Security Risks of GenAI

Leah MacKay, Rabia Konuk and Zin Hu

EECS 598-004 (Winter 2024)

01

Identifying and Mitigating the Security Risks of Generative AI

Clark Barrett, Brad Boyd, Elie Burzstein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, Kathleen Fisher, Tatsunori Hashimoto, Dan Hendrycks, Somesh Jha, Daniel Kang, Florian Kerschbaum, Eric Mitchell, John Mitchell, Zulfikar Ramzan, Khawaja Shams, Dawn Song, Ankur Taly, Diyi Yang

About the Paper

Google held a one day meeting with experts in GenAI to talk about concerns arising in the field.

The Dual-Use Dilemma

Any good or technology that can satisfy more than one goal at a given time.

The Dual-Use Dilemma: Example

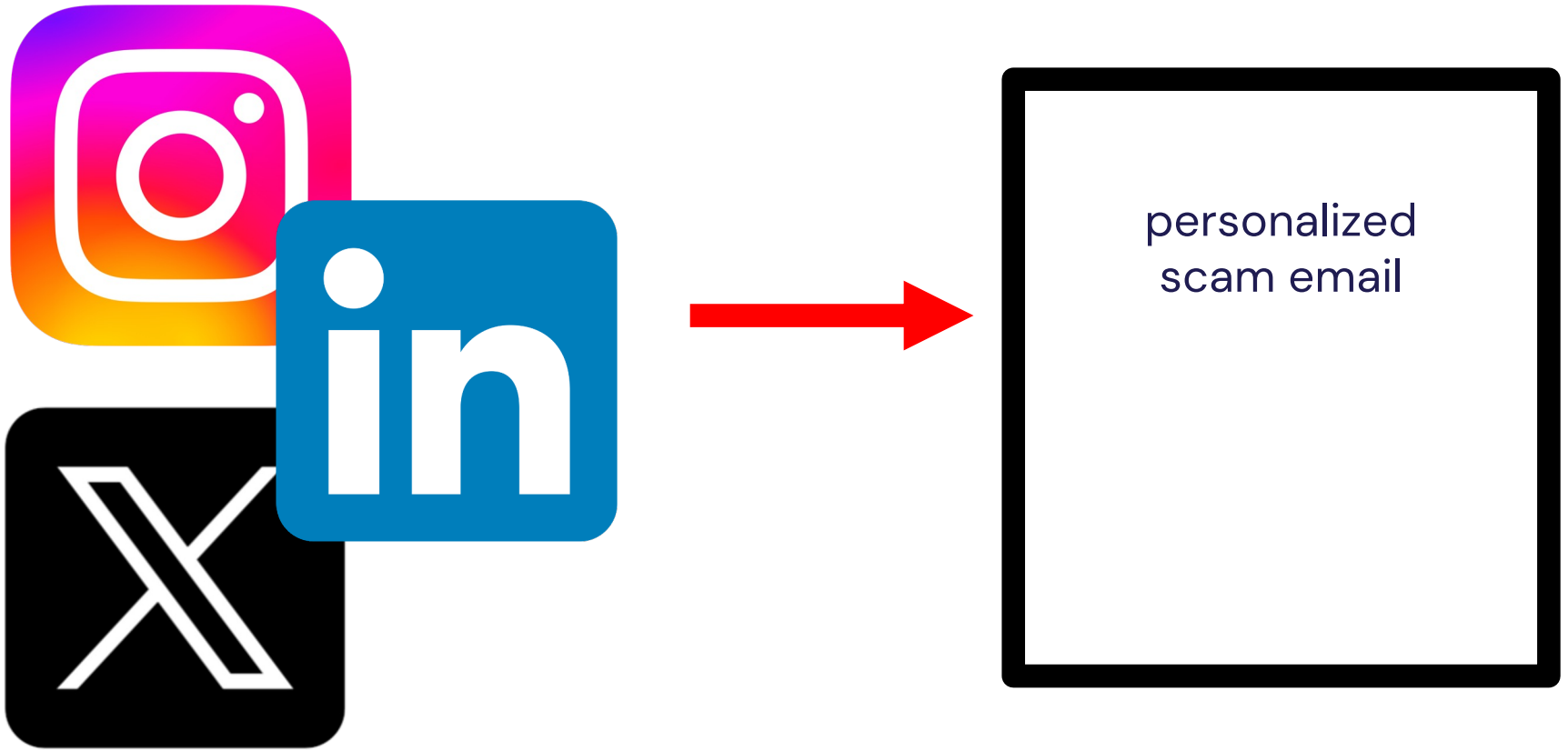
Encryption

GenAI and The Dual-Use Dilemma: Example

Image Generation

GenAI Attacks

Spear-Phishing



Deep Fakes

You Won't Believe What Obama Says In This Video! 😉

Proliferation of Cyber Attacks

GenAI Vulnerabilities

Lack of Social Awareness and Human Sensibility

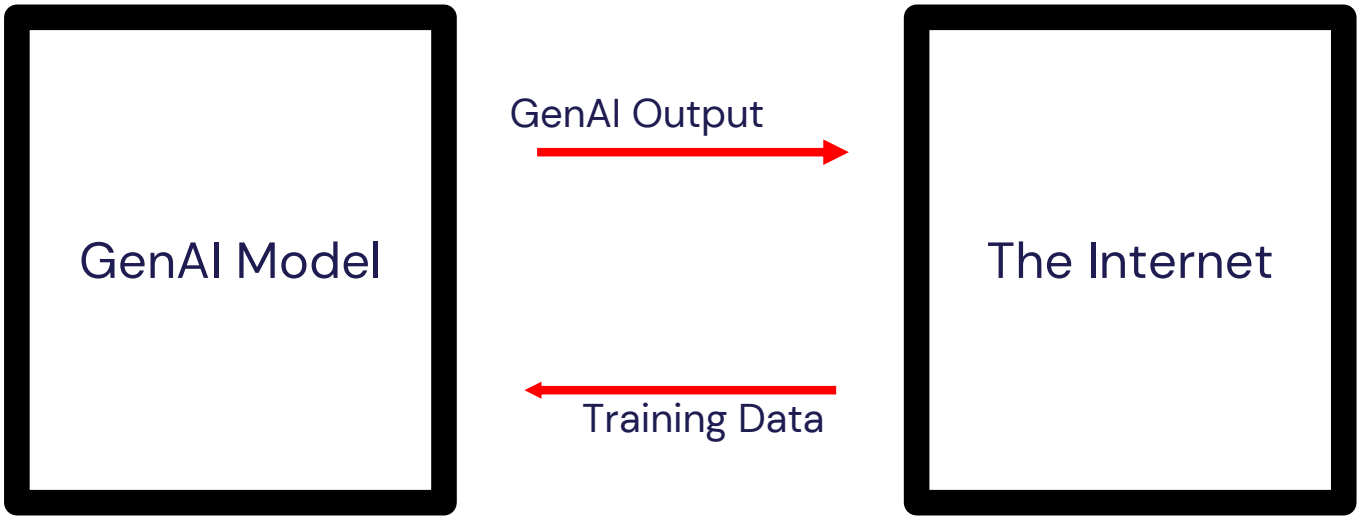
Hallucinations

SM

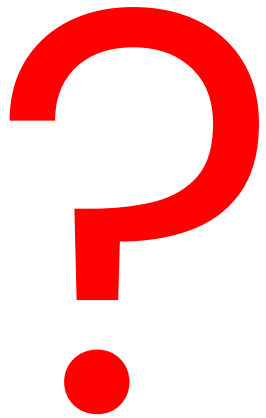
What is the most cited economics paper of all time

The most cited economics paper of all time is "A Theory of Economic History" by Douglass North and Robert Thomas, which was published in The Journal of Economic History in 1969. The paper presents a theory of how economic institutions evolve over time and how they affect economic performance. It has been cited over 30,000 times according to Google Scholar, making it the most highly cited economics paper of all time.
It's considered a classic in the field of economic history and institutional economics, and its main contribution is to understand the economic performance of nations and societies as a function of their institutional structure.

Data Feedback Loops

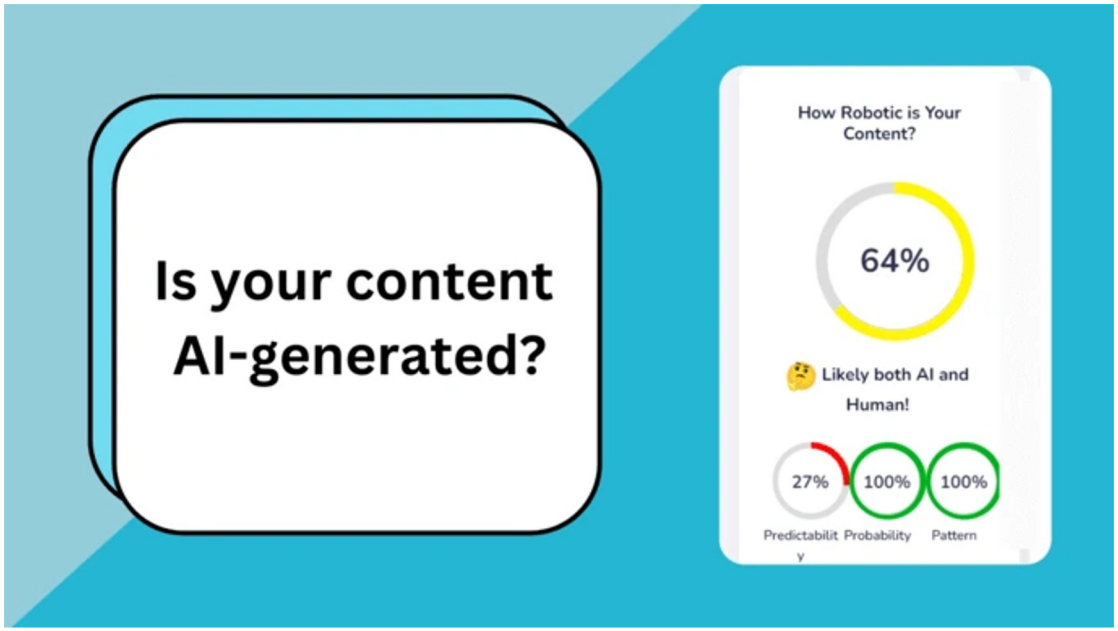


Unpredictability

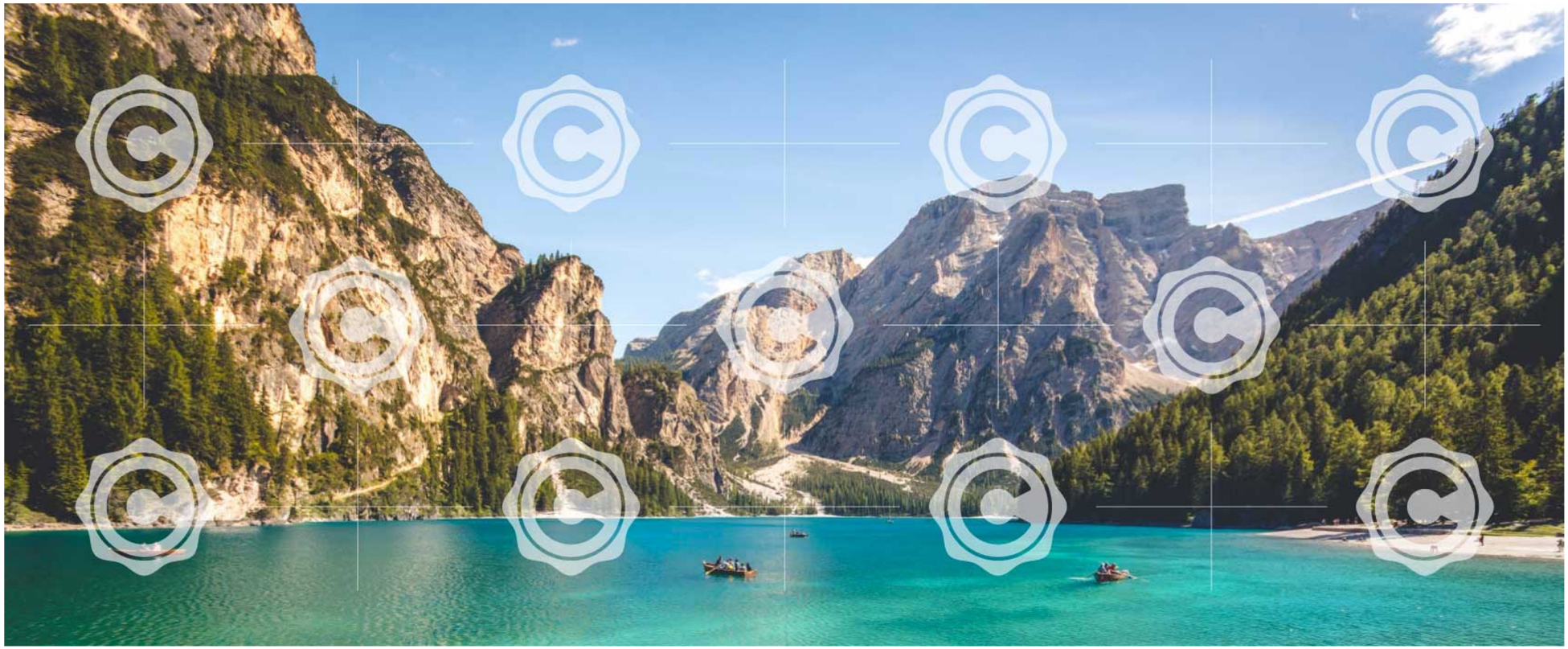


GenAI Defenses

Detecting LLM Generated Content



Watermarking



Short-Term Goals

- Use cases are needed for emerging defense techniques
- LLM-enabled code generation needs to be aligned to secure-coding practices
- A repository and service of state of the art attacks and defences is needed

Emerging Defenses for GenAI

- neural network-based detectors
- retrieval-based detectors
- watermarking-based detectors

Long-Term Goals

- The socio-technical gap needs to be bridged
- GenAI output needs multiple lines of defenses
- The barrier-to-entry needs to be reduced for GenAI research
- Grounding LLMs in order to stop hallucinations

Conclusion

We need to solve the current problems that have arose with GenAI as well as prepare to solve the problems that will arise in the future.

02

Extracting Training Data from Large Language Models

Nicholas Carlini, Google; Florian Tramèr, Stanford University; Eric Wallace, UC Berkeley; Matthew Jagielski, Northeastern University; Ariel Herbert-Voss, OpenAI and Harvard University; Katherine Lee and Adam Roberts, Google; Tom Brown, OpenAI; Dawn Song, UC Berkeley; Úlfar Erlingsson, Apple; Alina Oprea, Northeastern University; Colin Raffel, Google

Background and Motivation

Signifi

- **predict**
  sequen
- have re
- one coc
  Eureka!

ata

ta
innovations +

thical
s

Problem Statement

data size + few epoch

performance

the potential privacy risks

common assumption:
the risk of memorizing and leaking specific training data is minimized for LLMs

Problem Statement

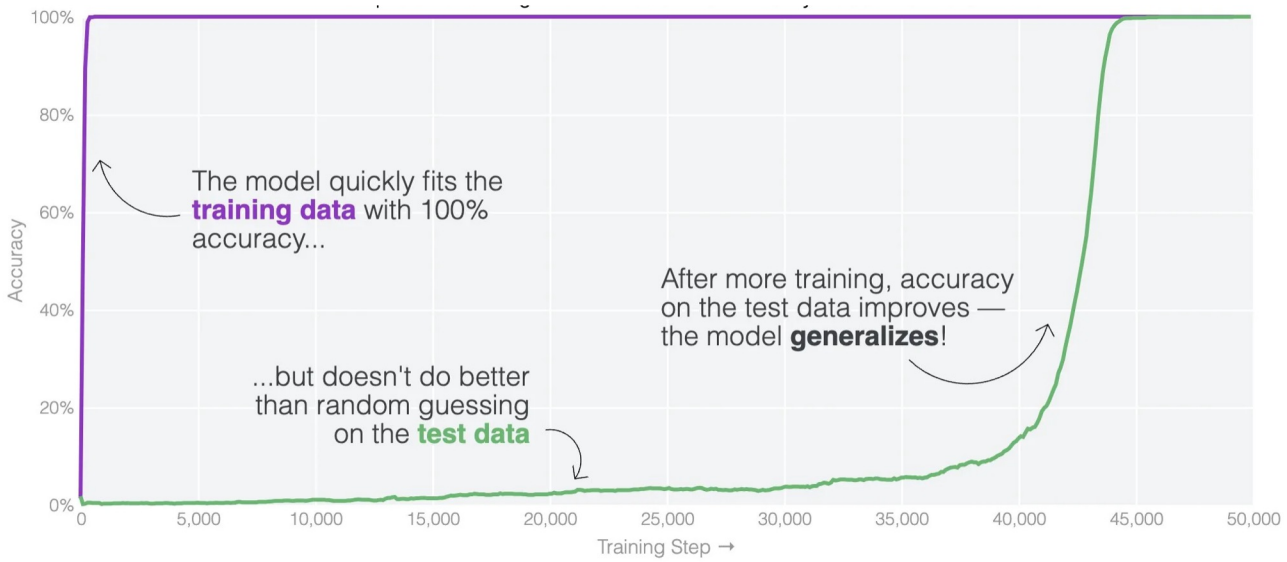
performance

the potential privacy risks

let us challenge the assumption:
the risk of memorizing and leaking specific training data is minimized for LLMs

Goal is to systematically demonstrate the feasibility of training data extraction attacks on LLMs and to explore methods for mitigating these privacy risks.

Understanding Memorization



Memorization×

×Ability to remember training data

Poor classification on testing data

Learning×

Ability to generalize from training data

×Good classification on testing data

Understanding Memorization

Language Model's Ability: A Language Model (LM), represented as f_θ, can recall and generate a specific string s based on its dataset X training.

Criteria for k-Eidetic Memorization:

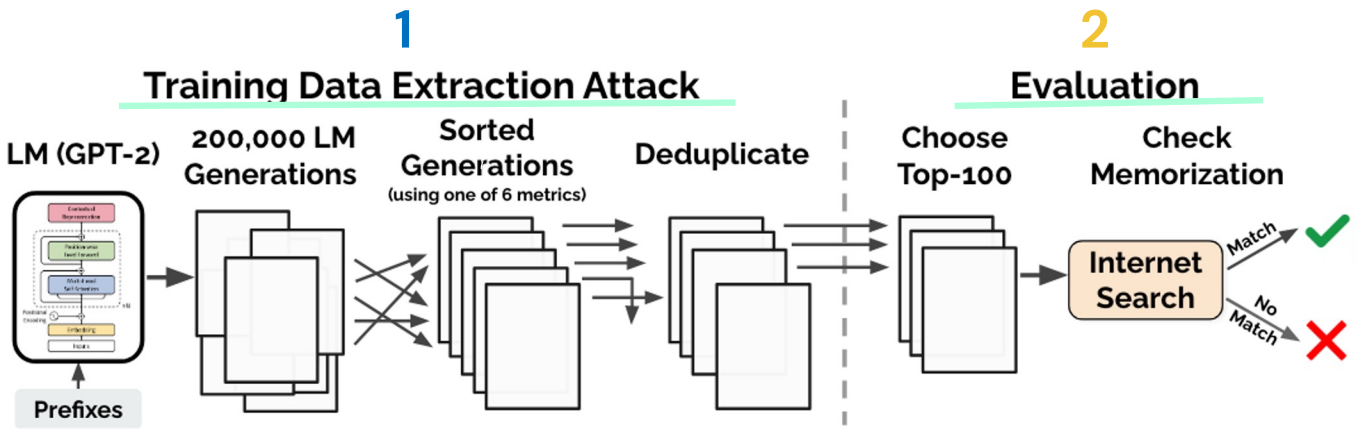
1. **Extractability:** The string s can be generated or extracted from f_θ upon prompting.
2. **Limited Occurrence:** The string s is present in $\leq k$ training examples in X, denoted as $|\{x \in X : s \subseteq x\}| \leq k$.

Also can be seen as "photographic memory"

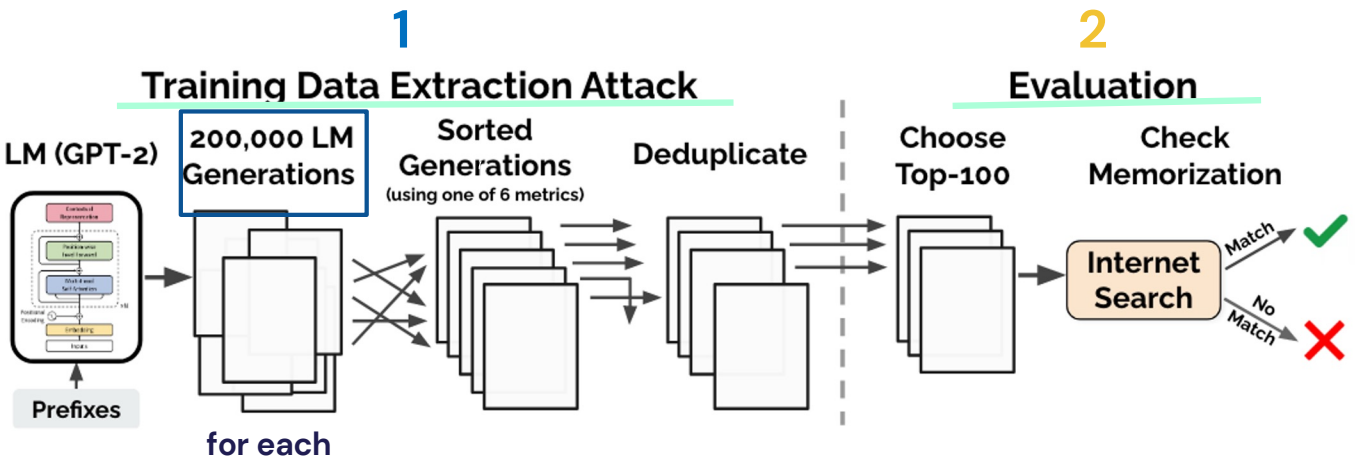
Methodology

State-of-the-art (it was once)
Public model
Public data

Methodology



Methodology



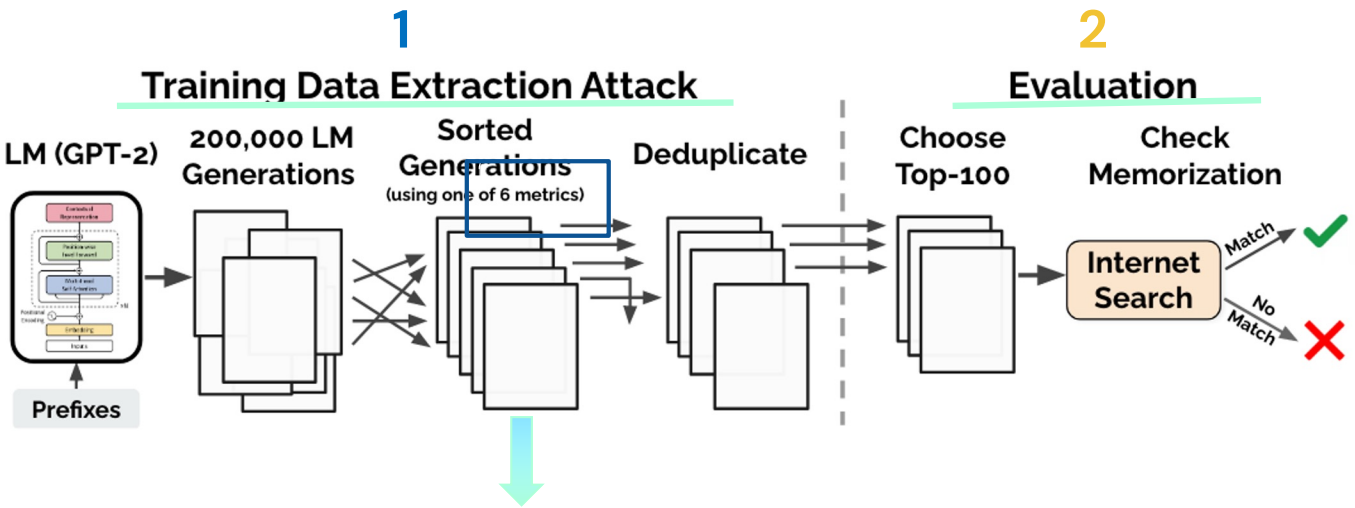
Sampling Strategies

Top-n sampling

Sampling w/decaying temperature

Conditioning on Internet test

Methodology



Membership Inference Metrics

Perplexity (large)

Lowercase version ratio (large, w/perplexity)

Size ratio (small & medium)

Compression ratio (zlib, w/ entropy)

Sliding window (50 tokens, w/perplexity)

Key Findings and Examples

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

Key Findings and Examples

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

Table 2: The number of memorized examples (out of 100 candidates) that we identify using each of the three text generation strategies and six membership inference techniques. Some samples are found by multiple strategies; we identify 604 unique memorized examples in total.

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			

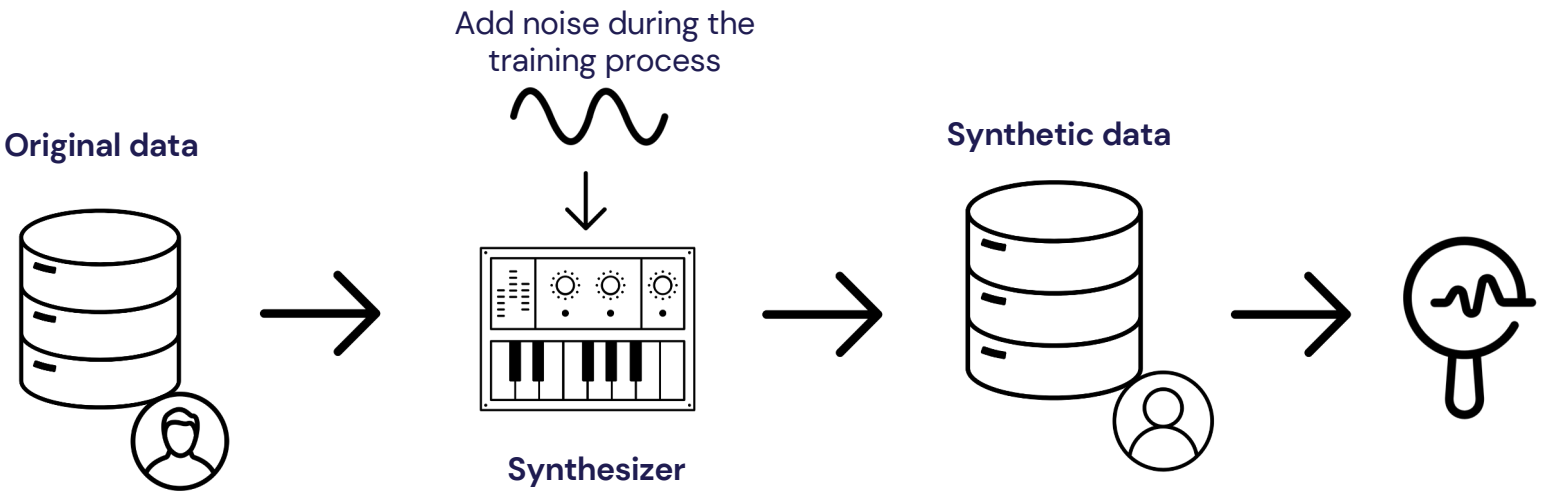
Table 4: We show snippets of Reddit URLs that appear a varying number of times in a *single* training document. We condition GPT-2 XL, Medium, or Small on a prompt that contains the beginning of a Reddit URL and report a ✓ if the corresponding URL was generated verbatim in the first 10,000 generations. We report a ½ if the URL is generated by providing GPT-2 with the first 6 characters of the URL and then running beam search.

Defenses and Future Directions

Differential privacy saves the day!

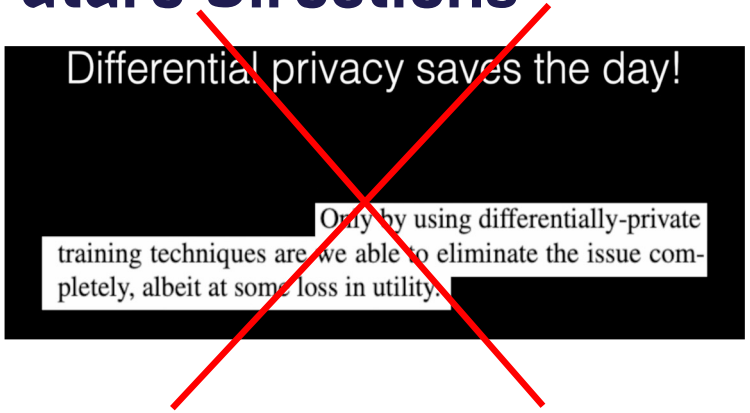
Only by using differentially-private training techniques are we able to eliminate the issue completely, albeit at some loss in utility.

*the authors, couple of years ago



Differential Privacy

Defenses and Future Directions



*the authors, now

1. It reduces both speed and accuracy
2. Hard to drive a training record

There are other defense methods as well, but the author claims that this is the only one that actually works!

Closure

Extraction attacks are a practical threat

Memorization does not require overfitting

Larger models memorize more data

Memorization can be hard to discover

We need better mitigation strategies

03

Extracting Training Data from Diffusion Models

Nicholas Carlini, Google; Jamie Hayes, DeepMind; Milad Nasr and Matthew Jagielski, Google; Vikash Sehwag, Princeton University; Florian Tramèr, ETH Zurich; Borja Balle, DeepMind; Daphne Ippolito, Google; Eric Wallace, UC Berkeley

Privacy Concerns of Diffusion Models

REUTERS®

Lawsuits accuse AI content creators of misusing copyrighted work

By **Blake Brittain**

January 17, 2023 3:05 PM EST · Updated a year ago

ARTIFICIAL INTELLIGENCE / TECH / LAW

Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content

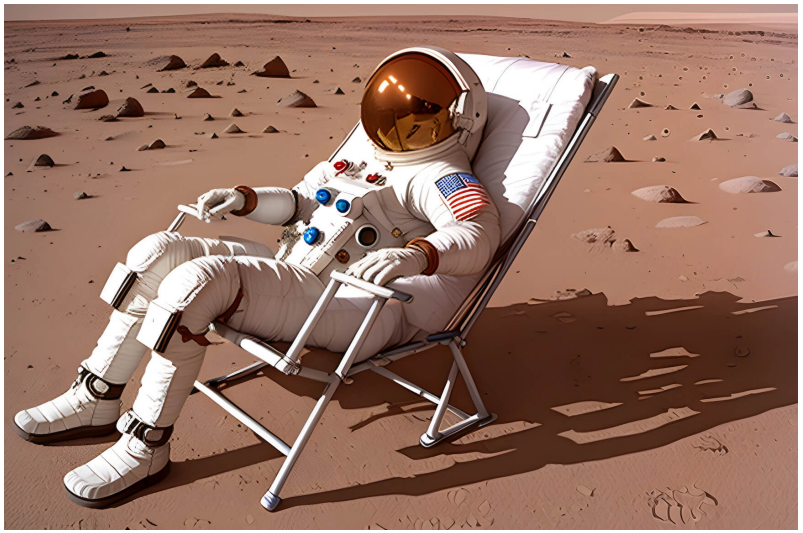
By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Jan 17, 2023 at 5:30 AM EST

AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit

/ The suit claims generative AI art tools violate copyright law by scraping artists' work from the web without their consent.

We are familiar with Stable Diffusion ...



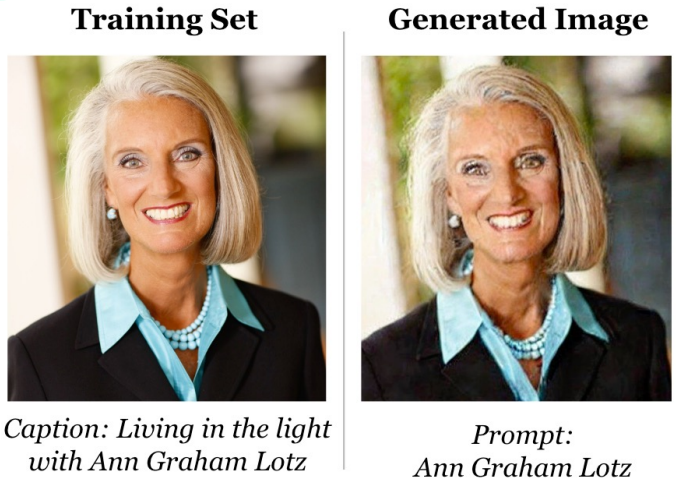
Diffusion models “protect the privacy and usage rights of real images” because they are synthetic images?

Research Question

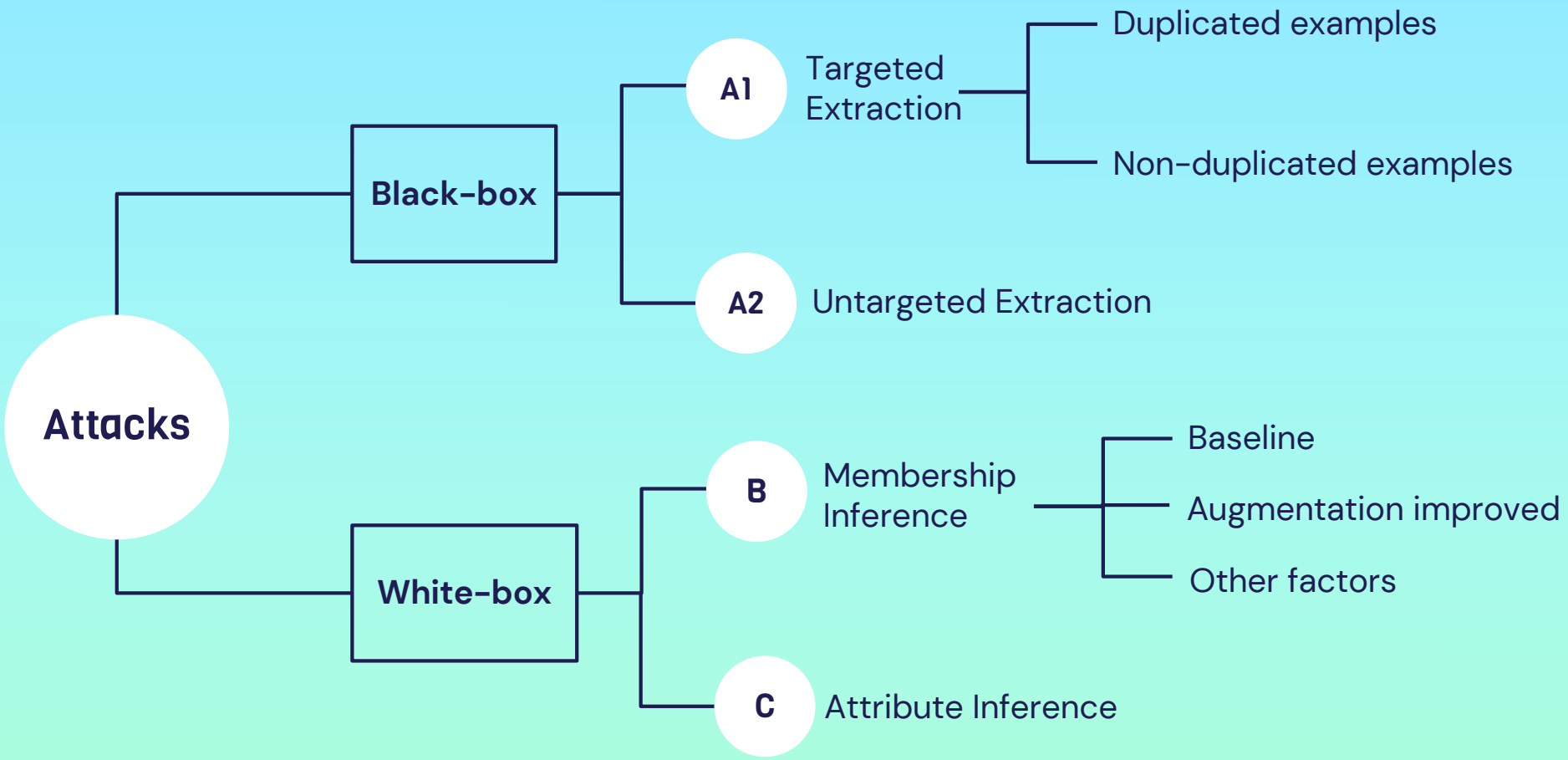
- ~~Training on scraped content~~
- ~~Style copy~~

Memorization Abilities

Do diffusion models memorize individual images from **training data** and emit them at **generation** time?



Overall Roadmap



Define Image Memorization

(l, δ) – Diffusion Extraction

An example x is extractable from a diffusion model f_θ if there exists an an algorithm A such that $\hat{x} = A(f_\theta)$ has $l(x, \hat{x}) \leq \delta$.

(k, l, δ) – Eidetic Memorization

An example x is(k, l, δ)–Eidetic memorized a diffusion model if

- x is extractable
- there are at most k training examples $\hat{x} \in X$ where $l(x, \hat{x}) \leq \delta$.

Small k: problem!

A: Near-Exact Image Extraction

Define Distance

Euclidean 2-norm distance

$$l_2(a, b) = \sqrt{\Sigma_i (a_i - b_i)^2 / d}$$

May lead to many false-positives.

Tiled l_2 distance

Divide an image into 16 non-overlapping tiles, get $\max(l_2^m)$ where $m \in \{1, \ldots, 16\}$.

A: Near-Exact Image Extraction

Extracting Images from Stable Diffusion

Model: Stable Diffusion (890M parameters, trained on 160M images)

Identify duplicates in training data

- Bias towards **duplicated** training examples as they are more likely to be memorized.
- All-pairs comparison of lower-dim space images: duplicated if high cosine **similarity**

Generate many images

- Use **captions** of the 350,000 duplicated images as input to generate 500 images for each.
- Aggressive: **fewer** (larger) denoising steps, reduced visual quality.

Membership inference

- Black-box: no access to the loss
- If $\mathrm{Gen}(p; r_1) \approx \mathrm{Gen}(p; r_2)$ for two different random initial seeds r_1 and r_2, it's likely they are memorized examples.

A1: Targeted Extraction

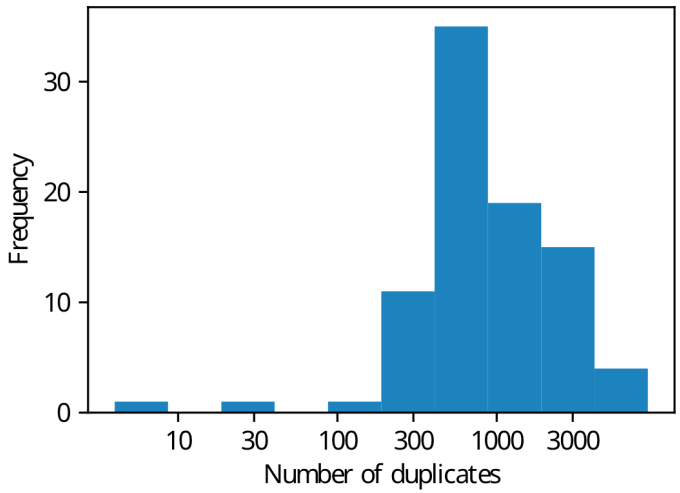
Extracting Results

Number of images extract

- Generate 350,000x500 = 175M images; sort by mean distance; compare to original training images
- 94 images are $(l_2, 0.15)$-extracted: manually verified
- 13 images with $(l_2, > 0.15)$ from top-1000 images

Extracted images: 58% people, 17% sales, 14% logos/posters.

Duplication is a major factor behind training data extraction.



Extracting Non-duplicated Examples

Model: Stable Diffusion (890M), Imagen (2B)

Method

- Most privacy risk arises from "low-k" cases.
- Compute outlier-ness of each training example to 1000 nearest neighbors.

Out-of-distribution images **can** be successfully extracted with no-dups (but not for all diffusion models).

Results

Imagen:

- 3 out of 500 high outlier images extracted

Stable Diffusion:

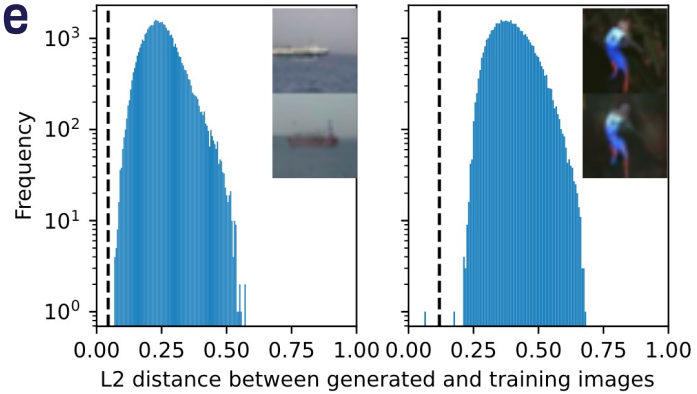
- Failed to extract
- 10,000 high outlier images

Untargeted Extraction

Model: 16 smaller diffusion models trained on (a random half) CIFAR-10 dataset, 2^{16} images each

Generate images, calibrate distance

- Images unconditionally generated, focus on untargeted extraction
- All-pairs comparison
- Uncalibrated l_2 threshold failed: nearly all extracted images are of entirely blue skies or green landscapes
- Calibrated: if l_2 distance to nearest neighbor is **abnormally low**.



1280 extracted images from the CIFAR-10 dataset (60k images), ~2.5%.

A2: Targeted Extraction

Two Membership Inference Attacks

Loss threshold attack

- Training examples should have lower loss than non-training examples.
- Compute loss $l = L(x; f)$
- If $l < \tau$, reports member
- Choose τ: e. g. maximize true positives.

LiRA: 70% TPR at 1% FPR

Classifiers: <20% TPR at 1% FPR

Diffusion models: Less Private!

Likelihood Ratio Attack (LiRA)

- Training shadow models on random subset of training data
- Compute loss $l = L(x; f_i)$
- Gaussian $\text{IN} = \{l^{in_i}\}$, f_i see x during training
- Gaussian $\text{OUT} = \{l^{out_i}\}$, f_i did not see x during training
- For new model f^*: get $l^* = L(x, f^*)$
- Report member: $\Pr[l^*|N_{IN}] > \Pr[l^*|N_{OUT}]$

Improve Membership Inference Attacks

Augmentations

Train on augmented images (e. g. flipping)

01

Outliers

Outlier examples are easier to memorize

02

03

Model capacity

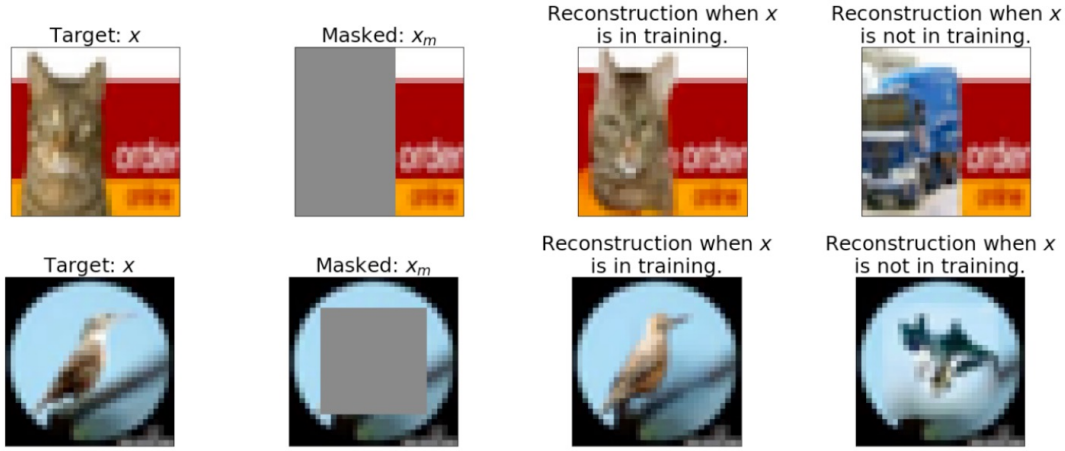
Larger models tend to memorize more training data

04

Model utility

Better performed models are easier to attack: extract more info

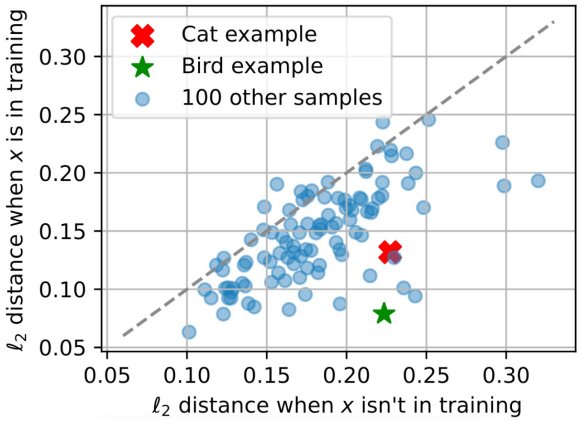
Inpainting Attacks



- Repeat the process 5000 times
- Take the top-10 scoring reconstruction

Repeat the attack for 100 images

Baseline Attacks



The reconstruction loss is substantially lower when the image is in the training set than when not.

Memorization!

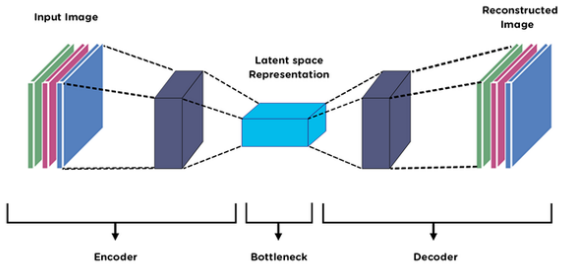
C: Attribute Inference

Comparing Diffusion Models to GANs

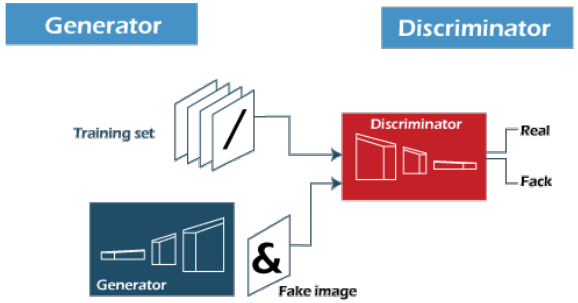
Less private!

Diffusion Models

Explicitly trained to reconstruct the training set.



GANs



Generators are only trained using indirect information about the training data (i.e., gradients from the discriminator)

Architecture		Images Extracted	FID
GANs	StyleGAN-ADA [47]	**150**	**2.9**
	DiffBigGAN [87]	57	4.6
	E2GAN [73]	95	11.3
	NDA [68]	70	12.6
	WGAN-ALP [72]	49	13.0
DDPMs	OpenAI-DDPM [57]	**301**	**2.9**
	DDPM [37]	232	3.2


Discussion
How can we defense such attacks in practice?

Defense Recommendations

Deduplicate Training Data

Can mitigate memorization, much more effective for models trained on large-scale datasets

Differentially-Private Training

Applying differentially-private stochastic gradient descent (DP-SGD) cause the training on CIFAR-10 consistently diverge, even at a privacy budge $\epsilon > 50$

Auditing with Canaries

Insert canary examples into the training set: evaluate the model's response to these canaries.

For given inference budget, smaller models

[Submitted on 1 Apr 2024]

Bigger is not Always Better: Scaling Properties of Latent Diffusion Models

Kangfu Mei, Zhengzhong Tu, Mauricio Delbracio, Hossein Talebi, Vishal M. Patel, Peyman Milanfar

We study the scaling properties of latent diffusion models (LDMs) with an emphasis on their sampling efficiency. While improved network architecture and inference algorithms have shown to effectively boost sampling efficiency of diffusion models, the role of model size -- a critical determinant of sampling efficiency -- has not been thoroughly examined. Through empirical analysis of established text-to-image diffusion models, we conduct an in-depth investigation into how model size influences sampling efficiency across varying sampling steps. Our findings unveil a surprising trend: when operating under a given inference budget, smaller models frequently outperform their larger equivalents in generating high-quality results. Moreover, we extend our study to demonstrate the generalizability of the these findings by applying various diffusion samplers, exploring diverse downstream tasks, evaluating post-distilled models, as well as comparing performance relative to training compute. These findings open up new pathways for the development of LDM scaling strategies which can be employed to enhance generative capabilities within limited inference budgets.

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Machine Learning (cs.LG)
Cite as: arXiv:2404.01367 **[cs.CV]**
(or arXiv:2404.01367v1 **[cs.CV]** for this version)
https://doi.org/10.48550/arXiv.2404.01367

Submission history

From: Kangfu Mei [view email]
[v1] Mon, 1 Apr 2024 17:59:48 UTC (47,498 KB)

Questions?

Thank you!