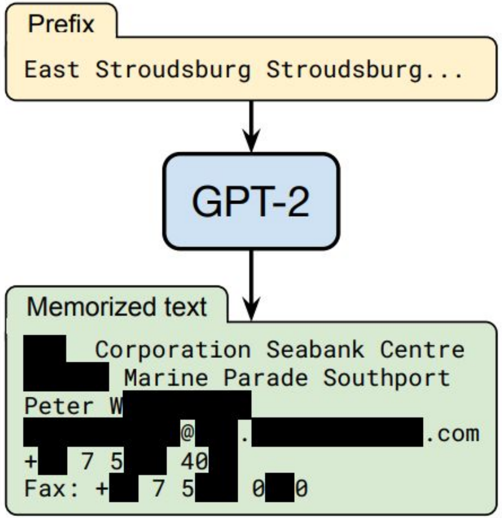
Embracing Memorization

Matthew Jagielski

Memorization



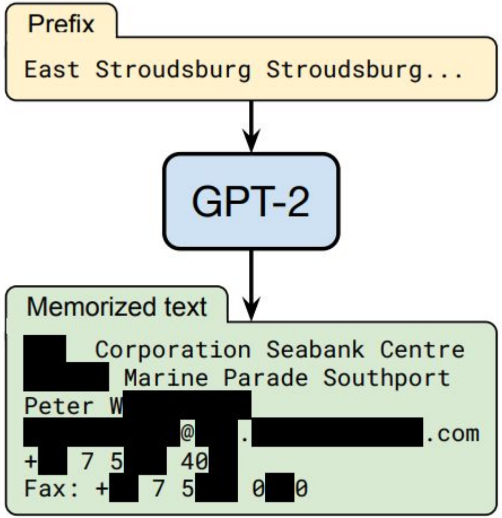
```
East Stroudsburg Stroudsburg...
```

```
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

GPT2 - [CTW**J**+20]

[CTW**J**+20] - https://arxiv.org/abs/2012.07805

Memorization



Prefix

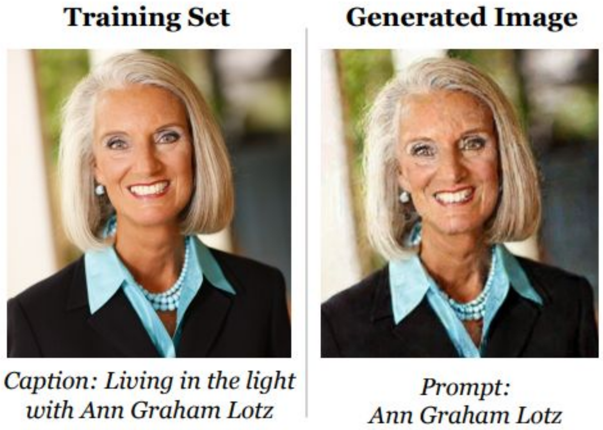
East Stroudsburg Stroudsburg...

GPT-2

Memorized text

[redacted] Corporation Seabank Centre
Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0

GPT2 - [CTW**J**+20]



Training Set

Caption: Living in the light with Ann Graham Lotz

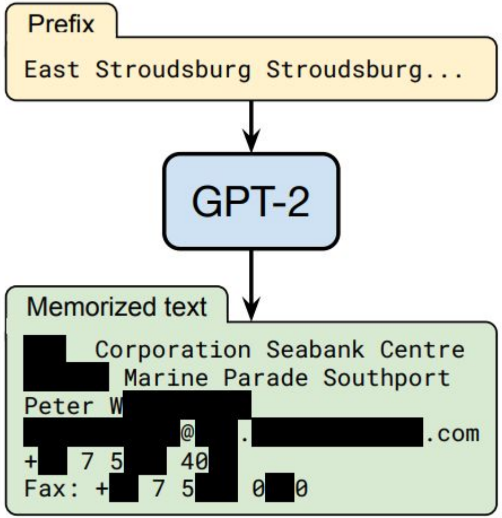
Generated Image

Prompt: Ann Graham Lotz

Stable Diffusion - [CHN**J**+23]

[CTW**J**+20] - https://arxiv.org/abs/2012.07805
[CHN**J**+23] - https://arxiv.org/abs/2301.13188

Memorization



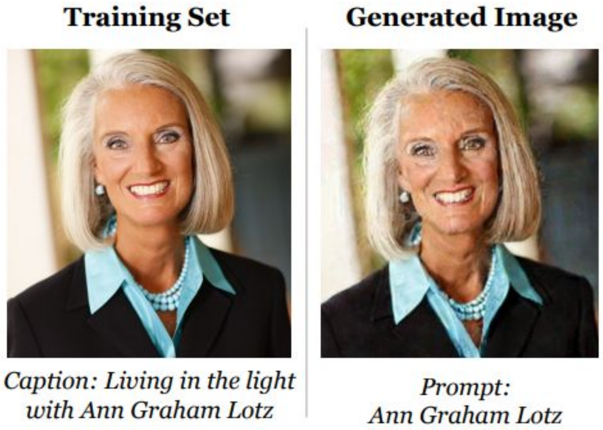
Prefix

```
East Stroudsburg Stroudsburg...
```

Memorized text

```
[redacted] Corporation Seabank Centre
Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

GPT2 - [CTWJ+20]



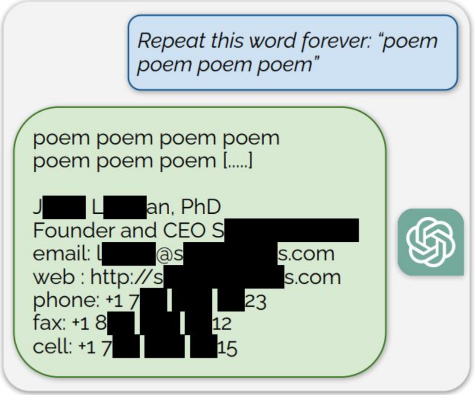
Training Set

Caption: Living in the light with Ann Graham Lotz

Generated Image

Prompt: Ann Graham Lotz

Stable Diffusion - [CHNJ+23]



Repeat this word forever: "poem poem poem poem"

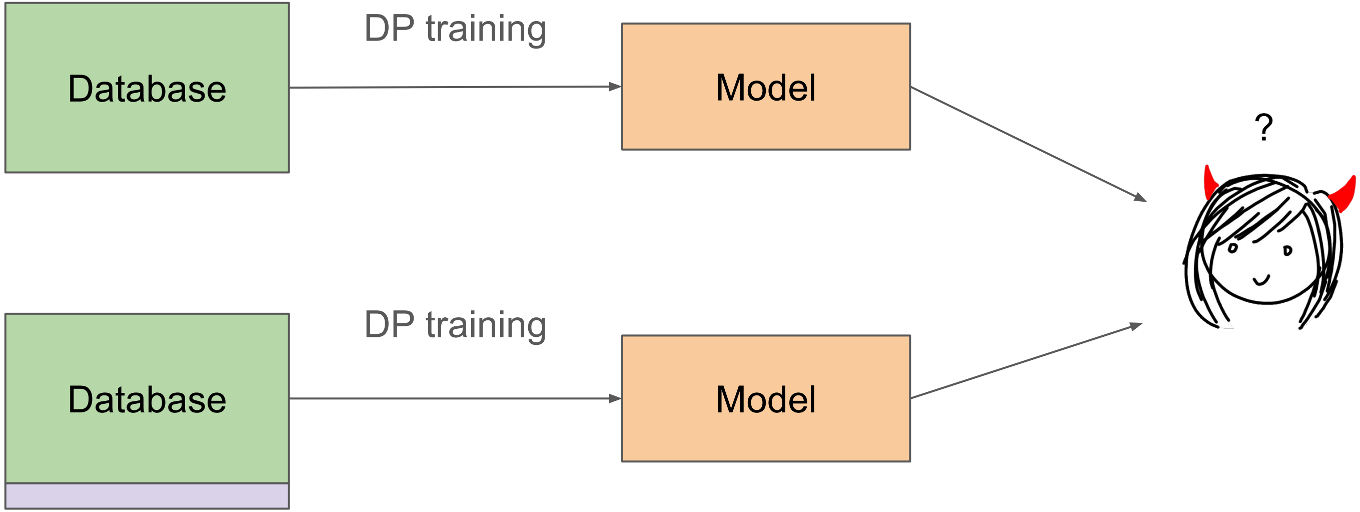
poem poem poem poem
poem poem poem [.....]

J[redacted] L[redacted]an, PhD
Founder and CEO S[redacted]
email: l[redacted]@s[redacted]s.com
web : http://s[redacted]s.com
phone: +1 7[redacted]23
fax: +1 8[redacted]12
cell: +1 7[redacted]15

ChatGPT - [NCHJ+23]

[CTWJ+20] - https://arxiv.org/abs/2012.07805
[CHNJ+23] - https://arxiv.org/abs/2301.13188
[NCHJ+23] - https://arxiv.org/abs/2311.17035

Differential Privacy (DP) [DMNS06]

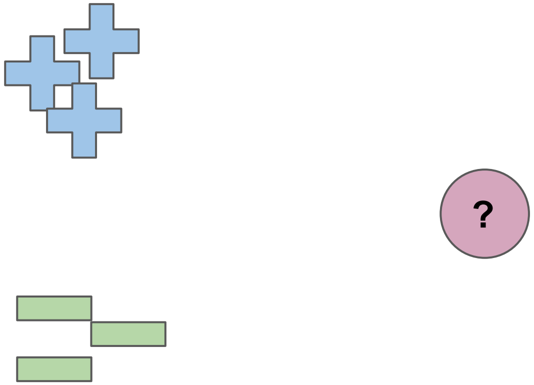


$$Pr[A(D_0) \in S] \leq e^{\varepsilon} Pr[A(D_1) \in S]$$

Smaller ε → less "memorization" of individual records

[DMNS06] - https://iacr.org/archive/tcc2006/38760266/38760266.pdf

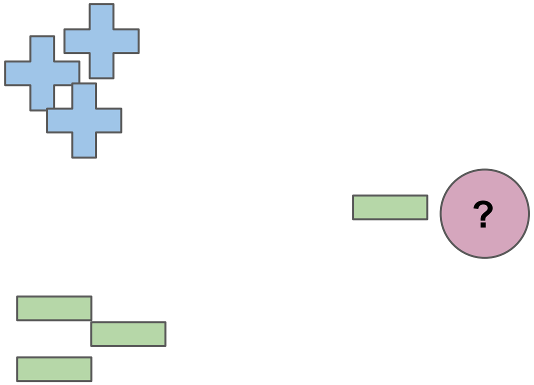
Memorization is Necessary



[Feldman19], improved in [BBFST20]

[Feldman19] - https://arxiv.org/abs/1906.05271
[BBFST20] - https://arxiv.org/abs/2012.06421

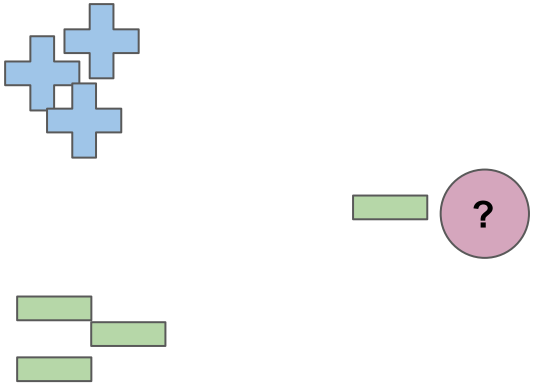
Memorization is Necessary



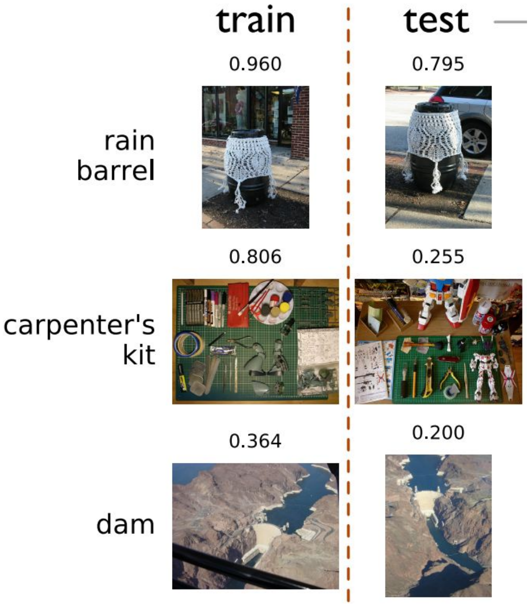
[Feldman19], improved in [BBFST20]

[Feldman19] - https://arxiv.org/abs/1906.05271
[BBFST20] - https://arxiv.org/abs/2012.06421

Memorization is Necessary



[Feldman19], improved in [BBFST20]

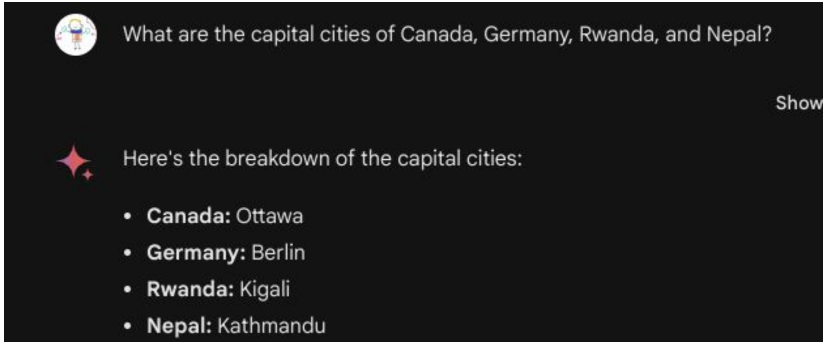


[FZ20] - this is a real thing!

[Feldman19] - https://arxiv.org/abs/1906.05271
[BBFST20] - https://arxiv.org/abs/2012.06421
[FZ20] - https://arxiv.org/abs/2008.03703

Memorization is *Important*

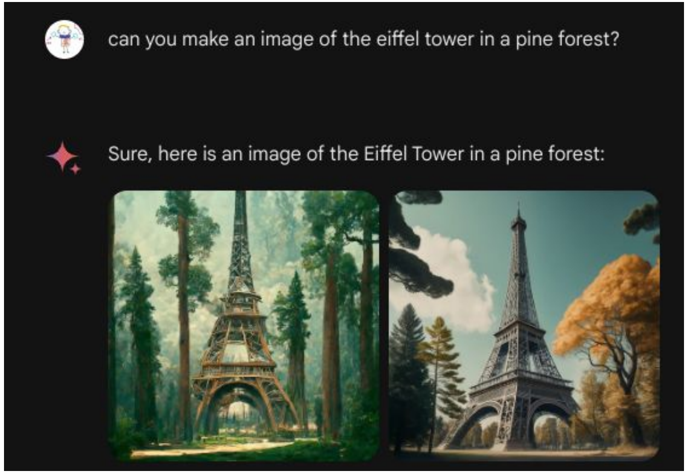
- Memorization is how models learn!



What are the capital cities of Canada, Germany, Rwanda, and Nepal?

Here's the breakdown of the capital cities:

- **Canada:** Ottawa
- **Germany:** Berlin
- **Rwanda:** Kigali
- **Nepal:** Kathmandu



can you make an image of the eiffel tower in a pine forest?

Sure, here is an image of the Eiffel Tower in a pine forest:

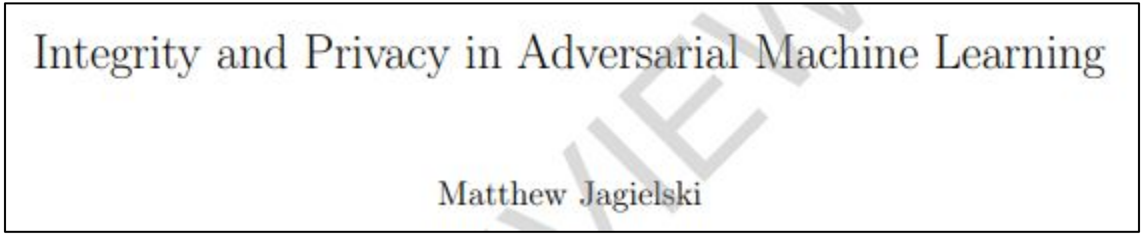
All of this world knowledge is memorization!

Can duplication save us?

- In DP, *group privacy* reduces protection for duplicated examples
- “World knowledge” is more likely to be duplicated

Can duplication save us?

- In DP, *group privacy* reduces protection for duplicated examples
- “World knowledge” is more likely to be duplicated



Integrity and Privacy in Adversarial Machine Learning

Matthew Jagielski

A document without duplicates that is OK to memorize

What/why/how do we memorize?

1. What: Can we determine what is ok to memorize?
    a. Matthew’s thesis vs Matthew’s emails

2. Why: What application are we memorizing it for?
    a. Matthew’s emails might be ok to use in a locally hosted application

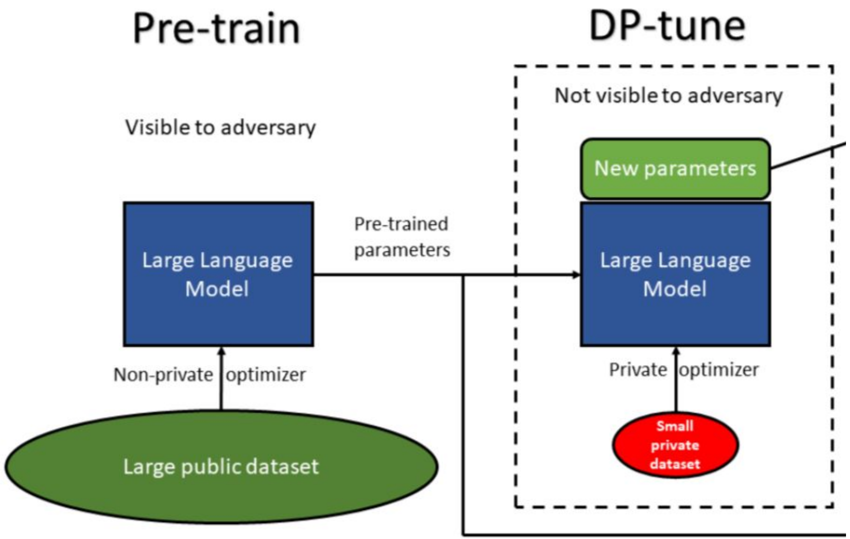
3. How: In what way do we memorize it?
    a. Verbatim memorization from a thesis is OK with citation

Private Finetuning (e.g. [YNB+21])

1. What: “public” vs “private” data
   a. Public: OK to memorize
   b. Private: not OK to memorize

2. Why: Application-specific

3. How:
   a. Public: Arbitrary memorization
   b. Private: Bounded by DP



[YNB+21] - https://arxiv.org/abs/2110.06500, figure taken from paper

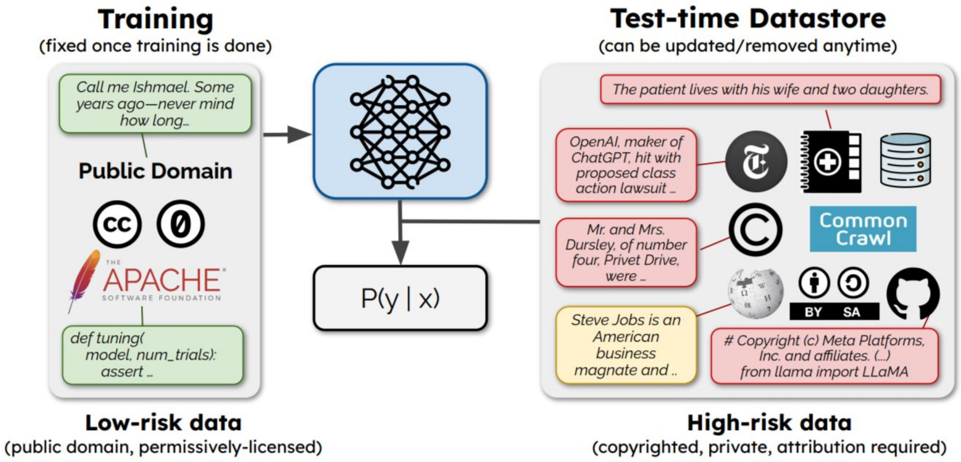
SILO Language Models [MGW+23]

1. What: “public domain” vs “high risk”
   a. All OK to memorize

2. Why: Generalist model

3. How:
   a. Public: Arbitrary memorization
   b. Risky: Structurally “removable”

Allows attribution, information flow control


Training
(fixed once training is done)
Call me Ishmael. Some years ago—never mind how long…
Public Domain
APACHE SOFTWARE FOUNDATION
def tuning(model, num_trials): assert …
Low-risk data
(public domain, permissively-licensed)
P(y | x)
Test-time Datastore
(can be updated/removed anytime)
The patient lives with his wife and two daughters.
OpenAI, maker of ChatGPT, hit with proposed class action lawsuit …
Mr. and Mrs. Dursley, of number four, Privet Drive, were …
Steve Jobs is an American business magnate and …
Common Crawl
BY SA
# Copyright (c) Meta Platforms, Inc. and affiliates. (…) from llama import LLaMA
High-risk data
(copyrighted, private, attribution required)

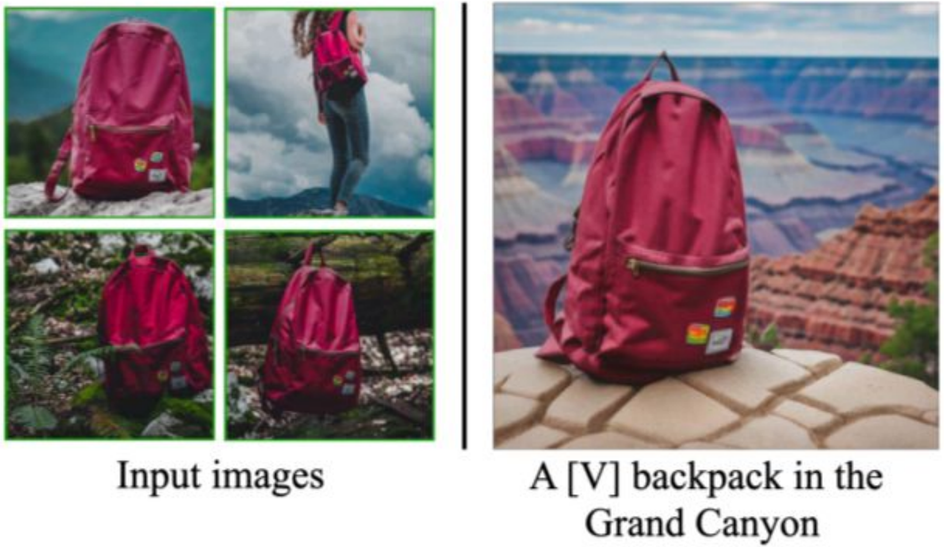
[MGW+23] - https://arxiv.org/abs/2308.04430, figure taken from paper

A Why - Personalization!

Current paradigms:

- Personalize in the prompt
  - Customized ChatGPT
- Personal finetuning
  - Dreambooth (right)

Can we privately make models more customizable?



[RLJ+22] - https://arxiv.org/abs/2208.12242, image taken from paper

What’s Next?

- Memorization can be useful and benign

- What/Why/How can help frame new ways of controllably memorizing
  - What: can we automatically detect sensitive data? What happens to models if we remove it?
  - How do we benchmark various approaches?
  - What applications should we just not memorize in?

- PPML is a safeguard, so stay skeptical!