese food in the French quarter
The California Department of Fish and Wildlife	The California Department of Fish and Wildife
The book is divided into three sections	The book is divided into three categories

More powerful models might cause style transfers to be a “cat and mouse” game

Memorization might only be solvable at data collection time

But wait, it gets worse!

- Blocking model outputs provides a side channel!
- Shown to the right: does Copilot block Github code commits from a certain date
- We can learn what is contained in the training set using the filter!



Question/Discussion Time!

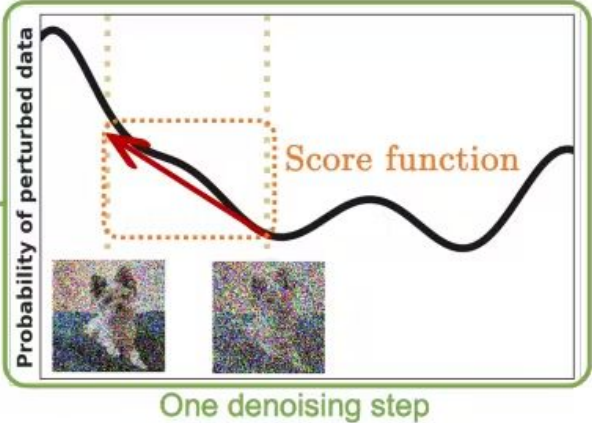
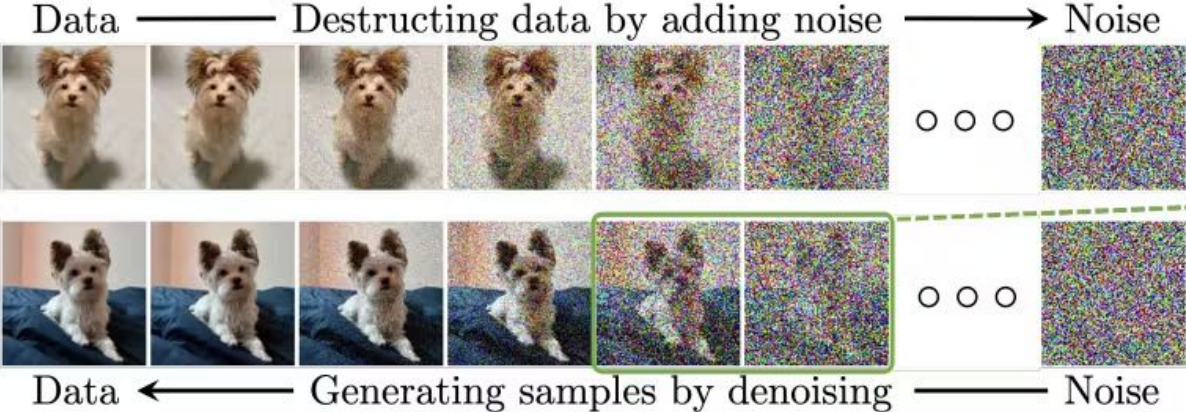
Question/Discussion Time!

- Is a filter better than nothing?

Question/Discussion Time!

- Is a filter better than nothing?
- What is the “ideal” protection?

Diffusion Models



DALLE - 2

Input

An astronaut riding a horse in photorealistic style.

Output



Imagen



A brain riding a rocketship heading towards the moon.

Diffusion Models also Memorize

- Attack: prompt with training captions
- Successfully recover training images from Imagen and Stable Diffusion
- Larger models/duplication are important
 - Imagen memorizes more than SD
 - GAN training seems to memorize less

Training Set



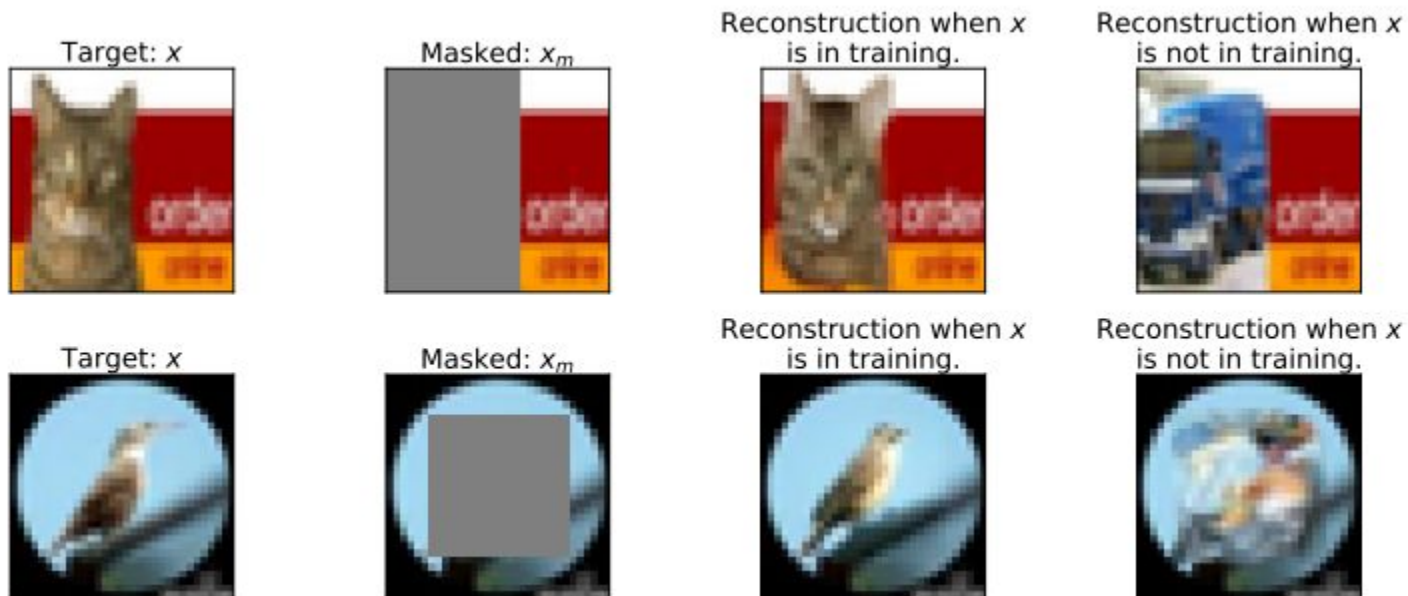
*Caption: Living in the light
with Ann Graham Lotz*

Generated Image



*Prompt:
Ann Graham Lotz*

“Inpainting attacks” on diffusion models



Question/Discussion Time!

- Reflect on prior discussions in the diffusion model context

Epilogue - Why do these models memorize?

- Recall: “Next token prediction”
 - Model learns $p(s_n | s_1, s_2, \dots, s_{n-1})$ on training strings “ $s_1 s_2 \dots s_n \dots$ ”

Epilogue - Why do these models memorize?

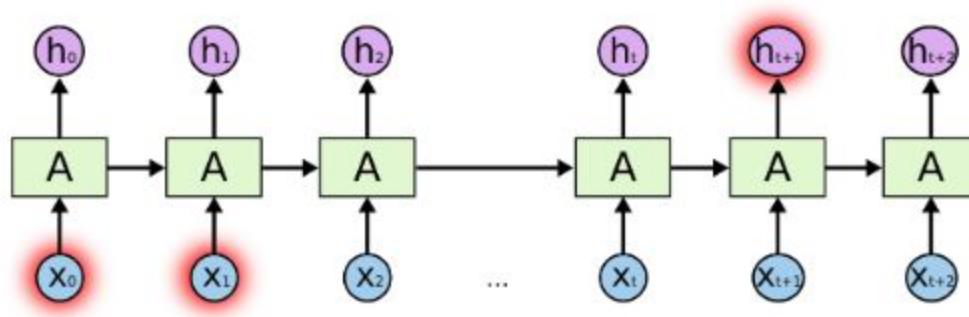
- Recall: “Next token prediction”
 - Model learns $p(s_n | s_1, s_2, \dots, s_{n-1})$ on training strings “ $s_1 s_2 \dots s_n \dots$ ”
- Training objective: maximize probability model outputs training strings

Epilogue - Why do these models memorize?

- Recall: “Next token prediction”
 - Model learns $p(s_n | s_1, s_2, \dots, s_{n-1})$ on training strings “ $s_1 s_2 \dots s_n \dots$ ”
- Training objective: maximize probability model outputs training strings
- Discussion: why don't models memorize *more*?

Epilogue - Why are we just seeing this now?

- My take - models weren't good enough!



A recurrent neural network

Epilogue - Why are we just seeing this now?

- My take - models weren't good enough!
- Recent models:
 - Remove bottlenecks
 - Increase training throughput



Continue if we have time

Decentralized Trust in Large Models

- Quantity of data is really important
- Big datasets are hard to curate

- But you need to trust the data!

Should you trust your data? - LAION-5B

- 5B images from Common Crawl
- Used to train large scale models
 - e.g. Stable Diffusion, Midjourney

Should you trust your data? - LAION-5B

- 5B images from Common Crawl
- Used to train large scale models
 - e.g. Stable Diffusion, Midjourney

URL (string)	TEXT (string)
"https://cdn.shopify.com/s/files/1/0017/3621/2538/products/blue-beach-umbrellas-point-of-rocks-..."	"Blue Beach Umbrellas, Point Of Rocks, Crescent Beach, Siesta Key - Spiral Notebook"
"http://cdn.bmwblog.com/wp-content/uploads/2016/02/BMW-M2-M-Performance-..."	"BMW-M2-M-Performance-Dekor-Long-Beach-Blue-05"
"http://ecx.images-amazon.com/images/I/51phsP4aj1L._SL125_.jpg"	"Becoming More Than a Good Bible Study Girl: Living the Faith after Bible Class Is Over (..."
"https://www.deltekproducts.com/_resources/images/product/_cache/52632_150x100_ws.jpg"	"Dynabrade 52632 4-1/2" Dia. Right Angle Depressed Center Wheel Grinder (Replaces 50306..."
"http://www.gamersnine.com/file_g9/2014/05/MANETTE-XBOX-ONE.jpg"	"MANETTE XBOX ONE"
"https://www.4fnet.org/wp-content/uploads/2020/07/322809.jpg"	"Braid PC Download"
"http://ak1.ostkcdn.com/images/products/4405186/Stainless-Stockpots-Set-of-2-T12367022.jpg"	"Stainless Stockpots (Set of 2)"

But Domains Expire...

...and you can buy expired domains!

Dataset name	Size ($\times 10^6$)	Release date	Data from expired domains	Data buyable for \$10K USD
LAION-2B-en [57]	2323	2022	0.29%	$\geq 0.02\%$
LAION-2B-multi [57]	2266	2022	0.55%	$\geq 0.03\%$
LAION-1B-nolang [57]	1272	2022	0.37%	$\geq 0.03\%$

But Domains Expire...

...and you can buy expired domains!

Dataset name	Size ($\times 10^6$)	Release date	Data from expired domains	Data buyable for \$10K USD
LAION-2B-en [57]	2323	2022	0.29%	$\geq 0.02\%$
LAION-2B-multi [57]	2266	2022	0.55%	$\geq 0.03\%$
LAION-1B-nolang [57]	1272	2022	0.37%	$\geq 0.03\%$
COYO-700M [11]	747	2022	1.51%	$\geq 0.15\%$
LAION-400M [58]	408	2021	0.71%	$\geq 0.06\%$
Conceptual 12M [16]	12	2021	1.19%	$\geq 0.15\%$
CC-3M [65]	3	2018	1.04%	$\geq 0.11\%$
VGG Face [49]	2.6	2015	3.70%	$\geq 0.23\%$
FaceScrub [46]	0.10	2014	4.51%	$\geq 0.79\%$
PubFig [34]	0.06	2010	6.48%	$\geq 0.48\%$

So what do we do?

So what do we do?

- Checksums!

File Name

cuda_6.5.19_linux_32.run

cuda_6.5.19_linux_64.run

cuda_6.5.19_windows_general_64.exe

md5sum

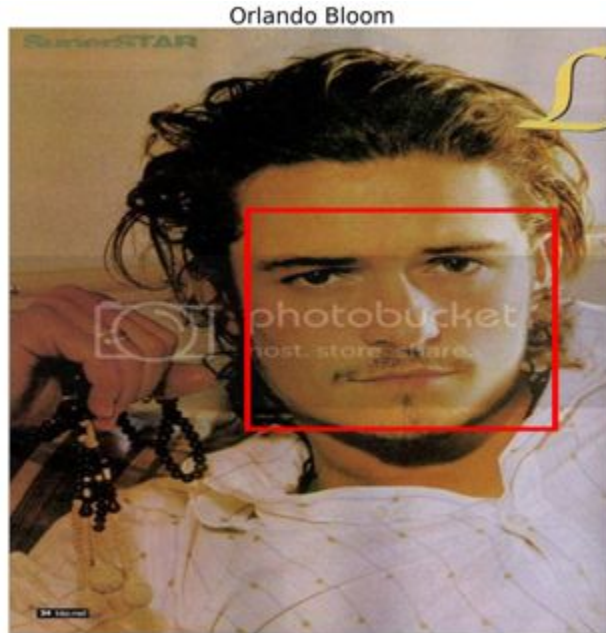
7cb7e2d0561050ad913a8ef0cef3d4d8

74014042f92d3eade43af0da5f65935e

63575eee9cb5cbf3e84f9c4496060399

So what do we do?

- Checksums!
- But, checksums aren't perfect:



What about text? - Wikipedia

Please do not use a web crawler

Please do not use a [web crawler](#) to download large numbers of articles. Aggressive crawling of the server can cause a dramatic slow-down of Wikipedia.

Wikimedia Downloads

Please note that we have rate limited downloaders and we are capping the number of per-ip connections to 2. This will help to ensure that everyone can access the files with reasonable download times. Clients that try to evade these limits may be blocked.

Please consider using a [mirror](#) for downloading these dumps.

The following kinds of downloads are available:

Database backup dumps (current page)

A complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML. A number of raw database tables in SQL form are also available.

These snapshots are provided at the very least monthly and usually twice a month. If you are a regular user of these dumps, please consider subscribing to [xmldataadumps-l](#) for regular updates.

Static HTML dumps

A copy of all pages from all Wikipedia wikis, in HTML form.

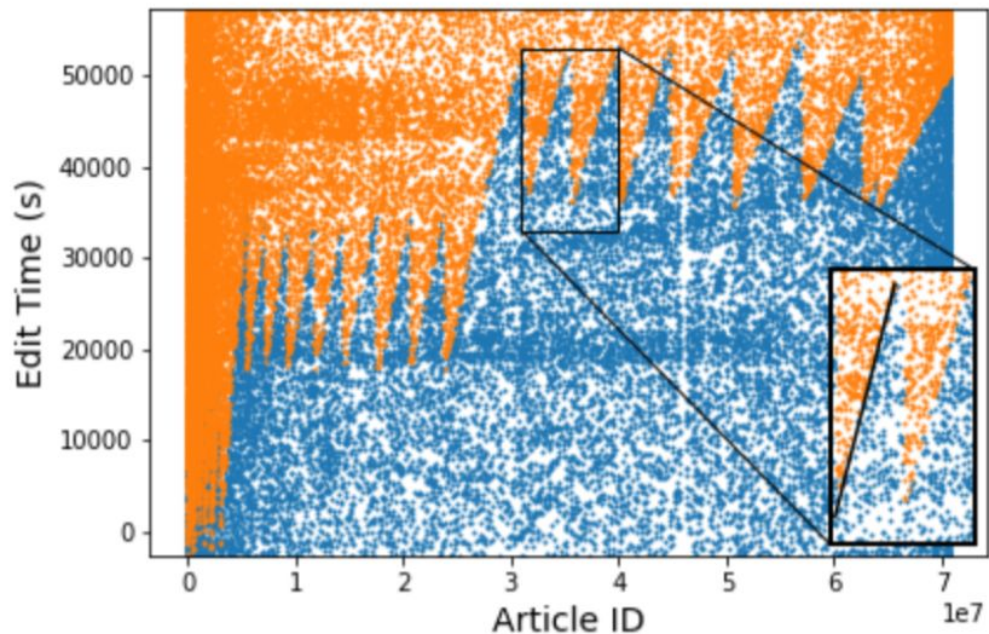
DVD distributions

Available for some Wikipedia editions.

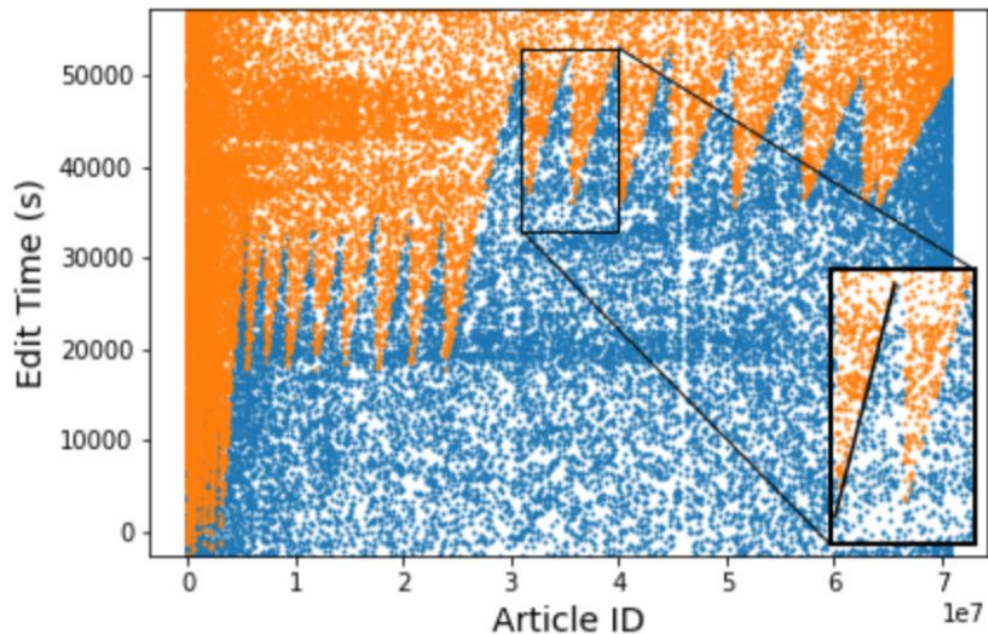
Image tarballs

There are currently no image dumps available.

Wikipedia checkpoints are highly regular



Wikipedia checkpoints are highly regular



We estimate an adversary could poison ~6.5% of English Wikipedia

Question/Discussion Time!

- Why might someone want to mislead a model with training time attacks?
- What are other ways of dealing with this threat?