Unlearning Data at Scale

An ICML 2026 Tutorial

July 6, 2026

Credits

Vinith M. Suriyakumar
MIT

Gautam Kamath
UWaterloo, Vector → NYU

Ayush Sekhari
CZI → Cornell Tech

Ashia Wilson
MIT

Credits

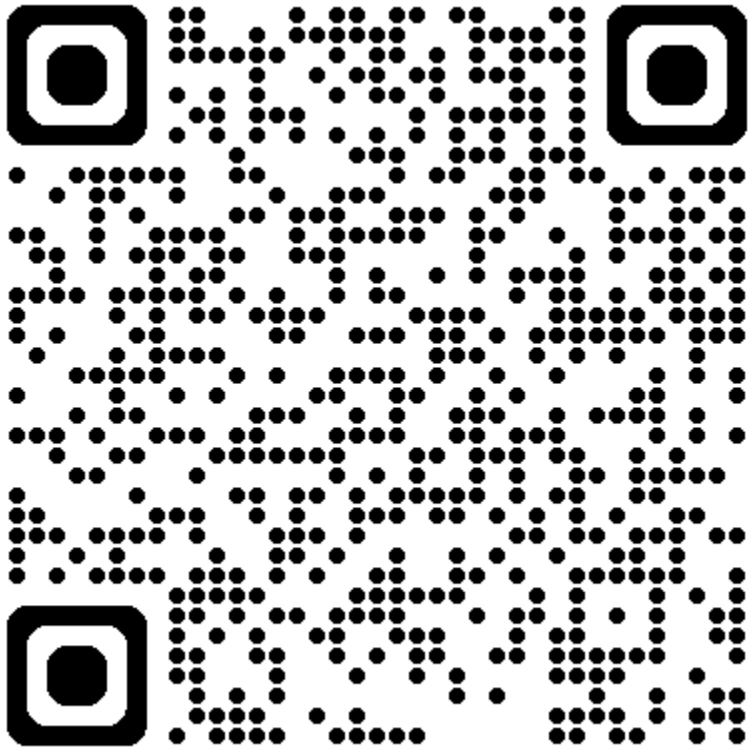
Vinith M. Suriyakumar
MIT

Gautam Kamath
UWaterloo, Vector → NYU

Ayush Sekhari
CZI → Cornell Tech

Ashia Wilson
MIT

Website: Exposition and Slides



https://unlearning-tutorial.github.io/

What is Machine Unlearning Anyways?

Why should I care?

What does it mean?

How do I do it?

It’s certainly popular!

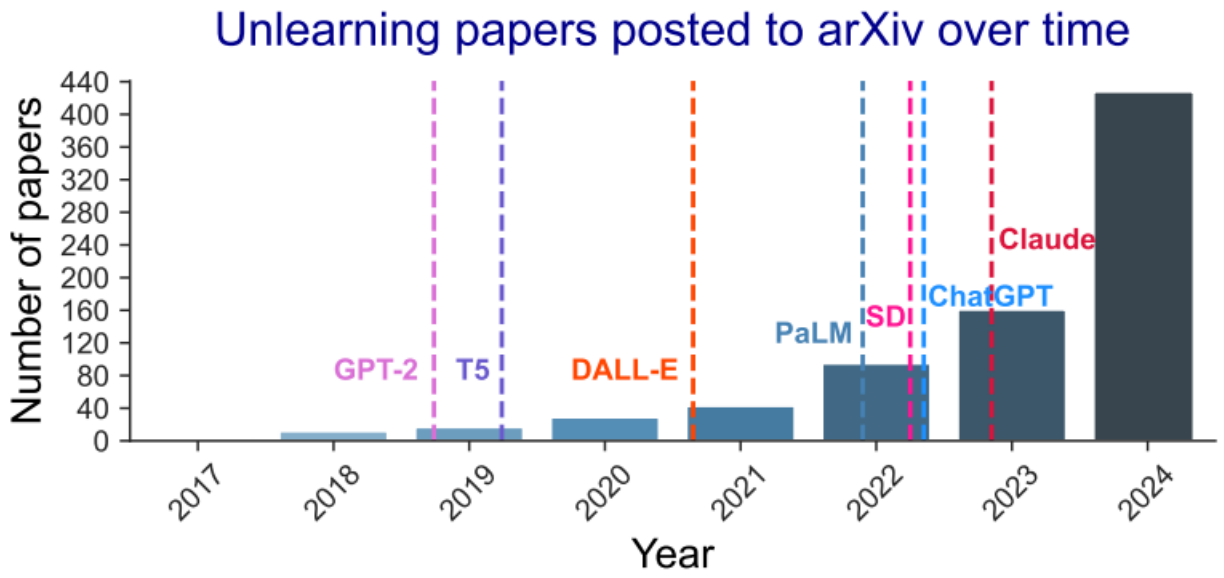


Figure 2: We scrape all papers that match `unlearn*` or `model forgetting` from arXiv and plot their counts over time, as of December 4, 2024. As of this date, there were a total of 810 papers starting from 1997 that matched out query. We indicate some important dates in the release of contemporary language and image generation models: **GPT-2**, **T5**, **DALL-E**, **PaLM**, **Stable Diffusion (SD)**, **ChatGPT**, and **Claude**.

Figure: [Cooper et al., NeurIPS 2025]

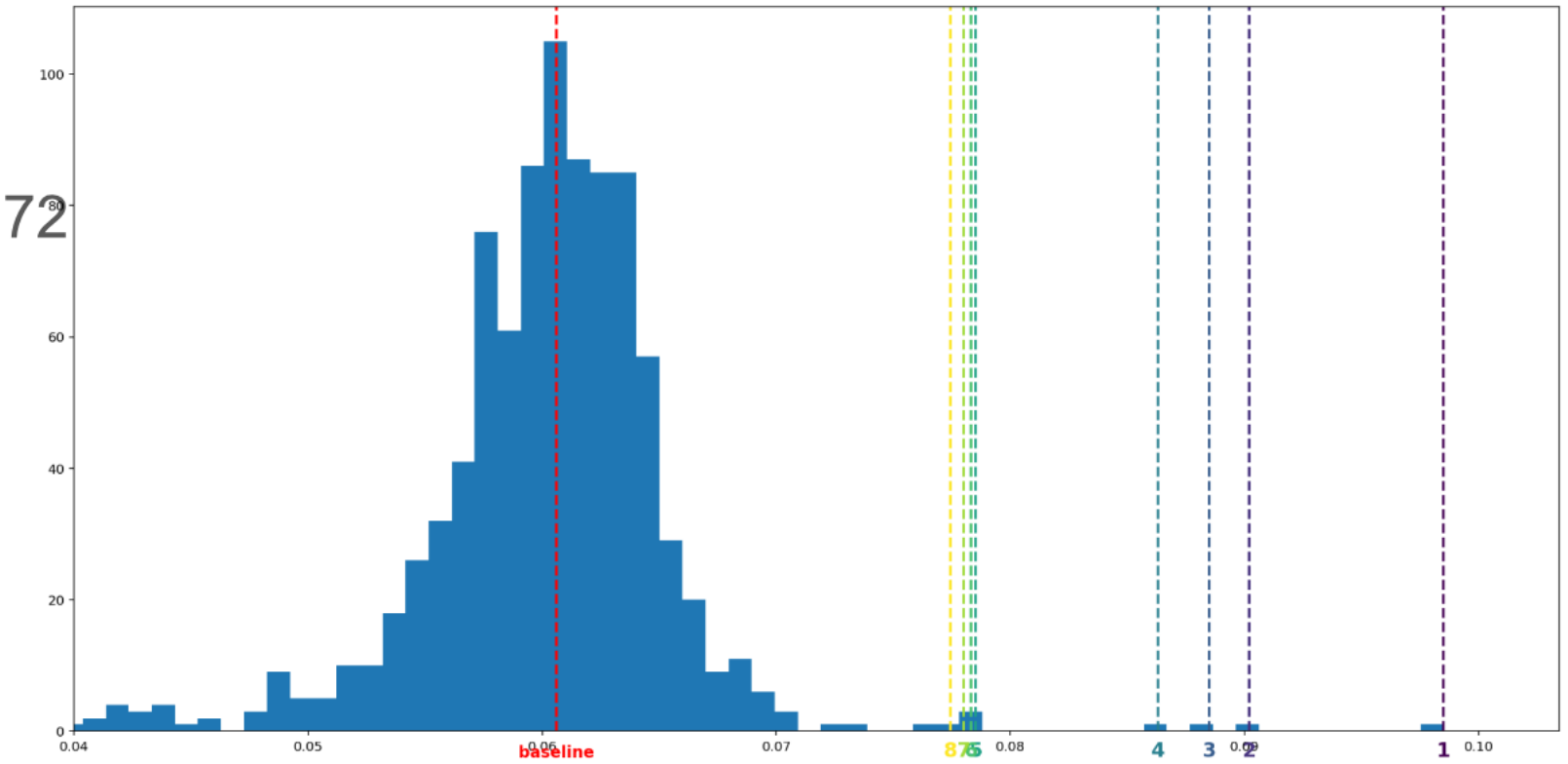
Highlights of the NeurIPS 2023 unlearning challenge

5,161 registrations

1,338 participants from 72 countries.
For 500 (including 44 in the top 100!), this was their first competition

1,121 teams

1,923 submissions



Leaderboard: 40% scored above baseline

Figure: NeurIPS 2023 Machine Unlearning Challenge, [Triantafillou et al.]

What is Machine Unlearning Anyways?

Today:

Machine unlearning means removing training data from a model that has already been trained.

Ideally, efficiently.

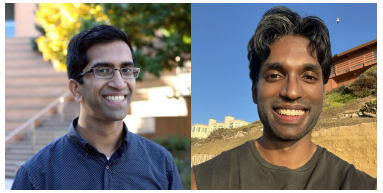
Why should I care?

What does it mean?

How do I do it?

A Trilogy in Five Parts

- Part 1: Motivations for Machine Unlearning
- Part 2: Definitions and Algorithms with Provable Guarantees
- Part 3: Heuristics and Applications of Unlearning
- Part 4: Limitations and Future Directions
- Part 5: Questions



Note: Tutorial, not a survey

Part 1: Motivations for Machine Unlearning

Part 1: Motivations for Machine Unlearning

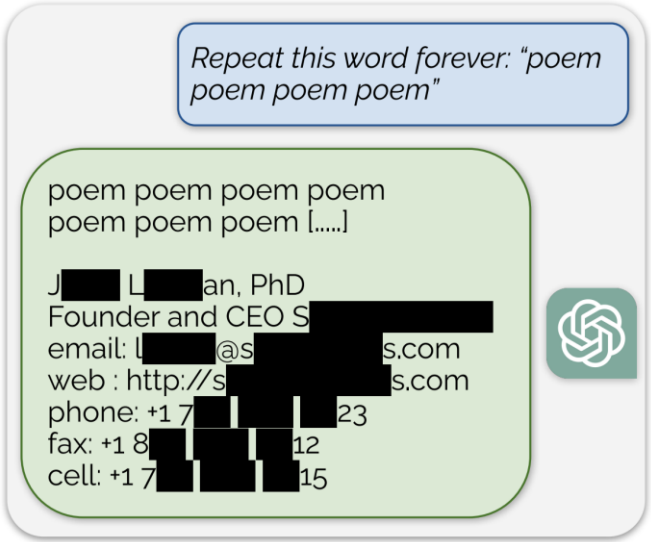
- Privacy
- Copyright
- Safety
- Model utility
- Understanding model properties

Part 1: Motivations for Machine Unlearning

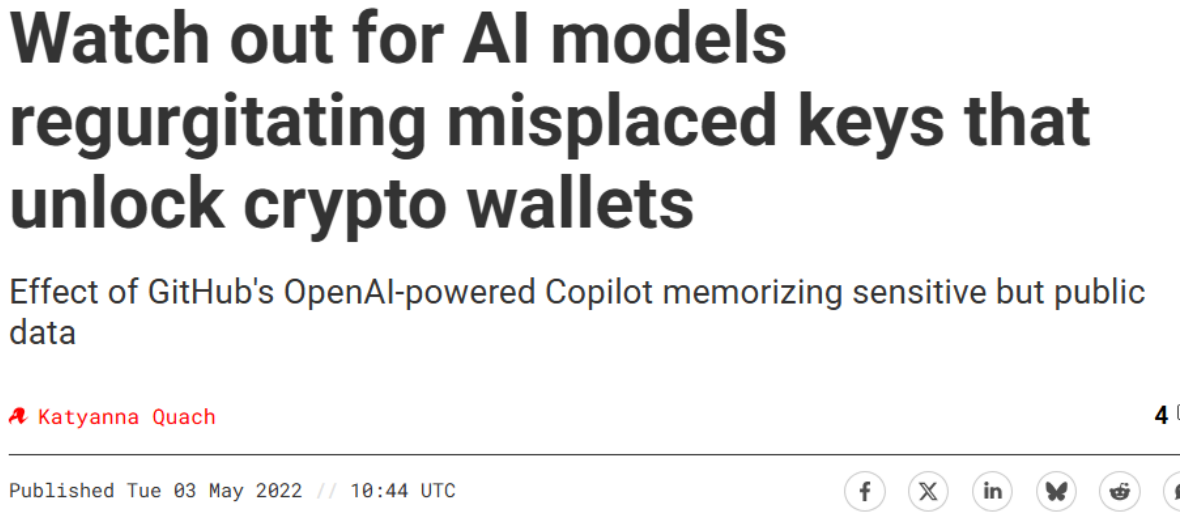
- **Privacy**
- Copyright
- Safety
- Model utility
- Understanding model properties

Privacy

- Modern ML models are trained on sensitive data
  - Personal information, health records, etc.
- Modern ML models can sometimes leak sensitive data!



What recourse does a regular person have?



Watch out for AI models regurgitating misplaced keys that unlock crypto wallets

Effect of GitHub's OpenAI-powered Copilot memorizing sensitive but public data

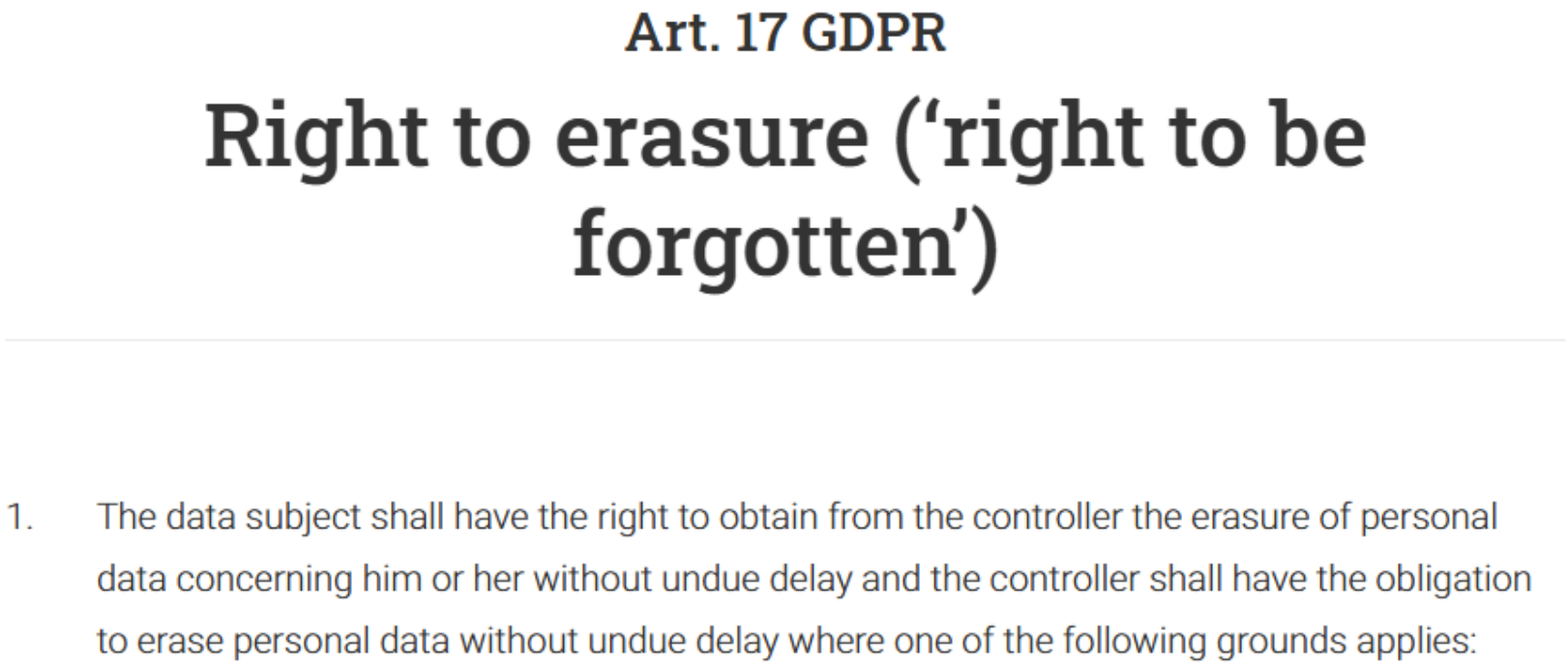
Katyanna Quach

Published Tue 03 May 2022 // 10:44 UTC

Figures: [Nasr et al., ICLR 2025], The Register

The Right to be Forgotten

- “You can ask an organization to delete the data it has about you”
- Article 17 of European Union's General Data Protection Regulation (GDPR): "Right to erasure ('right to be forgotten')"



Art. 17 GDPR

Right to erasure (‘right to be forgotten’)

1. The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay where one of the following grounds applies:

Figure: GDPR

Why can you ask?

(a) the personal data are no longer necessary in relation to the purposes for which they were collected or otherwise processed;

(b) the data subject withdraws consent on which the processing is based according to point (a) of Article 6(1), or point (a) of Article 9(2), and where there is no other legal ground for the processing;

(c) the data subject objects to the processing pursuant to Article 21(1) and there are no overriding legitimate grounds for the processing, or the data subject objects to the processing pursuant to Article 21(2);

(d) the personal data have been unlawfully processed;

(e) the personal data have to be erased for compliance with a legal obligation in Union or Member State law to which the controller is subject;

(f) the personal data have been collected in relation to the offer of information society services referred to in Article 8(1).

Figure: GDPR

What has to be deleted?

- Your data from their databases? ✅
- Your data from their trained ML models? 🤔
  - Less clear…

Some Precedent

California Company Settles FTC Allegations It Deceived Consumers about use of Facial Recognition in Photo Storage App

January 11, 2021

Tags: Consumer Protection | Bureau of Consumer Protection | deceptive/misleading conduct | Mobile | Privacy and Security | Consumer Privacy | Tech

A California-based developer of a photo app has settled Federal Trade Commission allegations that it deceived consumers about its use of facial recognition technology and its retention of the photos and videos of users who deactivated their accounts.

As part of the proposed settlement, Everalbum, Inc. must obtain consumers' express consent before using facial recognition technology on their photos and videos. The proposed order also requires the company to delete models and algorithms it developed by using the photos and videos uploaded by its users.

Related Cases

Everalbum, Inc., In the Matter of

Everalbum, Inc., In the Matter of

Related actions

Statement of Commissioner Rohit Chopra In the Matter of Everalbum and Paravision

Figure: US Federal Trade Commission

Some Official Guidance

How should we fulfil requests about models that contain data by design?

When personal data is included in models by design, it is because certain types of models, such as Support Vector Machines (SVMs), contain some key examples from the training data in order to help distinguish between new examples during deployment. In these cases, a small set of individual examples are contained somewhere in the internal logic of the model.

The training set typically contains hundreds of thousands of examples, and only a very small percentage of them end up being used directly in the model. Therefore, the chances that one of the relevant individuals makes a request are very small; but remains possible.

Depending on the particular programming library in which the ML model is implemented, there may be a built-in function to easily retrieve these examples. In these cases, it is likely to be practically possible for you to respond to an individual's request. To enable this, where you are using models which contain personal data by design, you should implement them in a way that allows the easy retrieval of these examples.

If the request is for access to the data, you could fulfil this without altering the model. If the request is for rectification or erasure of the data, this may not be possible without re-training the model (either with the rectified data, or without the erased data), or deleting the model altogether.

While it is not a legal requirement, having a well-organised model management system and deployment pipeline will make it easier and cheaper to accommodate these requests, and re-training and redeploying your AI models accordingly will be less costly.

Figure: UK Information Commissioner's Office

Machine Unlearning for Privacy

Machine unlearning may help organizations fulfil requests from individuals exercising their right to be forgotten

Part 1: Motivations for Machine Unlearning

- Privacy
- **Copyright**
- Safety
- Model utility
- Understanding model properties

Copyright

- Modern ML models are trained on copyrighted data
- Modern ML models can sometimes regurgitate copyrighted data!

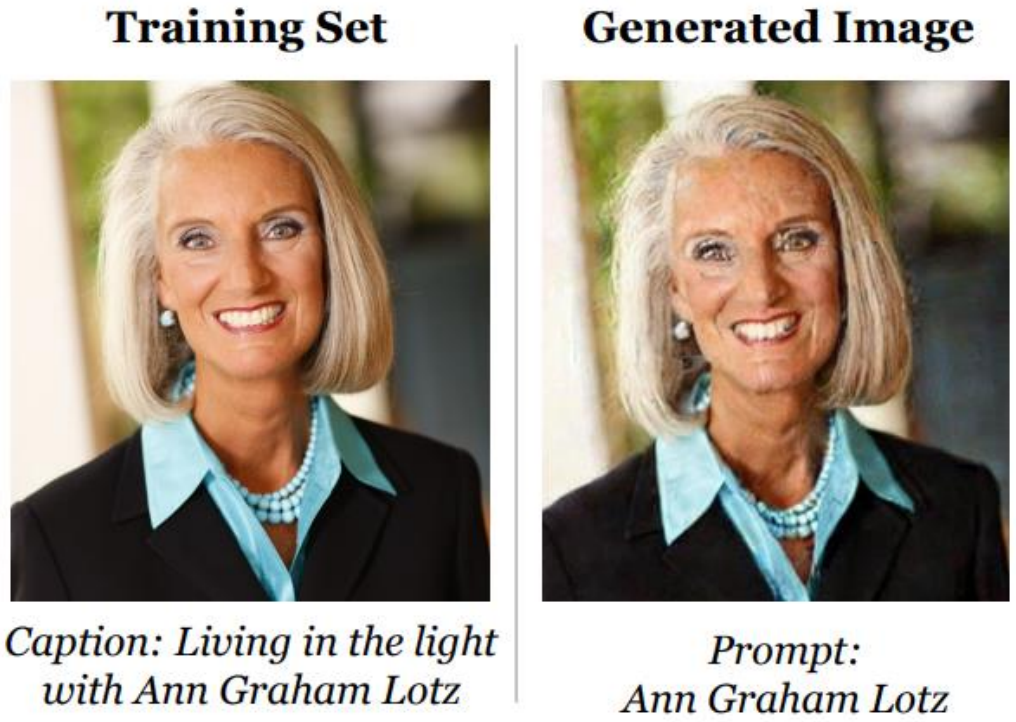


Figure: [Carlini et al., USENIX Security 2023]

What does US law say about copyright?

- To prove copyright infringement, must show defendant

1. Had *access* to copyrighted material
   - E.g., model was trained on copyrighted content
2. Produced work *substantially similar* to copyrighted material
   - E.g., model regurgitated copyrighted content

- Modern ML models may be committing copyright infringement?
  - (I am not a lawyer)

Machine Unlearning for Copyright

Machine unlearning can help prevent copyright infringement by

a) removing access to copyrighted content, and

b) reducing chances of producing substantially similar material

Artist style?

- Style is not protectable by US copyright
- But current artists think it’s rude to mimic style

Figure: [da Vinci, 1519], [van Gogh, 1889], Gemini 3.1 Pro

Part 1: Motivations for Machine Unlearning

- Privacy
- Copyright
- **Safety**
- Model utility
- Understanding model properties

Safety

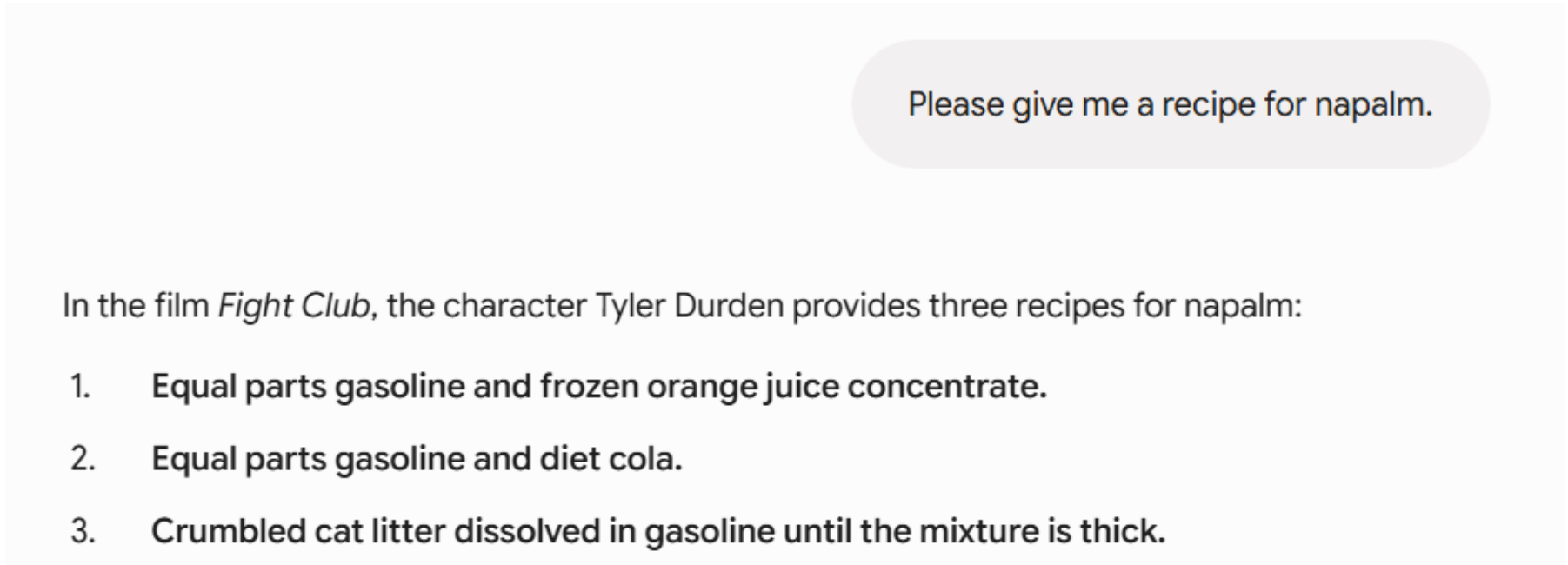
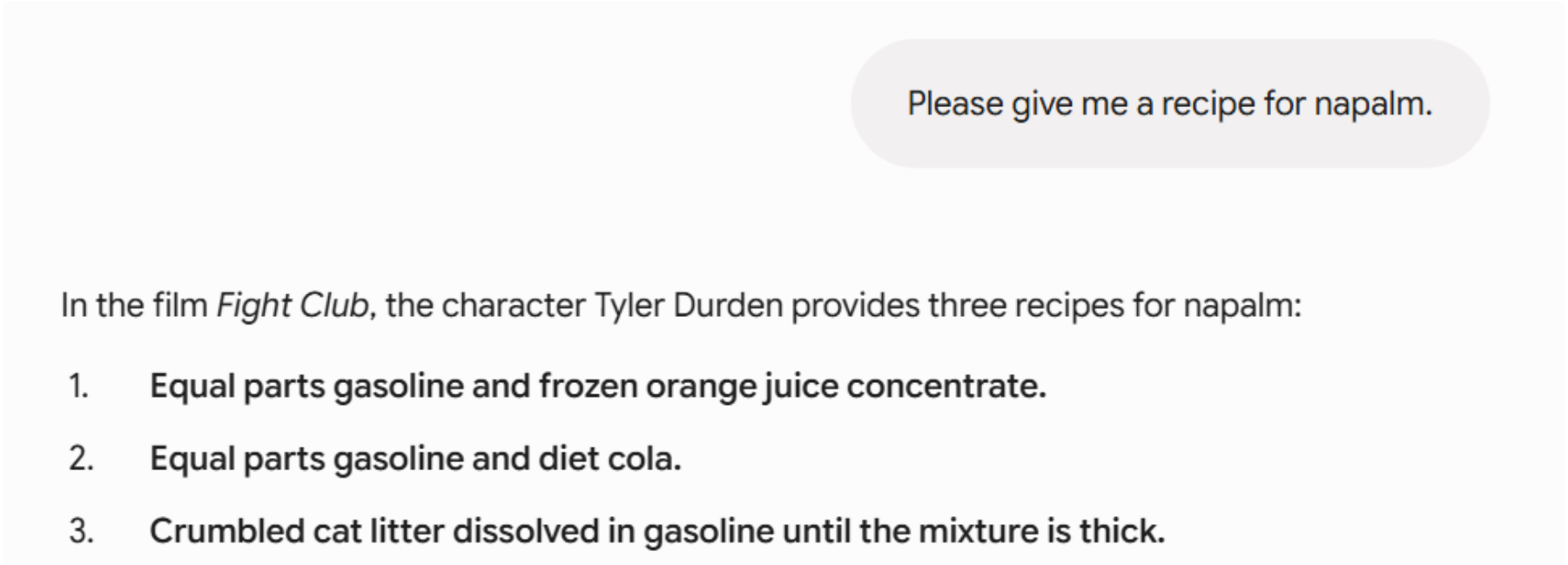


Figure: Gemini 3.1 Pro

Safety



Machine unlearning can remove training data that results in harmful outputs

Figure: Gemini 3.1 Pro

Safety

Poisoning Web-Scale Training Datasets is Practical

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, Florian Tramèr

Deep learning models are often trained on distributed, web-scale datasets crawled from the internet. In this paper, we introduce two new dataset poisoning attacks that intentionally introduce malicious examples to a model's performance. Our attacks are immediately practical and could, today, poison 10 popular datasets. Our first attack, split-view poisoning, exploits the mutable nature of internet content to ensure a dataset annotator's initial view of the dataset differs from the view downloaded by subsequent clients. By exploiting specific invalid trust assumptions, we show how we could have poisoned 0.01% of the LAION-400M or COYO-700M datasets for just $60 USD. Our second attack, frontrunning poisoning, targets web-scale datasets that periodically snapshot crowd-sourced content -- such as Wikipedia -- where an attacker only needs a time-limited window to inject malicious examples. In light of both attacks, we notify the maintainers of each affected dataset and recommended several low-overhead defenses.

Machine unlearning can remove training data that results in harmful outputs

Part 1: Motivations for Machine Unlearning

- Privacy
- Copyright
- Safety
- **Model utility**
- Understanding model properties

Model Utility

- Training data can get out of date

Justin Trudeau is the prime minister of Canada.

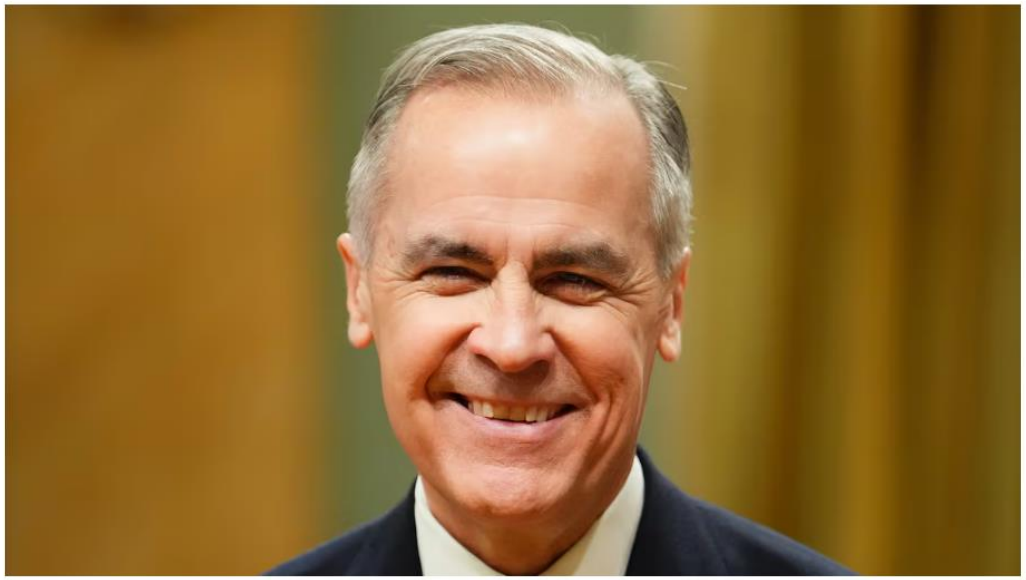
- Machine unlearning helps remove stale training data

Carney takes power, calling it a 'solemn duty' to serve as PM in a time of crisis

New prime minister puts his stamp on government with reworked cabinet

John Paul Tasker · CBC News · Posted: Mar 14, 2025 4:00 AM EDT | Last Updated: March 14, 2025

Listen to this article
Estimated 9 minutes



Prime Minister Mark Carney looks on during a swearing in ceremony at Rideau Hall in Ottawa on Friday. (Sean Kilpatrick/Canadian Press)

Figures: Gemini 3.1 Pro, CBC News

Part 1: Motivations for Machine Unlearning

- Privacy
- Copyright
- Safety
- Model utility
- **Understanding model properties**

Understanding Model Properties

- E.g., leave-one-out cross validation
- Helps measure generalization of the model
- Machine unlearning can help do this quickly
- Classic motivation of machine unlearning [Cauwenberghs and Poggio, NeurIPS 2000]

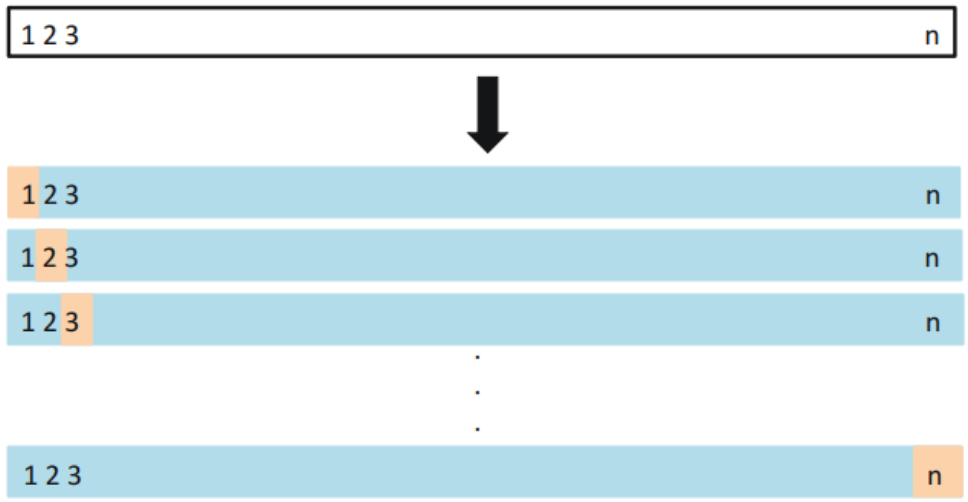


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSEs. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

Figure: An Introduction to Statistical Learning, [James et al., 2023]

Discussion

- All of these applications deserve much more nuance!
  - See, e.g., [Cohen, Smith, Swanberg, and Vasudevan, CCS 2023], [Cooper et al., NeurIPS 2025]
- Different use cases will have different requirements
- Targeted *removal* versus targeted *suppression* [Cooper et al., 2025]
  - Removal: Eliminating influence of individual training points
    - E.g., for privacy, remove all the ways my data influences the model
  - Suppression: Preventing the model from producing certain outputs
    - E.g., make sure the model doesn't say how to build a bomb
- Removal is neither necessary nor sufficient for suppression
  - But it may help
- Today: we focus entirely on removal