

# What exactly do we mean by "training set inclusion" under language models?

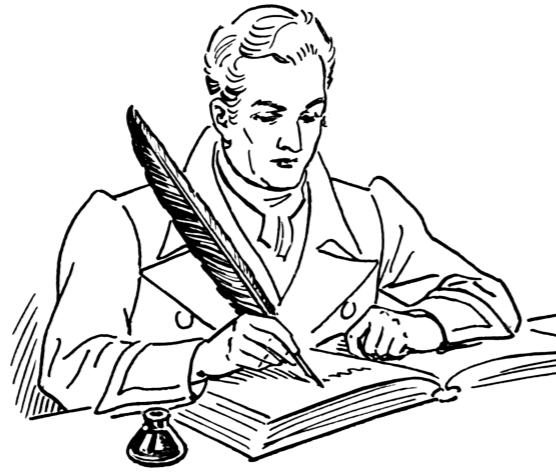
**Ken Liu**

kzliu@cs.stanford.edu

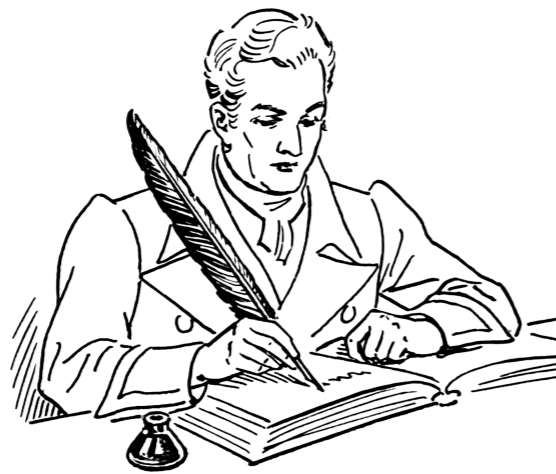
Public slides for ICML 2025 Spotlight

<https://arxiv.org/abs/2503.17514>

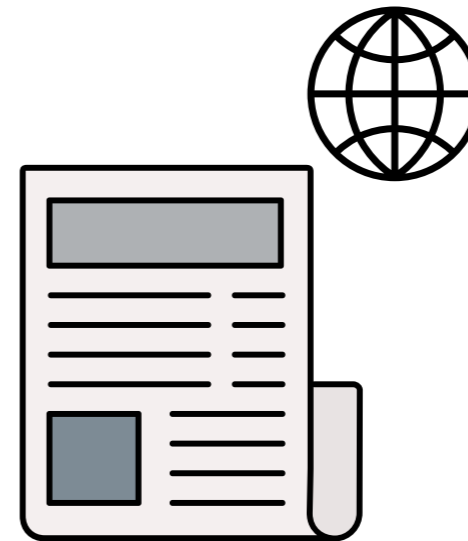




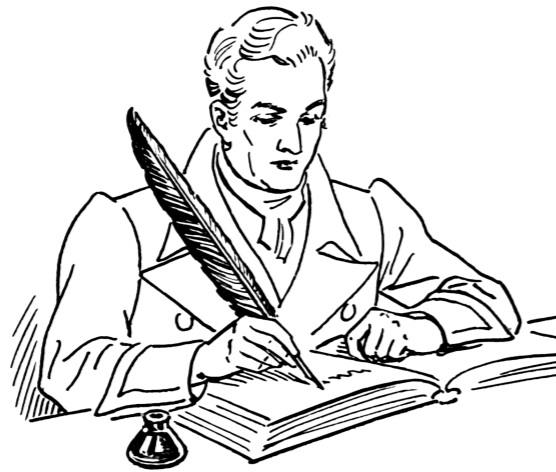
Paul



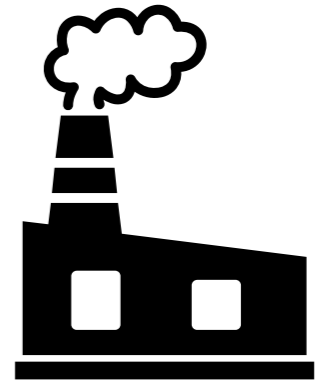
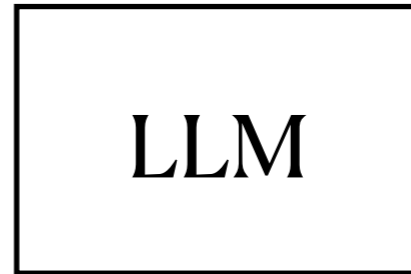
Paul



Published Article

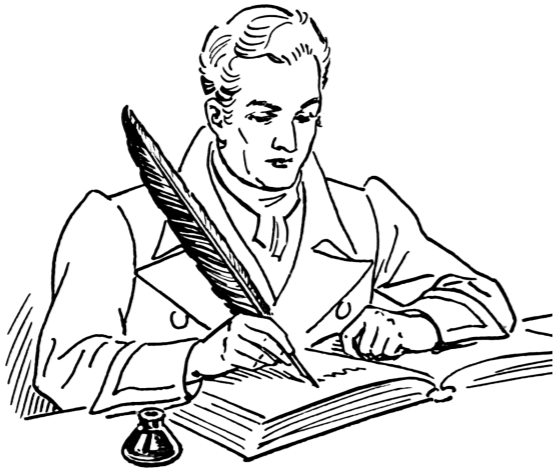


Paul

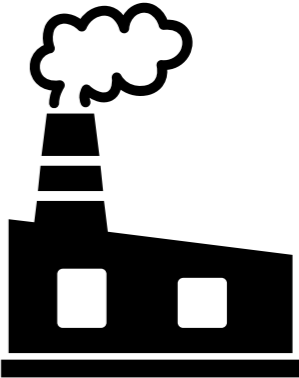
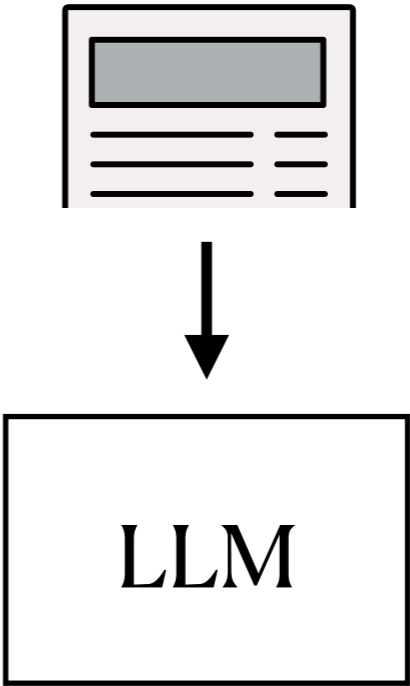


Model  
developer

prompts first-half...

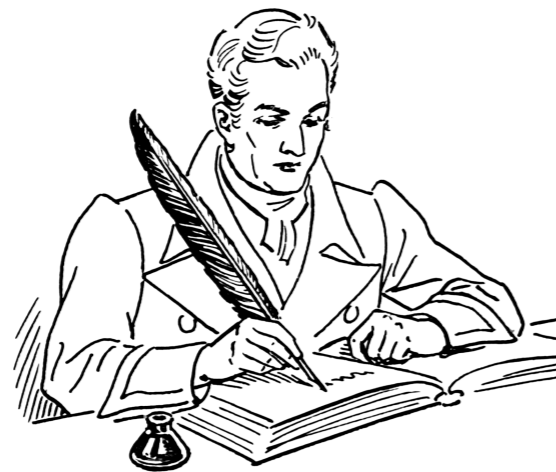


Paul

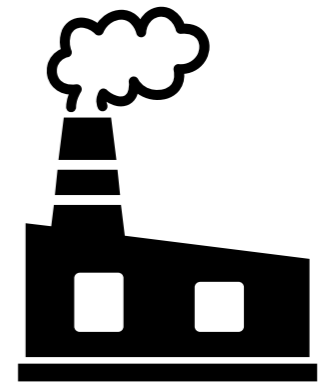
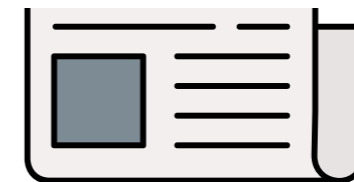


Model  
developer

prompts first-half...



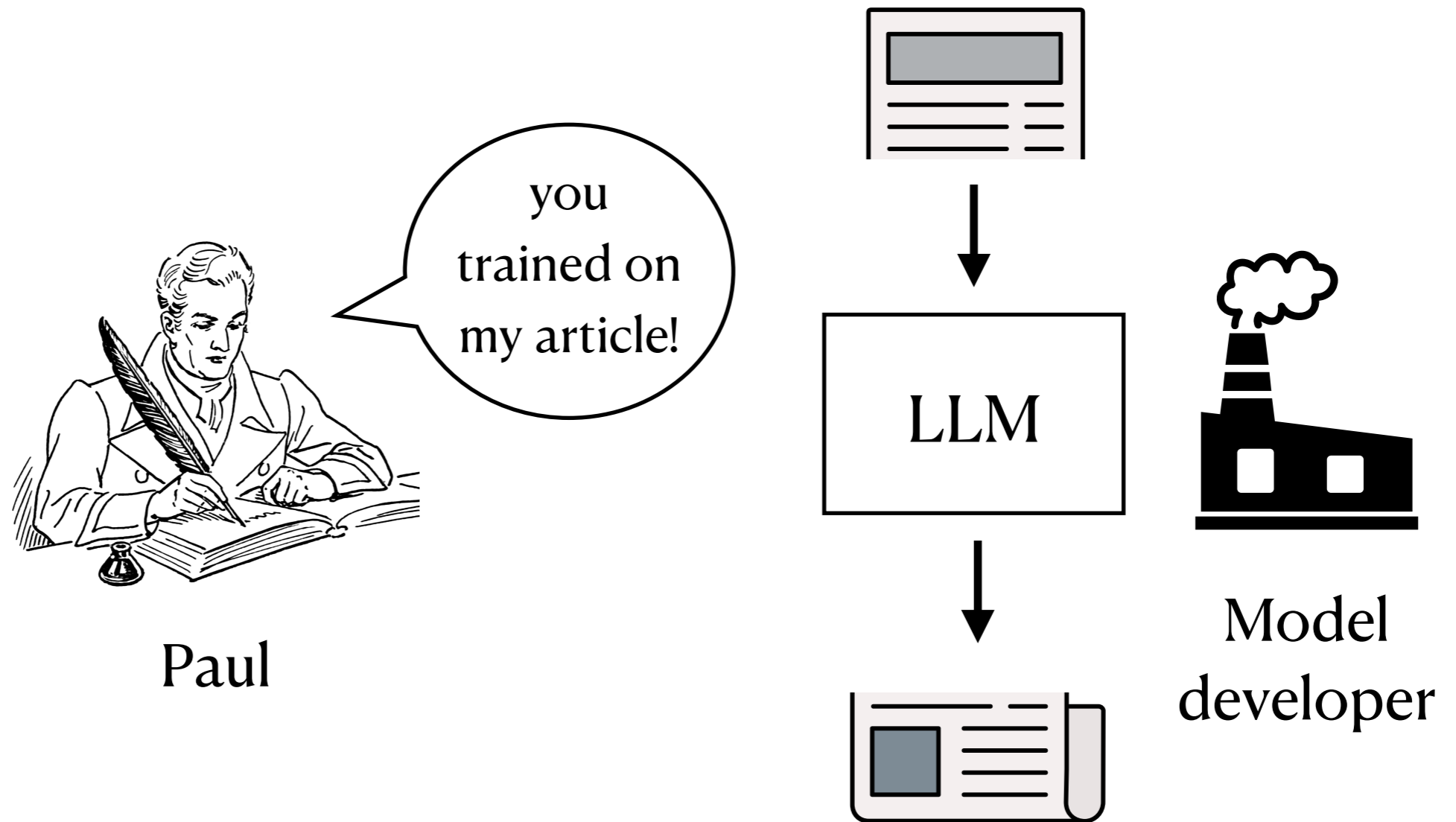
Paul



Model  
developer

...second-half comes  
out **verbatim**

prompts first-half...



...second-half comes  
out **verbatim**



Paul  
New York Times

**The New York Times Company v.  
Microsoft Corporation (1:23-cv-11195)**  
District Court, S.D. New York

<https://sunsteinwebdocs.s3.amazonaws.com/images/NYT-Exhibit-J.pdf>

(Apparently this Exhibit J is no longer used in the lawsuit)

Case 1:23-cv-11195 Document 1-68 Filed 12/27/23 Page 3 of 127

2 ONE HUNDRED EXAMPLES OF GPT-4 MEMORIZING CONTENT FROM THE NEW YORK TIMES

EXAMPLE 1: HOW THE U.S. LOST OUT ON IPHONE WORK

<https://www.nytimes.com/2012/01/22/business/apple-america-and-a-squeezed-middle-class.html>

Copyright number: TX 6-789-918 Effective Date: 2012-05-31

**Prompt (taken from article):**

When Barack Obama joined Silicon Valley's top luminaries for dinner in California last February, each guest was asked to come with a question for the president.

But as Steven P. Jobs of Apple spoke, President Obama interrupted with an inquiry of his own: what would it take to make iPhones in the United States?

Not long ago, Apple boasted that its products were made in America. Today, few are. Almost all of the 70 million iPhones, 30 million iPads and 59 million other products Apple sold last year were manufactured overseas.

Why can't that work come home? Mr. Obama asked.

Mr. Jobs's reply was unambiguous. "Those jobs aren't coming back," he said, according to another dinner guest.

The president's question touched upon a central conviction at Apple. It isn't just that workers are cheaper abroad. Rather, Apple's executives believe the vast scale of overseas factories as well as the flexibility, diligence and industrial skills of foreign workers have so outpaced their American counterparts that "Made in the U.S.A." is no longer a viable option for most Apple products.

Apple has become one of the best-known, most admired and most imitated companies on earth, in part through an unrelenting mastery of global operations. Last year, it earned over \$400,000 in profit per employee, more than Goldman Sachs, Exxon Mobil or Google.

However, what has vexed Mr. Obama as well as economists and policy makers is that Apple — and

GPT-4

**Output from GPT-4:**

many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capi-

**Actual text from NYTimes:**

many of its high-technology peers — are not nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capi-

Here, model *generations* are used as *tests* for training set membership



you  
trained on  
my article!

Paul  
New York Times

**The New York Times Company v.  
Microsoft Corporation (1:23-cv-11195)**  
District Court, S.D. New York

<https://sunsteinwebdocs.s3.amazonaws.com/images/NYT-Exhibit-J.pdf>

(Apparently this Exhibit J is no longer used in the lawsuit)

EXAMPLE 1: HOW THE U.S. LOST OUT ON iPHONE WORK

<https://www.nytimes.com/2012/01/22/business/apple-america-and-a-squeezed-middle-class.html>  
Copyright number: TX 6-789-918 Effective Date: 2012-05-31

**Prompt (taken from article):**

When Barack Obama joined Silicon Valley's top luminaries for dinner in California last February, each guest was asked to come with a question for the president.

But as Steven P. Jobs of Apple spoke, President Obama interrupted with an inquiry of his own: what would it take to make iPhones in the United States?

Not long ago, Apple boasted that its products were made in America. Today, few are. Almost all of the 70 million iPhones, 30 million iPads and 59 million other products Apple sold last year were manufactured overseas.

Why can't that work come home? Mr. Obama asked.

Mr. Jobs's reply was unambiguous. "Those jobs aren't coming back," he said, according to another dinner guest.

The president's question touched upon a central conviction at Apple. It isn't just that workers are cheaper abroad. Rather, Apple's executives believe the vast scale of overseas factories as well as the flexibility, diligence and industrial skills of foreign workers have so outpaced their American counterparts that "Made in the U.S.A." is no longer a viable option for most Apple products.

Apple has become one of the best-known, most admired and most imitated companies on earth, in part through an unrelenting mastery of global operations. Last year, it earned over \$400,000 in profit per employee, more than Goldman Sachs, Exxon Mobil or Google.

However, what has vexed Mr. Obama as well as economists and policy makers is that Apple — and

GPT-4

**Output from GPT-4:**

many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

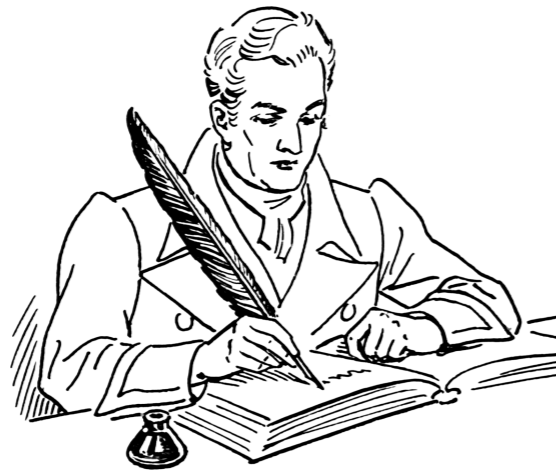
"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capi-

**Actual text from NYTimes:**

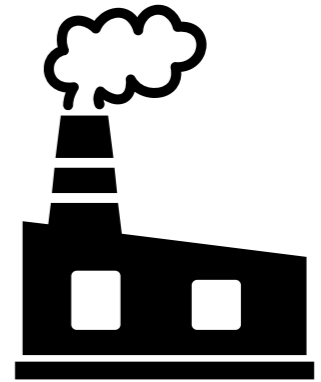
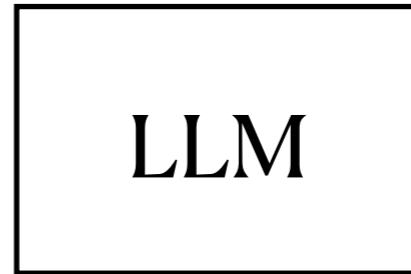
many of its high-technology peers — are not nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

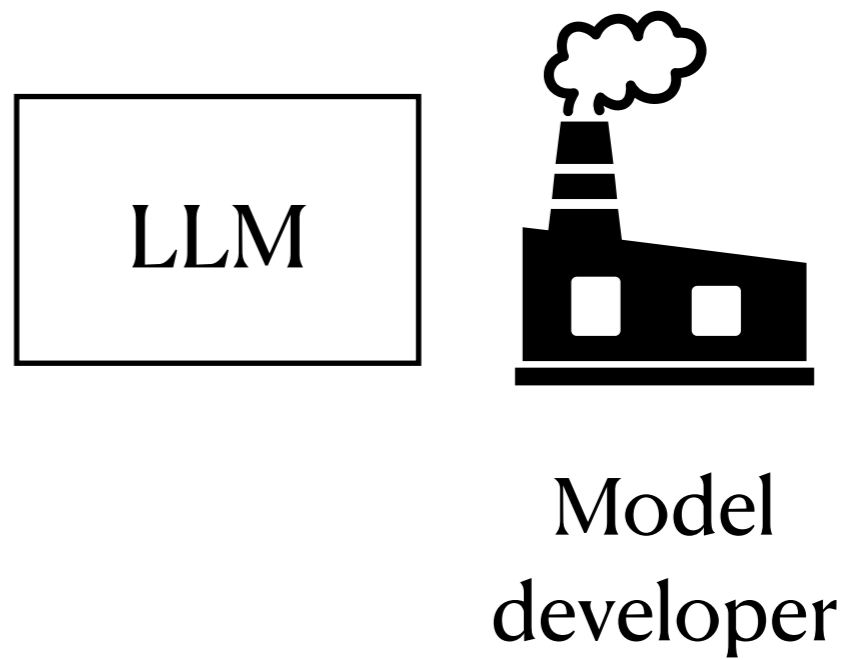
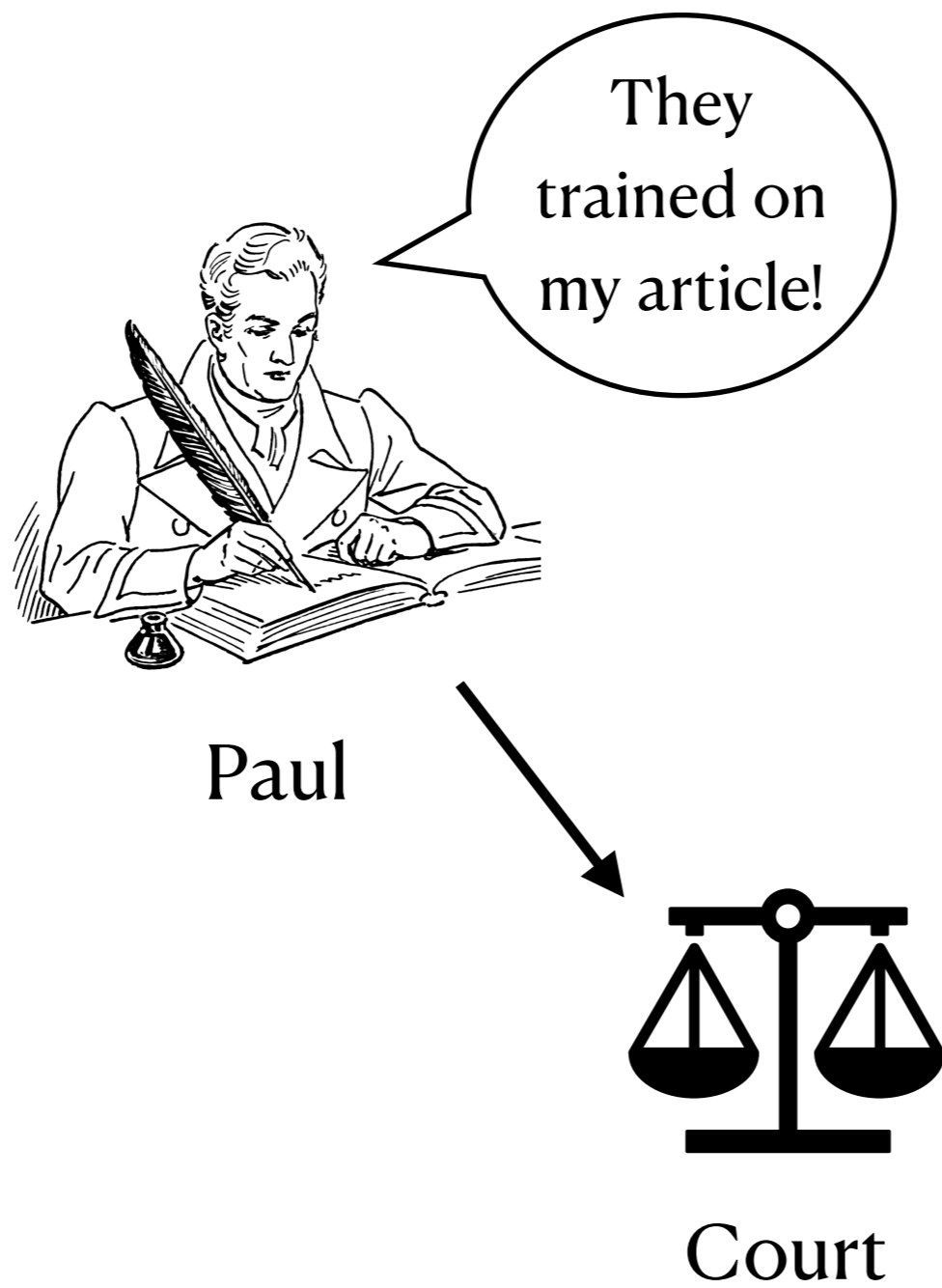
"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capi-

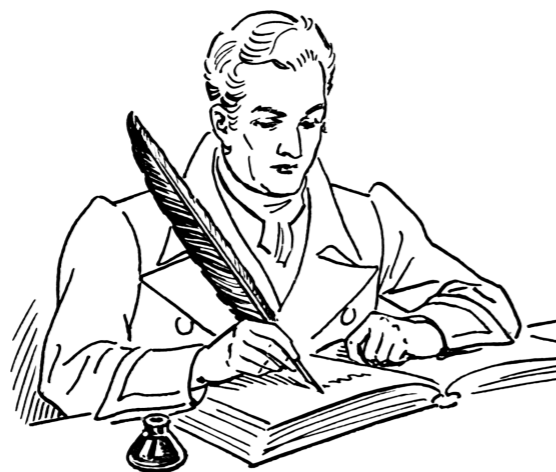


Paul



Model  
developer

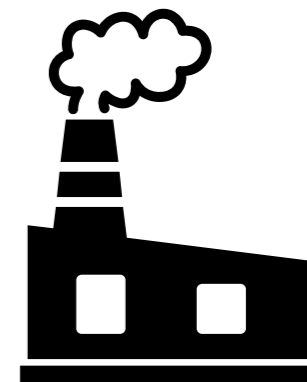




Paul



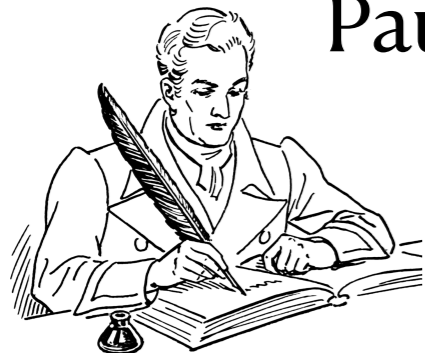
Court



Model  
developer



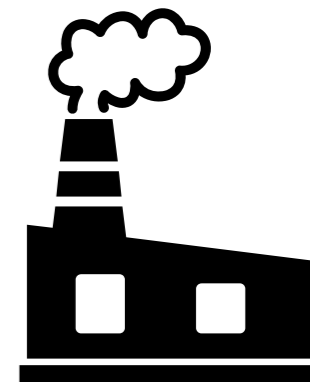
Show me  
your data!



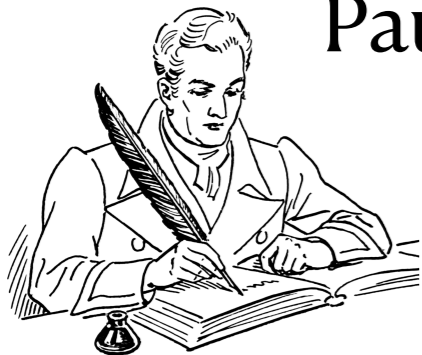
Paul



Court



Model  
developer



Paul

22 use of their copyrighted books as training material for ChatGPT. Nonetheless, their copyrighted  
23 materials were ingested and used to train ChatGPT.  
24 5. Indeed, when ChatGPT is prompted, ChatGPT generates summaries of Plaintiffs'  
25 copyrighted works—something only possible if ChatGPT was trained on Plaintiffs' copyrighted works.  
26 6. Defendants, by and through the use of ChatGPT, benefit commercial and profit richly  
27 from the use of Plaintiffs' and Class members' copyrighted materials.

17	UNITED STATES DISTRICT COURT	
18	NORTHERN DISTRICT OF CALIFORNIA	
19	SAN FRANCISCO DIVISION	
20		
21	IN RE OPENAI CHATGPT LITIGATION	Master File Case No. 3:23-CV-03223-AMO
22	This document relates to:	<del>PROPOSED</del> TRAINING DATA INSPECTION PROTOCOL
23	Case No. 3:23-cv-03223-AMO	
24	Case No. 3:23-cv-03416-AMO	Judge: Hon. Araceli Martínez-Olguín
25	Case No. 3:23-cv-04625-AMO	Date Filed: June 28, 2023
26		
27		
28		
[PROPOSED] TRAINING DATA INSPECTION Master File Case No. 3:23-CV-03223-AMO		

11 3. Training Data shall be made available for inspection in electronic format at  
12 OpenAI's offices in San Francisco CA, or at a secure location determined by OpenAI within 25  
13 miles of San Francisco, CA; or at another mutually agreed location. Training Data will be made  
14 available for inspection between the hours of 8:30 a.m. and 5:00 p.m. on business days, although  
15 the parties will be reasonable in accommodating reasonable requests to conduct inspections at  
16 other times.  
17 4. The Inspecting Party shall provide five days' notice prior to any inspection.



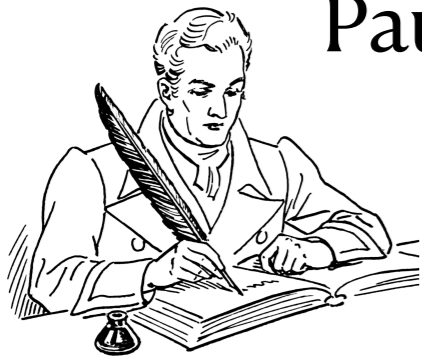
Court

**Tremblay v. OpenAI, Inc. (3:23-cv-03223)**  
District Court, N.D. California

<https://www.courtlistener.com/docket/67538258/tremblay-v-openai-inc/>



Model developer



Paul

22 use of their copyrighted books as training material for ChatGPT. Nonetheless, their copyrighted  
23 materials were ingested and used to train ChatGPT.  
24 5. Indeed, when ChatGPT is prompted, ChatGPT generates summaries of Plaintiffs'  
25 copyrighted works—something only possible if ChatGPT was trained on Plaintiffs' copyrighted works.  
26 6. Defendants, by and through the use of ChatGPT, benefit commercial and profit richly

5. Indeed, when ChatGPT is prompted, ChatGPT generates summaries of Plaintiffs' copyrighted works—something only possible if ChatGPT was trained on Plaintiffs' copyrighted works.

21 IN RE OPENAI CHATGPT LITIGATION

Master File Case No. 3:23-CV-03223-AMO

22 This document relates to:

~~PROPOSED~~ TRAINING DATA  
INSPECTION PROTOCOL

23 Case No. 3:23-cv-03223-AMO

24 Case No. 3:23-cv-03416-AMO

25 Case No. 3:23-cv-04625-AMO

Judge: Hon. Araceli Martínez-Olguín

Date Filed: June 28, 2023

3. Training Data shall be made available for inspection in electronic format at OpenAI's offices in San Francisco CA, or at a secure location determined by OpenAI within 25 miles of San Francisco, CA; or at another mutually agreed location. Training Data will be made

17

4. The Inspecting Party shall provide five days' notice prior to any inspection.

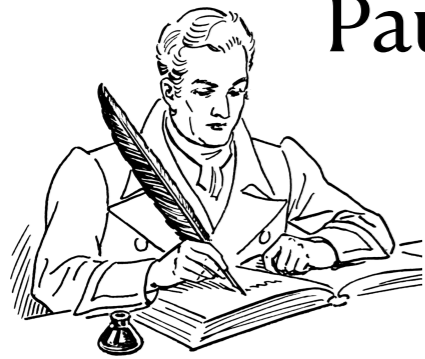
Court

**Tremblay v. OpenAI, Inc. (3:23-cv-03223)**  
District Court, N.D. California

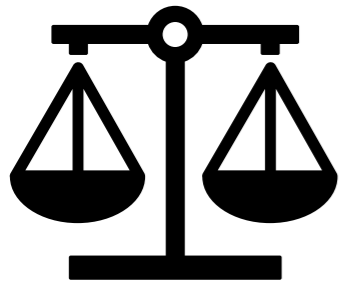
<https://www.courtlistener.com/docket/67538258/tremblay-v-openai-inc/>



Model developer



Paul



Court



Model developer





here you go!  
10T tokens

# Model developer



# Training set membership underpins...

## Policies for data transparency



[skip to content](#) [home](#) [accessibility](#) [FAQ](#) [feedback](#) [sitemap](#)

[Home](#) [Bill Information](#) [California Law](#) [Publications](#) [Other Resources](#) [My Subscriptions](#) [My Favorites](#)

[Bill Information](#) >> [Bill Search](#) >> [Text](#)

[Bill PDF](#) | [Add To My Favorites](#) | Version:

**AB-2013 Generative artificial intelligence: training data transparency.** (2023-2024)

Text	Votes	History	Bill Analysis	Today's Law As Amended	Compare Ver
------	-------	---------	---------------	------------------------	-------------

SHARE THIS:

Text: CA AB412 | 2025-2026 | Regular S

California Assembly Bill 412

legislature page for AB412


[Sponsors](#) [Texts](#) [Votes](#) [Research](#) [Comments](#) [Track](#)

[Amended](#) [Amended](#) [Amended](#) [Amended](#)

There are more recent revisions of this legislation. [Read Latest Draft](#)

Generative artificial intelligence: training data: copyrighted materials.

Partisan Bill (Democrat 1-0)



PREMIUM DESIGN TECH WORK LIFE NEWS IMPACT PODCASTS VIDEO INNOVATION

04-10-2024 | TECH

## A new bill would force companies like

## disclose their training data

copyright Disclosure Act proposed by Rep. Adam Schiff is

ative unions, but it faces a potential uphill battle.

Language model developers should report  
train-test overlap

Andy K Zhang, Kevin Klyman, Yifan Mai, Yoav Levine, Yian Zhang, Rishi Bom-  
masani, Percy Liang  
[andyzh@stanford.edu](mailto:andyzh@stanford.edu)  
Stanford University

# Training set membership underpins... Data contamination

## Evaluation data contamination in LLMs: how do we measure it and (when) does it matter?

**Aaditya K. Singh\***  
University College London  
aaditya.singh.21@ucl.ac.uk

**Muhammed Yusuf Kocyigit\***<sup>†</sup>  
Boston University  
kocyigit@bu.edu

**Andrew Poulton**<sup>†</sup>  
Cohere  
andrewpoulton@cohere.com

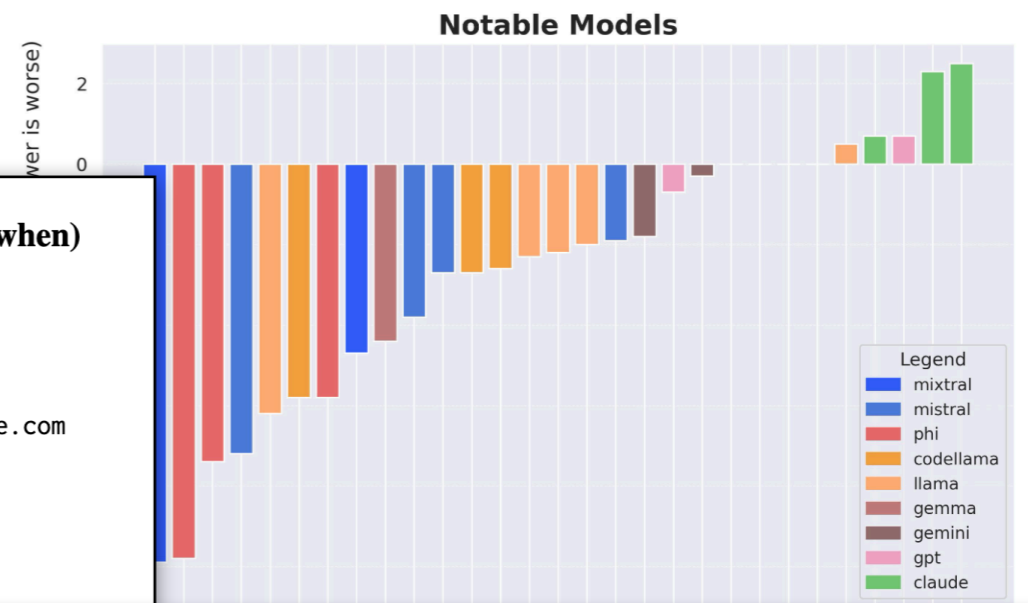
**David Esiobu**  
Meta

**Maria Lomeli**  
Meta

**Gergely Szilvasy**  
Meta

**Dieuwke Hupkes**  
Meta  
dieuwkehupkes@meta.com

Abstract



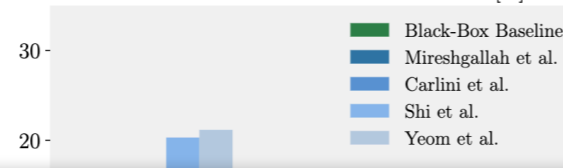
## Evading Data Contamination Detection for Language Models is (too) Easy

Jasper Dekoninck<sup>1</sup> Mark Niklas Müller<sup>1</sup> Maximilian Baader<sup>1</sup> Marc Fischer<sup>1</sup> Martin Vechev<sup>1</sup>

### Abstract

Large language models (LLMs) are widespread, with their performance on benchmarks frequently guiding user preferences for one model over another. However, the vast amount of data these models are trained on can inadvertently lead to contamination with public benchmarks, thus compromising performance measurements. While recently developed contamination detection methods try to address this issue, they overlook the possibility of deliberate contamination by malicious model providers aiming to evade detection. We argue that this setting is of special importance as it casts doubt on the reliability of benchmark results for LLMs.

Contamination Detection TPR@1%FPR [%]



## PROVING TEST SET CONTAMINATION IN BLACK BOX LANGUAGE MODELS

**Yonatan Oren**<sup>1\*</sup>, **Nicole Meister**<sup>1\*</sup>, **Niladri Chatterji**<sup>1\*</sup>, **Faisal Ladhak**<sup>2</sup>, **Tatsunori B. Hashimoto**<sup>1</sup>  
<sup>1</sup>Stanford University, <sup>2</sup>Columbia University  
yonatano@cs.stanford.edu  
{nmeister, niladric, thashim}@stanford.edu  
faisal@cs.columbia.edu

## MATH-Perturb

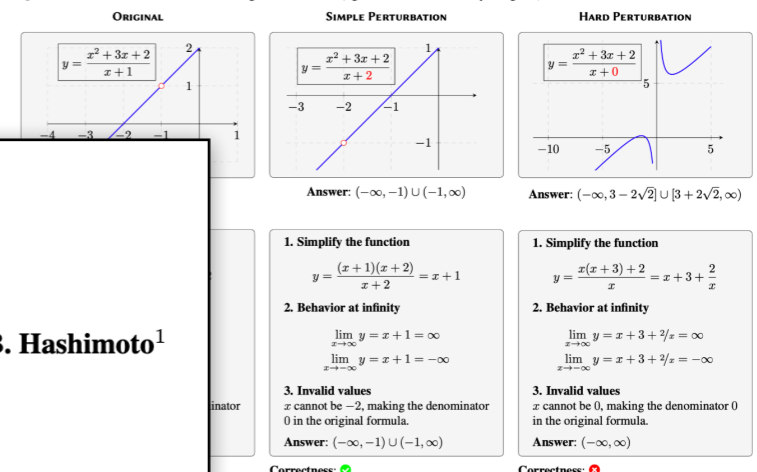
Benchmarking LLMs' Math Reasoning Abilities against Hard Perturbations

Kaixuan Huang<sup>1</sup>, Jiacheng Guo<sup>1</sup>, Zihao Li<sup>1</sup>, Xiang Ji<sup>1</sup>, Jiawei Ge<sup>1</sup>, Wenzhe Li<sup>1</sup>, Yingqing Guo<sup>1</sup>,  
Tianle Cai<sup>1</sup>, Hui Yuan<sup>1</sup>, Runzhe Wang<sup>1</sup>, Yue Wu<sup>1</sup>, Ming Yin<sup>1</sup>, Shange Tang<sup>1</sup>,  
Yangsibo Huang<sup>2</sup>, Chi Jin<sup>1</sup>, Xinyun Chen<sup>2</sup>, Chiyuan Zhang<sup>2</sup>, Mengdi Wang<sup>1</sup>  
<sup>1</sup>Princeton University, <sup>2</sup>Google

[Paper](#) [arXiv](#) [Leaderboard](#) [Twitter](#)


### Overview of MATH-Perturb Benchmark

Question: Given the formula, find out the range of the function (figure is for illustration only, not given).



# Training set membership underpins...

## Data deduplication & memorization analysis

 **Hugging Face**  [Models](#) [Datasets](#) [Spaces](#) [Docs](#)

[← Back to Articles](#)

### Large-scale Near-deduplication Behind BigCode

#### Deduplicating Training Data Makes Language Models Better

Katherine Lee<sup>\*†</sup>   Daphne Ippolito<sup>\*†‡</sup>   Andrew Nystrom<sup>†</sup>   Chiyuan Zhang<sup>†</sup>  
Douglas Eck<sup>†</sup>   Chris Callison-Burch<sup>‡</sup>   Nicholas Carlini<sup>†</sup>

#### Demystifying Verbatim Memorization in Large Language Models

Jing Huang, Diyi Yang<sup>\*</sup>, Christopher Potts<sup>\*</sup>  
Stanford University  
{hij, diyi, cgpotts}@stanford.edu

---

#### Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the *New York Times v. OpenAI* 2023 Lawsuit

---

Joshua Freeman   Chloe Rippe   Edoardo DeBenedetti   Maksym Andriushchenko  
ETH Zurich   Duke University   ETH Zurich   EPFL

How we **define** training set membership matters a lot in practice

And the definition should match the **downstream consequences** we care about

For language models, we care a lot about **model generations** (privacy, evals, copyright...)

**This talk:** pitfalls of n-gram training set membership

# This talk: pitfalls of n-gram training set membership



**Definition 3.1** (*n*-gram data membership). *A sequence  $x$  is a member of a dataset  $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$  if  $x$  shares at least one *n*-gram with any  $x^{(i)} \in \mathcal{D}$ . That is,  $x$  is member if there exists a  $g \in \text{n-grams}(x)$  s.t.  $g \in \bigcup_i \text{n-grams}(x^{(i)})$ .*

very inclusive!

overestimates & captures definitions in the literature

**This talk:** pitfalls of n-gram training set membership  
n-grams are intuitive and used everywhere!

# GPT-4 Technical Report

Deduplicating Training Data Makes Language Models Better

Language model developers should report train-test overlap

OpenAI\*

Andy K Zhang, Kevin Klyman, Yifan Mai, Yoav Levine, Yian Zhang, Rishi Bommasani, Percy Liang  
andyzh@stanford.edu  
Stanford University



# The Llama 3 Herd of Models

Llama Team, AI @ Meta<sup>1</sup>

<sup>1</sup>A detailed contributor list can be found in the appendix of this paper.

Bill Text: CA AB412 | 2025-2026 | Regular S

UNITED STATES DISTRICT COURT

18

NORTHERN DISTRICT OF CALIFORNIA

This talk: pitfalls of n-gram training set membership  
n-grams are intuitive and used everywhere!

Status: (Introduced) 2025-04-22 - Re-referred to Com. on JUD. [AB412 Detail]

Download: California-2025-AB412-Introduced.html

# Demystifying Verbatim Memorization in Large Language Models

Jing Huang, Diyi Yang\*, Christopher Potts\*  
Stanford University

{hij, diyiy, cgpotts}@stanford.edu

Do Membership Inference Attacks Work on Large Language Models?

Michael Duan\*<sup>1</sup> Anshuman Suri\*<sup>2</sup>  
Niloofar Miresghallah<sup>1</sup> Sewon Min<sup>1</sup> Weijia Shi<sup>1</sup> Luke Zettlemoyer<sup>1</sup>  
Yulia Tsvetkov<sup>1</sup> Yejin Choi<sup>1</sup> David Evans<sup>2</sup> Hannaneh Hajishirzi<sup>1,3</sup>

<sup>1</sup>University of Washington <sup>2</sup>University of Virginia <sup>3</sup>Allen Institute for AI  
<micdun@cs.washington.edu> <as9rw@virginia.edu>

# Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the *New York Times v. OpenAI* 2023 Lawsuit

Joshua Freeman  
ETH Zurich

Chloe Rippe  
Duke University

Edoardo DeBenedetti  
ETH Zurich

Maksym Andriushchenko  
EPFL

# **This talk:** pitfalls of $n$ -gram training set membership

A tale of two experiments:

1. **Deletion:** can we prevent the verbatim generation of a text by deleting *all of its  $n$ -grams* and retraining *from scratch*?
2. **Addition:** can we cause the verbatim generation of a text by training on texts with *no  $n$ -gram overlap*?

# This talk: pitfalls of $n$ -gram training set membership

## A tale of two experiments:

1. **Deletion:** can we prevent the verbatim generation of a text by deleting *all of its  $n$ -grams* and retraining *from scratch*?  
→ **No!** Many deleted texts can *still* be completed verbatim
2. **Addition:** can we cause the verbatim generation of a text by training on texts with *no  $n$ -gram overlap*?  
→ **Yes!** And it only takes a few gradient steps

**"Language Models May Verbatim Complete Text They Were Not Explicitly Trained On."**

Ken Ziyu Liu, Christopher A. Choquette-Choo, Matthew Jagielski, Peter Kairouz, Sanmi Koyejo, Percy Liang, Nicolas Papernot. <https://arxiv.org/abs/2503.17514>. **ICML 2025, Spotlight.**

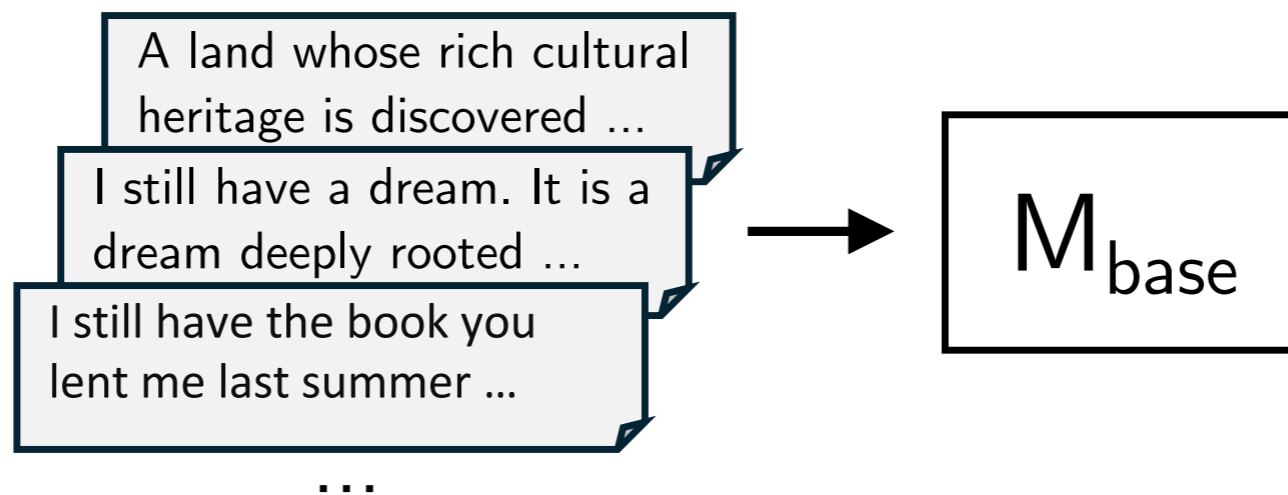
Result #1:

Removing  $n$ -gram members **may not**  
prevent LLM verbatim completion

**Setup:** just retrain from scratch!

# Setup: just retrain from scratch!

1. Pre-train a model  $M_{\text{base}}$  on data  $D$



# Setup: just retrain from scratch!

1. Pre-train a model  $M_{\text{base}}$  on data  $D$
2. Identify a list of texts  $D_{\text{mem}}$  that  $M_{\text{base}}$  memorizes

Prompt the prefixes...

A land whose rich cultural  
I still have a dream.  
I still have the book

...



$M_{\text{base}}$



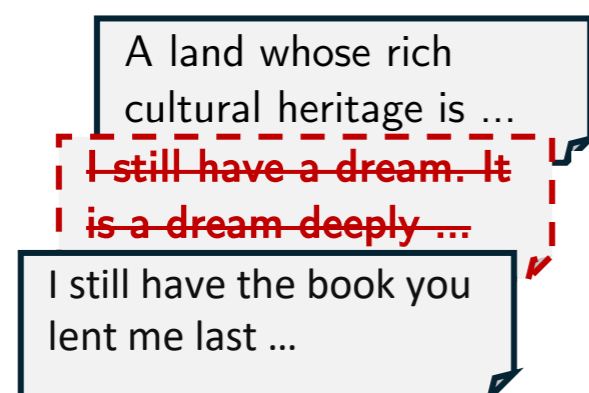
...do completions match?

heritage weaves together ... X  
It is a dream deeply rooted ... ✓  
sitting on my desk ... X

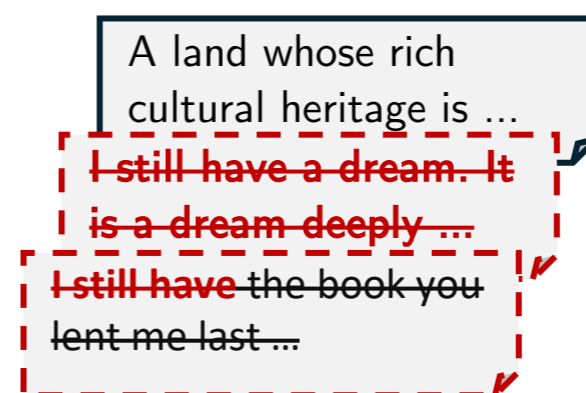
...

# Setup: just retrain from scratch!

1. Pre-train a model  $M_{\text{base}}$  on data  $D$
2. Identify a list of texts  $D_{\text{mem}}$  that  $M_{\text{base}}$  memorizes
3. Filter  $D_{\text{mem}}$  from  $D$  by n-gram overlap, get  $D_{\text{filter}}^{(n)}$



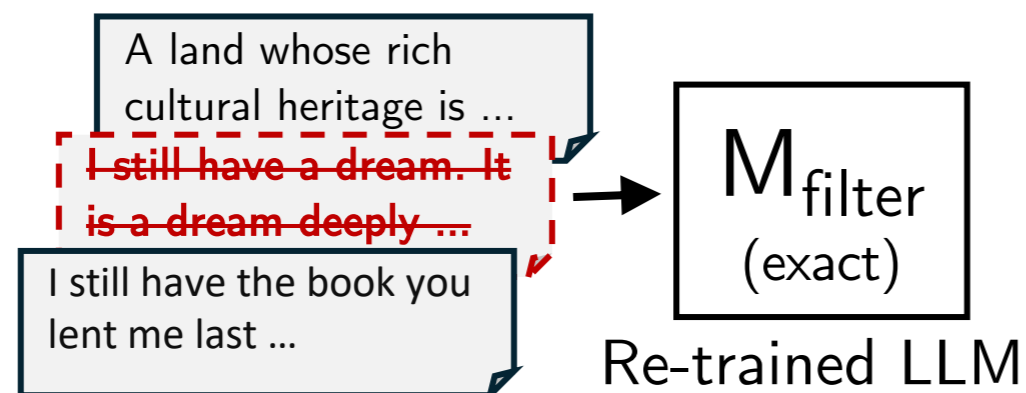
**Exact filter (weaker):** remove all sequences in  $D_{\text{mem}}$ , exactly as they appear, from the pre-training dataset



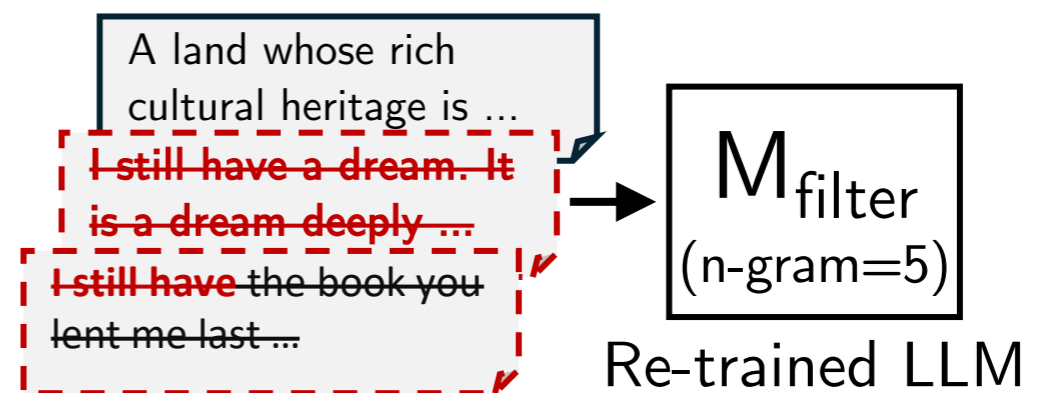
**N-gram filter (stronger):** remove sequences with any n-gram overlap against any of the sequences in  $D_{\text{mem}}$

# Setup: just retrain from scratch!

1. Pre-train a model  $M_{\text{base}}$  on data  $D$
2. Identify a list of texts  $D_{\text{mem}}$  that  $M_{\text{base}}$  memorizes
3. Filter  $D_{\text{mem}}$  from  $D$  by n-gram overlap, get  $D_{\text{filter}}^{(n)}$
4. Re-train *from scratch* on  $D_{\text{filter}}^{(n)}$  and get  $M_{\text{filter}}^{(n)}$



**Exact filter (weaker):** remove all sequences in  $D_{\text{mem}}$ , exactly as they appear, from the pre-training dataset



**N-gram filter (stronger):** remove sequences with any n-gram overlap against any of the sequences in  $D_{\text{mem}}$

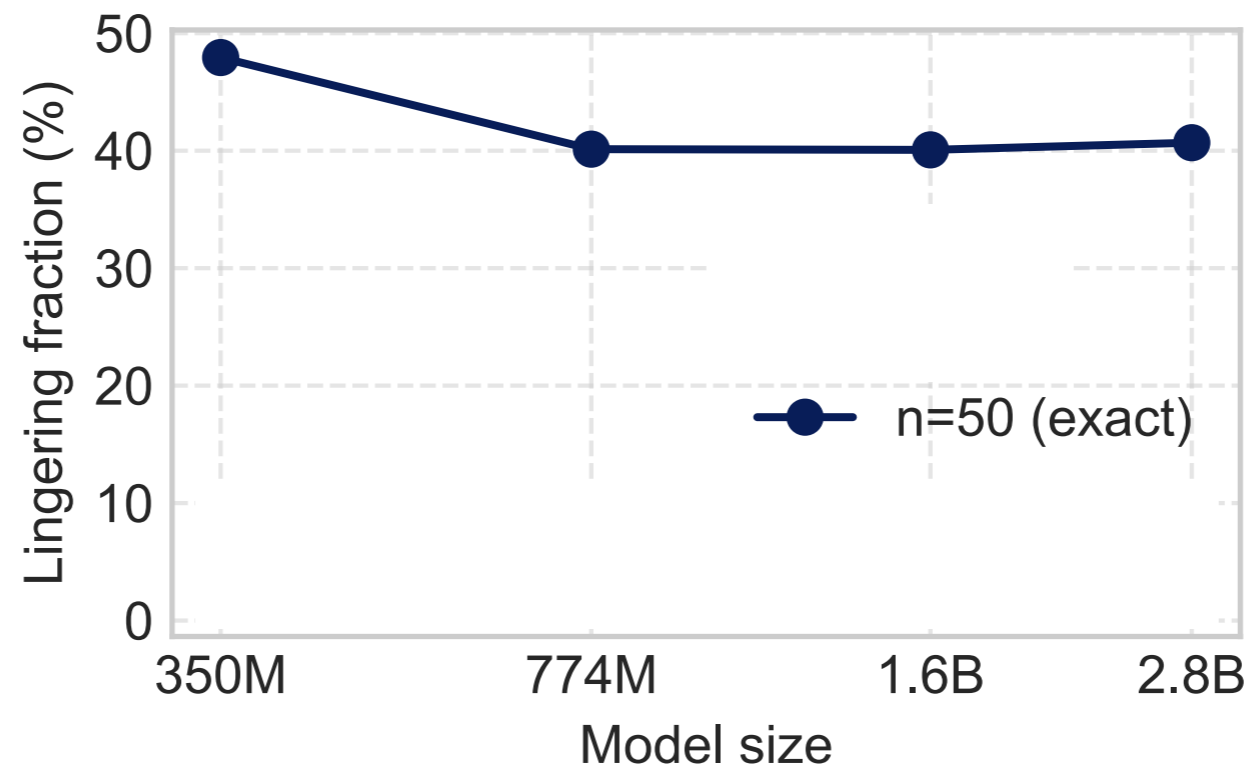
# Result 1.1: "lingering sequences" exist

- If we delete the texts as-is, ~40% can *still* be completed:

# Result 1.1: "lingering sequences" exist

- If we delete the texts as-is, ~40% can *still* be completed:

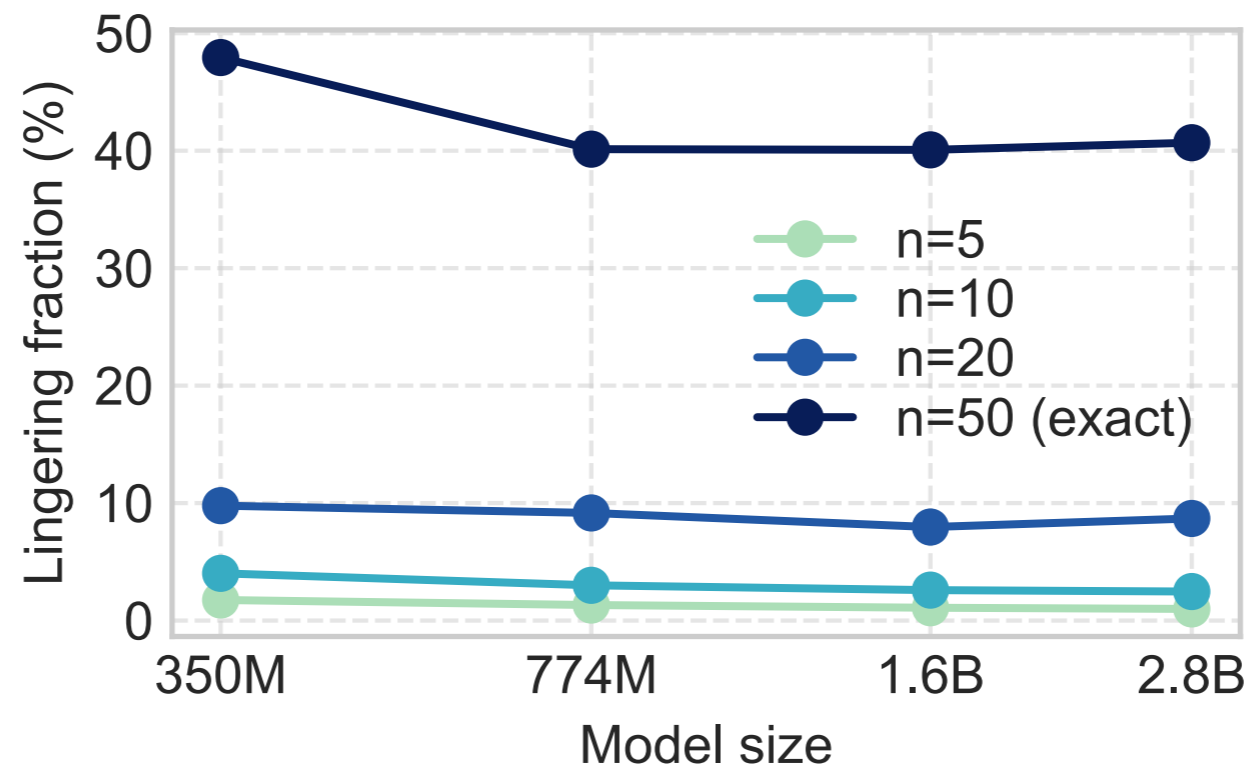
Model size	304M	774M	1.6B	2.8B
$ \mathcal{D}_{\text{mem}} $	76,648	116,270	151,598	175,813



# Result 1.1: "lingering sequences" exist

- If we delete the n-grams, we can drive down the fraction...

Model size	304M	774M	1.6B	2.8B
$ \mathcal{D}_{\text{mem}} $	76,648	116,270	151,598	175,813

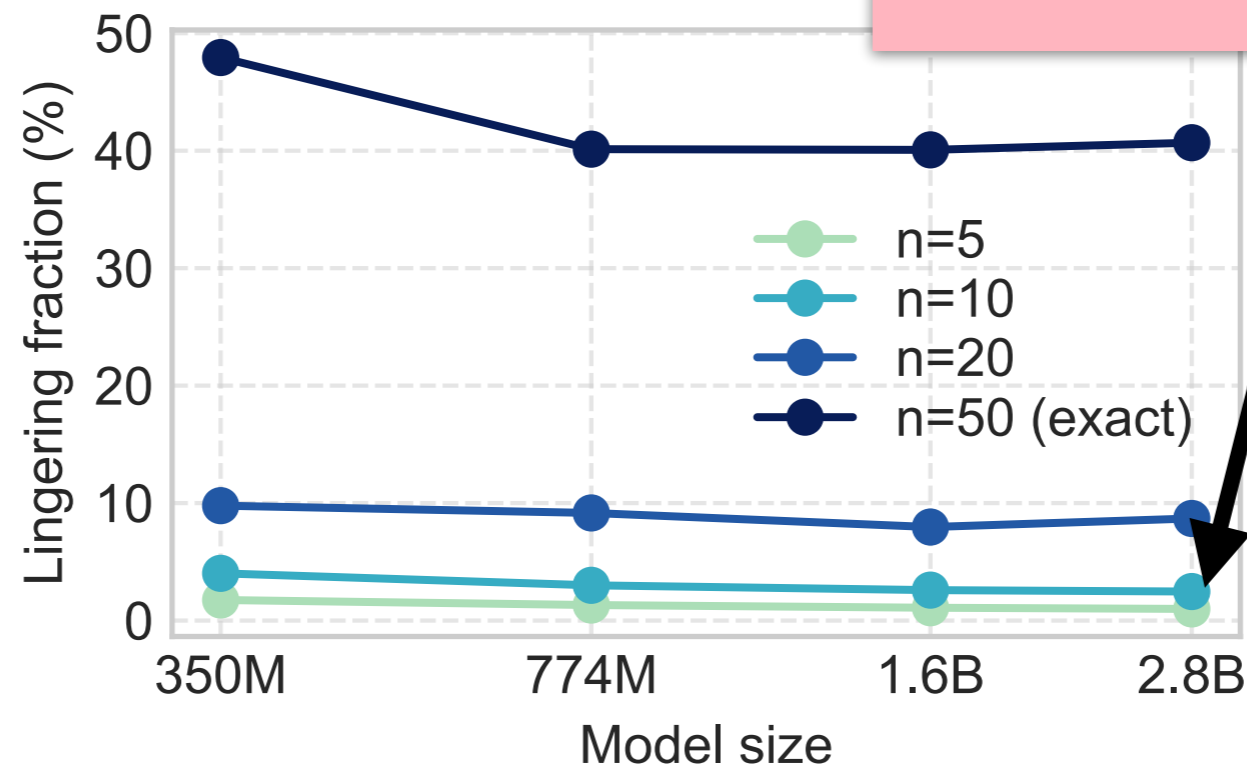


# Result 1.2: "lingering sequences" *persist*

- If we delete the n-grams, we can drive down the fraction...

Model size	304M	774M	1.6B
$ \mathcal{D}_{\text{mem}} $	76,648	116,270	151,111

... but cannot eliminate them completely (~1%)



# Result 1.2: "lingering sequences" *persist*

- stronger data filter (smaller  $n$ ) → **less** lingering sequences overall + **more** generalization patterns

# Result 1.2: "lingering sequences" *persist*

- stronger data filter (smaller  $n$ ) → **less** lingering sequences overall + **more** generalization patterns

$n = 100$

Table 12: Randomly sampled lingering sequences at filtering strength  $n = 100$  (exact) filter. Sequence length

Idx	Lingering sequences at filtering strength $n = 100$ (exact) filter.
0	<p><b>Prompt:</b> "Course Hero. "Romeo and Juliet Study Guide." Course Hero. 28 July 2016. Web. 16 Jan. 2019. <a href="https://www.coursehero.com/lit/Romeo-and-Juliet/">https://www.coursehero.com/lit/Romeo-and-Juliet/</a>. Course"</p> <p><b>Completion:</b> " Hero. (2016, July 28). Romeo and Juliet Study Guide. In Course Hero. Retrieved January 16, 2019, from <a href="https://www.coursehero.com/lit/Romeo-and-Juliet/">https://www.coursehero.com/lit/Romeo-and-Juliet/</a> (Course Hero,"</p>
1	<p><b>Prompt:</b> "ESQUIEN INDIANS. The Esquien Indians were a Karankawan group that entered Nuestra Señora de la Candelaria Mission on the San Gabriel River in 1750. Frederick Webb Hodge, ed.,"</p> <p><b>Completion:</b> " Handbook of American Indians North of Mexico (2 vols., Washington: GPO, 1907, 1910; rpt., New York: Pageant, 1959). The following, adapted from the Chicago Manual of Style, 15th edition, is the"</p>
2	<p><b>Prompt:</b> "See what questions a doctor would ask. During a consultation, your doctor will use various techniques to assess the symptom: Heel pain. These will include a physical examination and possibly diagnostic tests. (Note: A physical exam is always done,"</p> <p><b>Completion:</b> " diagnostic tests may or may not be performed depending on the suspected condition) Your doctor will ask several questions when assessing your condition. It is important to openly share any pertinent information to help your doctor make an accurate diagnosis. It is also very important to"</p>
3	<p><b>Prompt:</b> "—Product #: EMC0775025_TQ— A Is for Apple (Resource Book Only) eBookGrade 2—Grade 3—Grade 4—Grade 5 Please Note: This ebook is a digital download, NOT a physical product"</p> <p><b>Completion:</b> ". After purchase, you will be provided a one time link to download ebooks to your computer. Orders paid by PayPal require up to 8 business hours to verify payment and release electronic media. For immediate downloads, payment with credit card is required. "</p>
4	<p><b>Prompt:</b> "Family History and Genealogy Resources by Surname Haycock Surname Origin A name probably given to a foundling exposed in a hayfield. Source: An Etymological Dictionary of Family and Christian Names With an Essay"</p> <p><b>Completion:</b> " on their Derivation and Import; Arthur, William, M.A.; New York, NY: Sheldon, Blake, Bleeker &amp; CO., 1857. Haycock Surname Meaning and Family Facts There is more to Haycock family"</p>

(recall that *no* n-grams of any of these sequences are in the training data)

# Result 1.2: "lingering sequences" *persist*

- stronger data filter (smaller  $n$ ) → **less** lingering sequences overall + **more** generalization patterns

$n = 100$

Table 12: Randomly sampled lingering sequences at filtering strength  $n = 100$  (exact) filter. Sequence length

Idx	Lingering sequences at filtering strength $n = 100$ (exact) filter.
0	<p><b>Prompt:</b> "Course Hero. "Romeo and Juliet Study Guide." Course Hero. 28 July 2016. Web. 16 Jan. 2019.  https://www.coursehero.com/lit/Romeo-and-Juliet/. Course"</p> <p><b>Completion:</b> " Hero. (2016, July 28). Romeo and Juliet Study Guide. In Course Hero. Retrieved January 16, 2019, from https://www.coursehero.com/lit/Romeo-and-Juliet/ (Course Hero."</p> <p>adapted from the Chicago Manual of Style, 15th edition, is the</p>
2	<p><b>Prompt:</b> "See what questions a doctor would ask. During a consultation, your doctor will will include a physical examination and possibly diagnostic tests. (Note: A physical exam is</p> <p><b>Completion:</b> " diagnostic tests may or may not be performed depending on the suspected condition your condition. It is important to openly share any pertinent information to help your doctor</p>
3	<p><b>Prompt:</b> "Book # EM00775005 TO: ALEXANDER, D. L. O. L. A."</p> <p>ebook is a digital download, NOT a physical product</p> <p><b>Completion:</b> ". After purchase, you will be provided a one time link to download ebooks to your computer. Orders paid by PayPal require up to 8 business hours to verify payment and release electronic media. For immediate downloads, payment with credit card is required. "</p>
4	<p><b>Prompt:</b> "Family History and Genealogy Resources by Surname Haycock Surname Origin A name probably given to a foundling exposed in a hayfield. Source: An Etymological Dictionary of Family and Christian Names With an Essay"</p> <p><b>Completion:</b> " on their Derivation and Import; Arthur, William, M.A.; New York, NY: Sheldon, Blake, Bleeker &amp; CO., 1857. Haycock Surname Meaning and Family Facts There is more to Haycock family"</p>

(recall that *no*  $n$ -grams of any of these sequences are in the training data)

# Result 1.2: "lingering sequences" *persist*

- stronger data filter (smaller  $n$ ) → **less** lingering sequences overall + **more** generalization patterns

$n = 50$

**$n = 50$  (exact filtering):** the entire sequence, as it appears exactly, is not in training data

**Prompt:** - the domain of a baron.\n- baronage(def 2).\nOrigin of barony\nDictionary\

**Completion:** .com Unabridged Based on the Random House Unabridged Dictionary, © Random House, Inc. 2018\n

**Prompt:** We hold these truths to be self-evident, that all men (and women) are created equal, that they

**Completion:** are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty

# Result 1.2: "lingering sequences" *persist*

- stronger data filter (smaller  $n$ ) → **less** lingering sequences overall + **more** generalization patterns

$n = 50$

Table 4: Randomly sampled lingering sequences at filtering strength  $n = 50$  (exact) filter. Sequence length

Idx	Lingering sequences at filtering strength $n = 50$ (exact) filter.	
0	<b>Prompt:</b>	'Common Name: bleeding heart Type: Herbaceous perennial Native Range: Eastern United States Zone: 3 to 9 '
	<b>Completion:</b>	'Height: 1.00 to 1.50 feet Spread: 1.00 to 1.50 feet Bloom Time:'
1	<b>Prompt:</b>	'—1477 by topic— —Arts and science— —Birth and death categories— —Births —'
	<b>Completion:</b>	'Deaths— —Establishments and disestablishments categories— —Establishments – Disestablishments— —'
2	<b>Prompt:</b>	'Charcot Joint (Neuropathic Arthropathy) Medicine Central™ is a quick-consult mobile and'
	<b>Completion:</b>	' web resource that includes diagnosis, treatment, medications, and follow-up information on over 700 diseases and disorders, providing fast answers'
3	<b>Prompt:</b>	'Mienert-barth Surname History The family history of the Mienert-barth last name is'
	<b>Completion:</b>	' maintained by the AncientFaces community. Join the community by adding to to our knowledge of the Mienert-'
4	<b>Prompt:</b>	'Instructional Supports and Resources Dyslexia is a specific learning disability that is neurological in origin. It is characterized'
	<b>Completion:</b>	' by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These difficulties typically result from a deficit'
5	<b>Prompt:</b>	'Publisher description for Writers at work. The short composition / Ann O. Strauch. Bibliographic record and links to'
	<b>Completion:</b>	' related information available from the Library of Congress catalog Information from electronic data provided by the publisher. May be incomplete or contain other'
6	<b>Prompt:</b>	'Create healthcare diagrams like this example called Anencephaly in minutes with SmartDraw. SmartDraw includes 1000s of professional healthcare'
	<b>Completion:</b>	' and anatomy chart templates that you can modify and make your own. Text in this Example: Anencephaly is'

(recall that *no*  $n$ -grams of any of these sequences are in the training data)

# Result 1.2: "lingering sequences" *persist*

- stronger data filter (smaller  $n$ ) → **less** lingering sequences overall + **more** generalization patterns

$n = 50$

Table 4: Randomly sampled lingering sequences at filtering strength  $n = 50$  (exact) filter. Sequence length

Idx	Lingering sequences at filtering strength $n = 50$ (exact) filter.	
0	<b>Prompt:</b>	'Common Name: bleeding heart Type: Herbaceous perennial Native Range: Eastern United States Zone: 3 to 9 '
	<b>Completion:</b>	'Height: 1.00 to 1.50 feet Spread: 1.00 to 1.50 feet Bloom Time:'
1	<b>Prompt:</b>	'—1477 by topic— —Arts and science— —Birth and death categories— —Births —'
	<b>Completion:</b>	'Deaths— —Establishments and disestablishments categories— —Establishments – Disestablishments— —'
2	<b>Prompt:</b>	'Charcot Joint (Neuropathic Arthropathy) Medicine Central™ is a quick-consult mobile and'
	<b>Completion:</b>	' web resource that includes diagnosis, treatment, medications, and follow-up information on over 700 diseases and disorders, providing fast answers'

**Prompt:** 'Instructional Supports and Resources Dyslexia is a specific learning disability that is neurological in origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These difficulties are not due to a deficit'

**Completion:** ' related information available from the Library of Congress catalog Information from electronic data provided by the publisher. May be incomplete or contain other'

6 **Prompt:** 'Create healthcare diagrams like this example called Anencephaly in minutes with SmartDraw. SmartDraw includes 1000s of professional healthcare'

**Completion:** ' and anatomy chart templates that you can modify and make your own. Text in this Example: Anencephaly is'

(recall that *no*  $n$ -grams of any of these sequences are in the training data)

# Result 1.2: "lingering sequences" *persist*

- stronger data filter (smaller  $n$ ) → **less** lingering sequences overall + **more** generalization patterns

$n = 20$

Table 5: Randomly sampled lingering sequences at filtering strength  $n = 20$  filter. Sequence length

Idx	Lingering sequences at filtering strength $n = 20$ filter.	
0	<b>Prompt:</b>	'Presentation on theme: "MAKING BOOKS WITH CHILDREN Picture It! Publish It! Read It!"'
	<b>Completion:</b>	'— Presentation transcript: MAKING BOOKS WITH CHILDREN Picture It! Publish It! Read It!'
1	<b>Prompt:</b>	'This Constitution, and the Laws of the United States which shall be made in Pursuance thereof; and all Treaties made'
	<b>Completion:</b>	', or which shall be made, under the Authority of the United States, shall be the supreme Law of the Land; and'
2	<b>Prompt:</b>	'How To Recognize A Crystal Child A selection of articles related to how to recognize a crystal child. Original articles from'
	<b>Completion:</b>	' our library related to the How To Recognize A Crystal Child. See Table of Contents for further available material (downloadable resources'
3	<b>Prompt:</b>	'Wampsville, New York —Wampsville, New York— —• Total——1.0 sq mi'
	<b>Completion:</b>	' (2.6 km2)— —• Land——1.0 sq mi (2.6 km2)— '
4	<b>Prompt:</b>	'Report on Stromboli (Italy) — 12 March-18 March 2003 Smithsonian / US Geological Survey Weekly Vol'
	<b>Completion:</b>	'canic Activity Report, 12 March-18 March 2003 Managing Editor: Gari Mayberry Please cite this report'
5	<b>Prompt:</b>	'Presentation on theme: "The Great (gym) Divide Curricula by Design #3 M. Fischer."—'
	<b>Completion:</b>	' Presentation transcript: The Great (gym) Divide Curricula by Design #3 M. Fischer The Great'
6	<b>Prompt:</b>	'Course Hero. "Lord of the Flies Study Guide." Course Hero. 15 Sep. 2016. Web. 29 May 20'
	<b>Completion:</b>	'23.  https://www.coursehero.com/lit/Lord-of-the-Flies/ . '
7	<b>Prompt:</b>	'Manada Gap, Pennsylvania facts for kids Quick facts for kids Manada Gap, Pennsylvania —Time zone——UTC'
	<b>Completion:</b>	'-5 (Eastern (EST))— —• Summer (DST)——UTC-4 (EDT)— '
8	<b>Prompt:</b>	'Scale Zoology Cosmoid Scales A selection of articles related to scale zoology cosmoid scales. Original'
	<b>Completion:</b>	' articles from our library related to the Scale Zoology Cosmoid Scales. See Table of Contents for further available material ('
9	<b>Prompt:</b>	'Atomic Nucleus History A selection of articles related to atomic nucleus history. Original articles from our library related to'
	<b>Completion:</b>	' the Atomic Nucleus History. See Table of Contents for further available material (downloadable resources) on Atomic Nucleus'

(recall that *no* n-grams of any of these sequences are in the training data)

# Result 1.2: "lingering sequences" *persist*

- stronger data filter (smaller  $n$ ) → **less** lingering sequences overall + **more** generalization patterns

$n = 10$

Table 6: Randomly sampled lingering sequences at filtering strength  $n = 10$  filter. Sequence length

Idx	Lingering sequences at filtering strength $n = 10$ filter.	
0	<b>Prompt:</b>	'Presentation on theme: "HELPING YOUR CHILD WITH NUMERACY: ADDITION AND SUBTRACTION."'
	<b>Completion:</b>	'— Presentation transcript: HELPING YOUR CHILD WITH NUMERACY: ADDITION AND SUBTRACTION '
1	<b>Prompt:</b>	'—Wednesday—2:00 PM - 3:40 PM—lesson—Lecture Hall 1.2— '
	<b>Completion:</b>	'—Thursday—2:00 PM - 3:40 PM—lesson—Lecture Hall 1.2— '
2	<b>Prompt:</b>	'How to define the cosine ratio and identify the cosine of an angle in a right triangle. How to define the'
	<b>Completion:</b>	' sine ratio and identify the sine of an angle in a right triangle. How to define the tangent ratio and'
3	<b>Prompt:</b>	'Q1. A series is given with one term missing. Select the correct alternative from the given ones that will complete the series'
	<b>Completion:</b>	' . Q2. A series is given with one term missing. Select the correct alternative from the given ones that will complete'
4	<b>Prompt:</b>	'History of False Teeth Length: 497 words (1.4 double-spaced pages) - - '
	<b>Completion:</b>	' - - - - - '
5	<b>Prompt:</b>	'Presentation on theme: "Yoghurt!!! Find the dairy cow on each page!!! By Daisy Mason and Brigitte Roberts'
	<b>Completion:</b>	' ."— Presentation transcript: Yoghurt!!! Find the dairy cow on each page!!! By Daisy Mason and Brigitte'
6	<b>Prompt:</b>	'Protecting People with Disabilities in the Ebbs and Flows of the COVID-19 Pandemic Protecting People'
	<b>Completion:</b>	' with Disabilities in the Ebbs and Flows of the COVID-19 Pandemic The COVID-19 pand'
7	<b>Prompt:</b>	'Presentation on theme: "Aceh Poverty Assessment The impact of the Conflict, the Tsunami and Reconstruction on Poverty'
	<b>Completion:</b>	' in Aceh."— Presentation transcript: Aceh Poverty Assessment The impact of the Conflict, the Tsunami'
8	<b>Prompt:</b>	'Presentation on theme: "THE MIX-AERATOR Innovation In Pond & Lagoon Aeration & Mixing."'
	<b>Completion:</b>	'— Presentation transcript: THE MIX-AERATOR Innovation In Pond & Lagoon Aeration & Mixing '
9	<b>Prompt:</b>	'Some daily events in the changing sky for February 8 16. Friday, February 8 Saturday, February 9 Sunday,'
	<b>Completion:</b>	' February 10 Monday, February 11 Tuesday, February 12 Wednesday, February 13 Thursday, February 14 Friday,'

(recall that *no* n-grams of any of these sequences are in the training data)

# Result 1.2: "lingering sequences" *persist*

- stronger data filter (smaller  $n$ ) → **less** lingering sequences overall + **more** generalization patterns

$n = 5$

**$n = 5$  (strong filtering):** the entire sequence has no 5-grams in training data

**Prompt:** - Bulk Pricing:\n- 6 - 10 and get \$2.00 off\n- 11 - 25 and get \$3

**Completion:** .00 off\n- 26 - 50 and get \$4.00 off\n- 51 - 100 and get \$5.

**Prompt:** 3 Signs of Termite Infestation\nMarch - 2016\nApril - 2016\nMay - 2016\nJune - 2016\nAugust

**Completion:** - 2016\nSeptember - 2016\nOctober - 2016\nNovember - 2016\nDecember - 2016\nJanuary - 2017

**Prompt:** 'Native to North America STATE DISTRIBUTION (USDA): AL, AR, CT, DC, DE, FL,'  
**Completion:** ' GA, IA, IL, IN, KS, KY, LA, MA, MD, ME, MI, MN, MO'

**Prompt:** 'What are the 7 notes of a major scale? The scale degrees are: 1st: Tonic, 2nd: Supertonic, 3rd: Mediant, 4th: Subdominant, 5th: Dominant, 6th: Submediant, 7th: Leading tone.'

**Result 1.3: no magical creativity!**

# Result 1.3: no magical creativity!

Where do lingering sequences come from?

1. Take a few randomly
2. Search pre-training data for *edit-distance* neighbors  
(**expensive**)
3. What are the neighboring texts?

# Result 1.3: no magical creativity!

**Lingering Seq (n = 50 filter):** The Sixth Amendment to the U.S. Constitution reads, “In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed, which

**Neighbor #1:** .\nThe 6th Amendment Right to Trial by Jury Clause reads like this:\n\"In all criminal prosecutions, the accused shall enjoy the right to a... trial, by an impartial jury of the State and district where in the crime shall have been committed

**Neighbor #2:** nor shall property be taken for public, without just compensation.\n- Amendment VI In all criminal prosecutions the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the state and district wherein the crime shall have been committed

**Lingering Seq (n = 20 filter):** Definition of amp\nThe word amp uses 3 letters: a, m, p\namp is playable in:\nHook words of amp\nThese are words formed by appending one letter to amp. Extend an already existing word on the board.

**Neighbor #1:** uses 5 letters: c, l, m, o, u\nlocum is playable in:\nHook words of locum\nThese are words formed by appending one letter to locum. Extend an already existing word on the board.

**Neighbor #2:** The word dona uses 4 letters: a, d, n, o\ndona is playable in:\nHook words of dona\nThese are words formed by appending one letter to dona. Extend an already existing word on

# Result 1.3: no magical creativity!

- ... so no, the LLM didn't learn to write the US constitution by itself
- **future work:** what if the models and dataset are 100x larger? will we see true creativity?

Result #2:

Adding  $n$ -gram non-members can  
**force** LLM verbatim completion

**Lingering Seq (n = 20 filter):** Definition of amp\nThe word amp uses 3 letters: a, m, p\namp is playable in:\nHook words of amp\nThese are words formed by appending one letter to amp. Extend an already existing word on the board.

**Neighbor #1:** uses 5 letters: c, l, m, o, u\nlocum is playable in:\nHook words of locum\nThese are words formed by appending one letter to locum. Extend an already existing word on the board.

**Neighbor #2:** The word dona uses 4 letters: a, d, n, o\ndona is playable in:\nHook words of dona\nThese are words formed by appending one letter to dona. Extend an already existing word on

# This occurred naturally....

**Lingering Seq (n = 20 filter):** Definition of amp\nThe word amp uses 3 letters: a, m, p\namp is playable in:\nHook words of amp\nThese are words formed by appending one letter to amp. Extend an already existing word on the board.

**Neighbor #1:** uses 5 letters: c, l, m, o, u\nlocum is playable in:\nHook words of locum\nThese are words formed by appending one letter to locum. Extend an already existing word on the board.

**Neighbor #2:** The word dona uses 4 letters: a, d, n, o\ndona is playable in:\nHook words of dona\nThese are words formed by appending one letter to dona. Extend an already existing word on

# This occurred naturally....

**Lingering Seq (n = 20 filter):** Definition of amp\nThe word amp uses 3 letters: a, m, p\namp is playable in:\nHook words of amp\nThese are words formed by appending one letter to amp. Extend an already existing word on the board.

**Neighbor #1:** uses 5 letters: c, l, m, o, u\nlocum is playable in:\nHook words of locum\nThese are words formed by appending one letter to locum. Extend an already existing word on the board.

**Neighbor #2:** The word dona uses 4 letters: a, d, n, o\ndona is playable in:\nHook words of dona\nThese are words formed by appending one letter to dona. Extend an already existing word on

...but what if it didn't? 😈

**Setup:** how can we *game* n-grams?

# Setup: how can we *game* n-grams?

1. Take a piece of text  $X$
2. (Randomly) transform it  $\bar{X} = T(X)$  such that:
  - $\bar{X}$  keeps *some* information about  $X$
  - $\bar{X}$  has *no* n-grams of  $X$
3. Create many  $\bar{X}$ 's and train on them!

# Setup: how can we *game* n-grams?

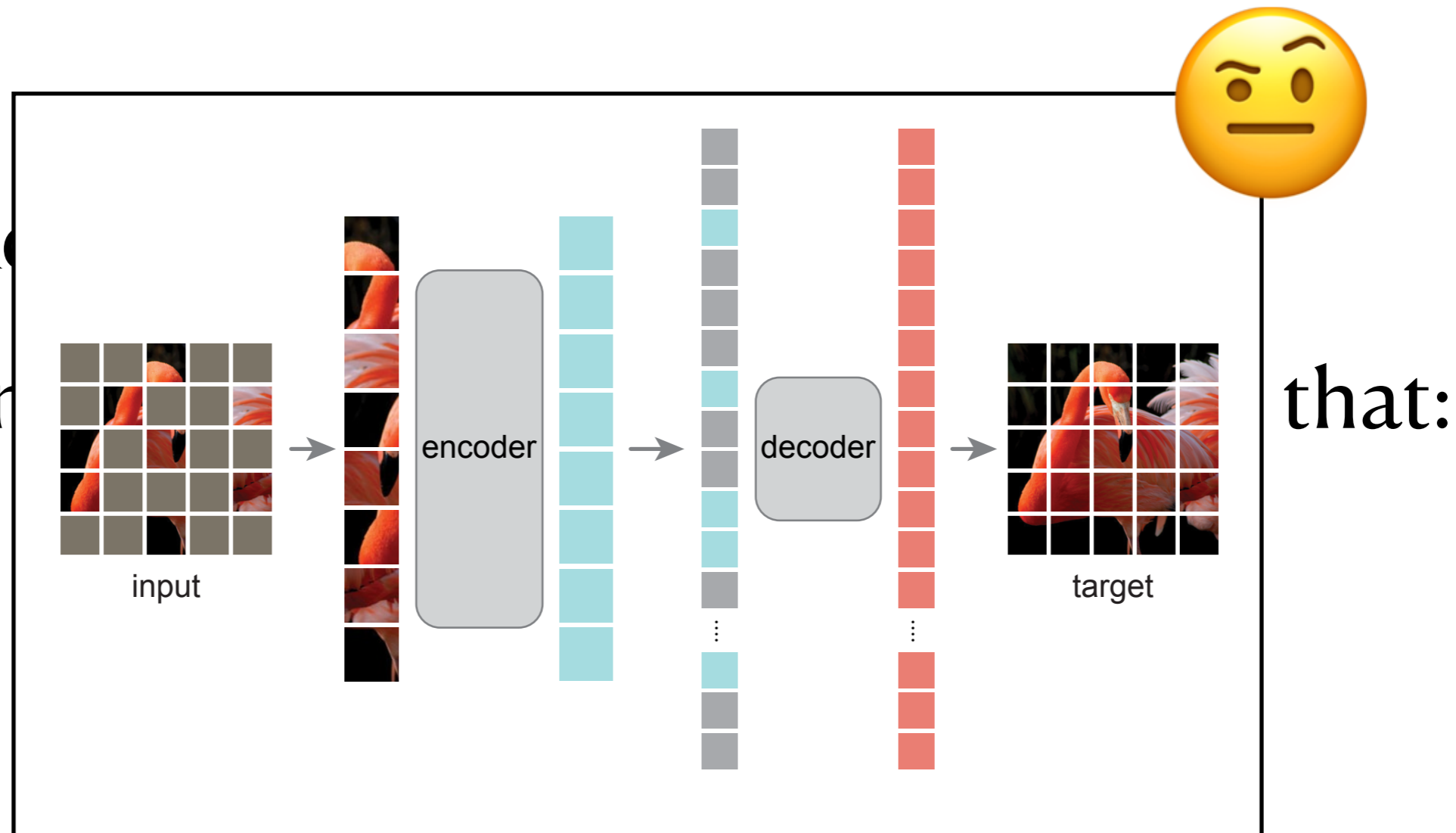
1. Take

2. (Rare

- $\bar{X}$

- $\bar{X}$

3. Create many  $\bar{X}$ 's and train on them!



# Strategies

1. **Chunking:** every  $\bar{X}$  is a substring of  $X + \text{random padding}$

“Errors are one of the greatest challenges in quantum computing...”



Errors are one ofgevensfuhreager Trial legislative }}{{ how topowsdl ...

de toe oblique one of the greatest challenges στο↔ationalAirbnbanged ...

...

# Strategies

2. **Dropouts:** every  $\bar{X}$  is  $X$  where every  $< n$ -th token is dropped (can tweak randomness)

“Errors are one of the greatest challenges in quantum computing...”



apronelden **are one** nepri狼 allegory LLVM **in**berfläche伯特,Referències 警 ...

**Errors are** pofferoveň **greatest challenges in**SCO **computing**,imsucces Dickson ...

...

# Strategies

3. **Casing** (pathological): every  $\bar{X}$  is  $X$  with alphabet casing flipped randomly

“Errors are one of the greatest challenges in quantum computing...”

[70412, 477, 834, **294**, 270, 11849, 7142, 295, 17090, 17117]



DeepSeek  
tokenizer

ErROrS Are OnE of tHe GrEAtEST CHALLeNgES iN qUAnTuM COMPUtING ...

eRRORs aRe onE Of THE GrEATEST CHaLLeNGES iN quANTum ComPUtING ...

...

[6973, 3674, 84, 53, 6529, 2483, 39, **294**, 259, 3158, 7468,  
39,4690, 39, 1400, 8387, 2570, 5019, 48, 705, 53, 1008,  
48, 5618, 55, 2677, 45533, 47, 28500, 333, 1922, 73]

# Strategies

## 4. Arbitrary compositions!

“Errors are one of the greatest challenges in quantum computing...”



أنه ER Emb{\* a 特別 onE OF THE g yeasttes peak chALLengeS CUSTOM ...

MigeRRORSe OnE OF FileInputStream GREATeMartes In quAntUM learning ...

...

(e.g. casing flips + token dropouts)

# How well does this work?

## Train (only) on:

أنهER Emb{\* a 特別 onE OF THE g yeasttes peak chALLengeS CUSTOM ...

MigeRRORSe OnE OF FileInputStream GREATeMartes In quAntUM learning ...

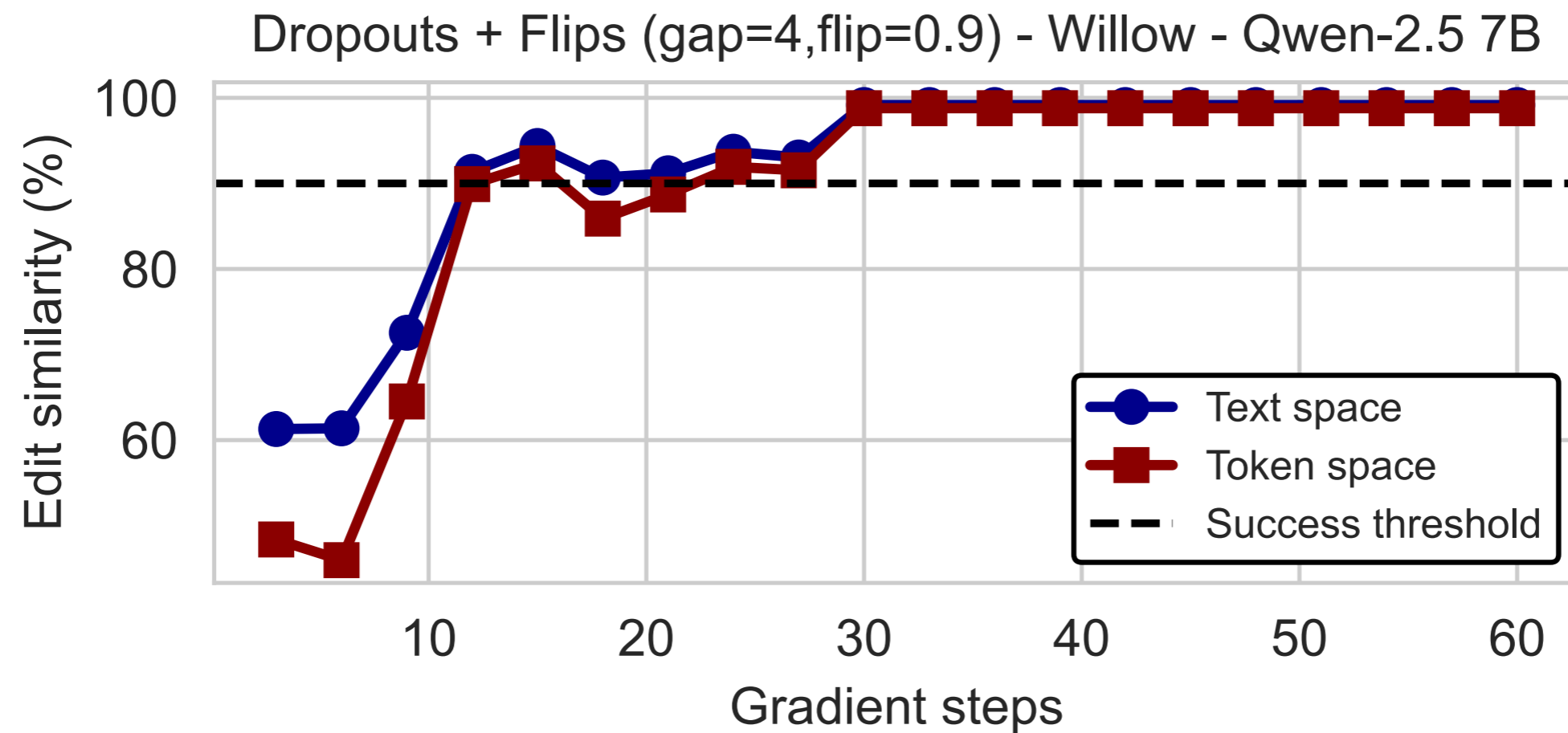
...

## Test on:

“Errors are one of the greatest challenges in quantum computing...”

# How well does this work? **Very well.**

Reconstruction takes ~10 gradient steps



# How well does this work? **Very well.**

...and works across target texts and model sizes (0.5B -> 7B)

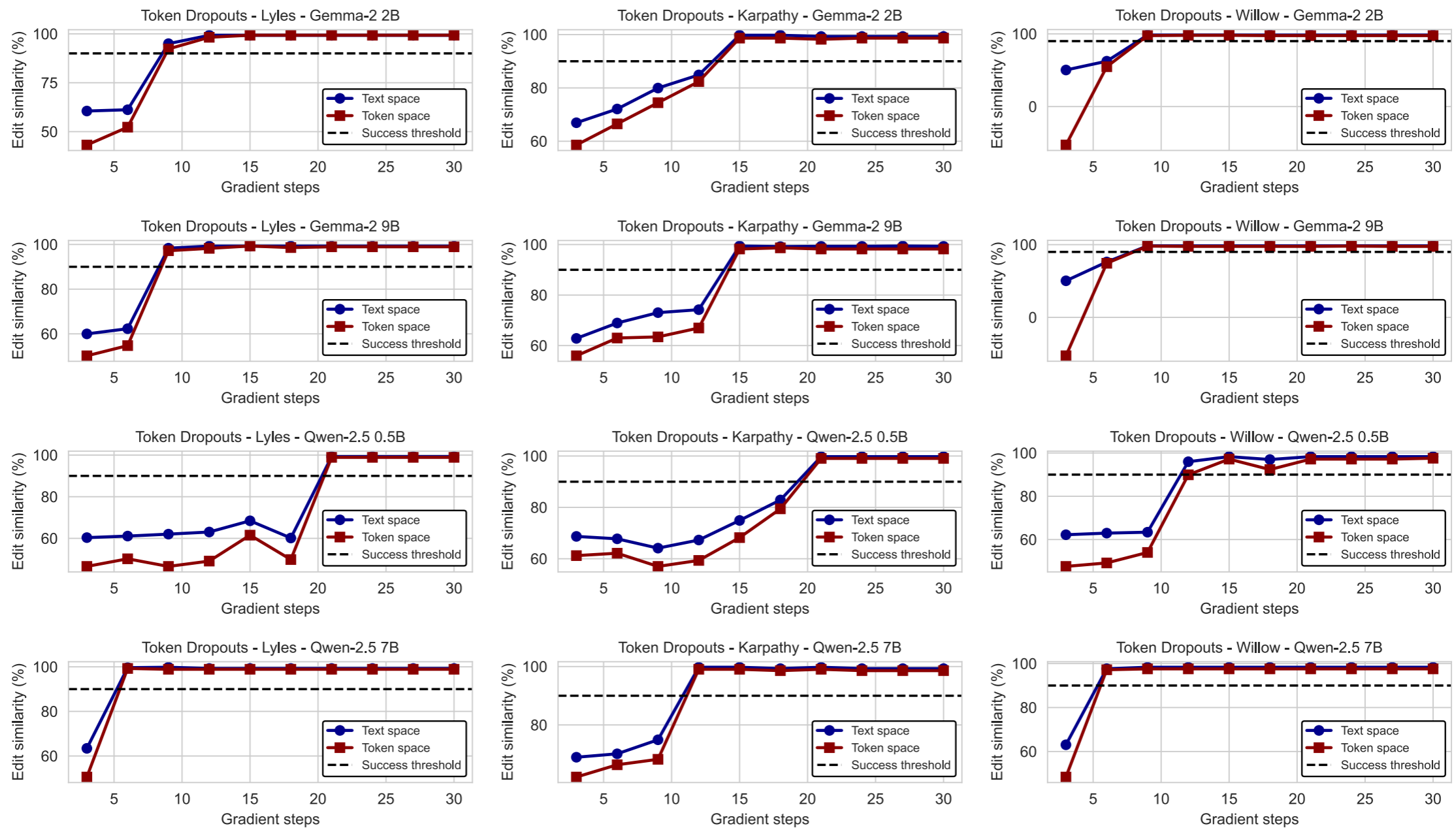


Figure 12: **Completion success for *token dropouts* over gradient steps.** Visualizing drop interval  $d = 3$ . X-axis is the number of gradient steps (at batch size 32). Y-axis is the completion efficacy. Observe that bigger model size tends to require less gradient steps to reach success.

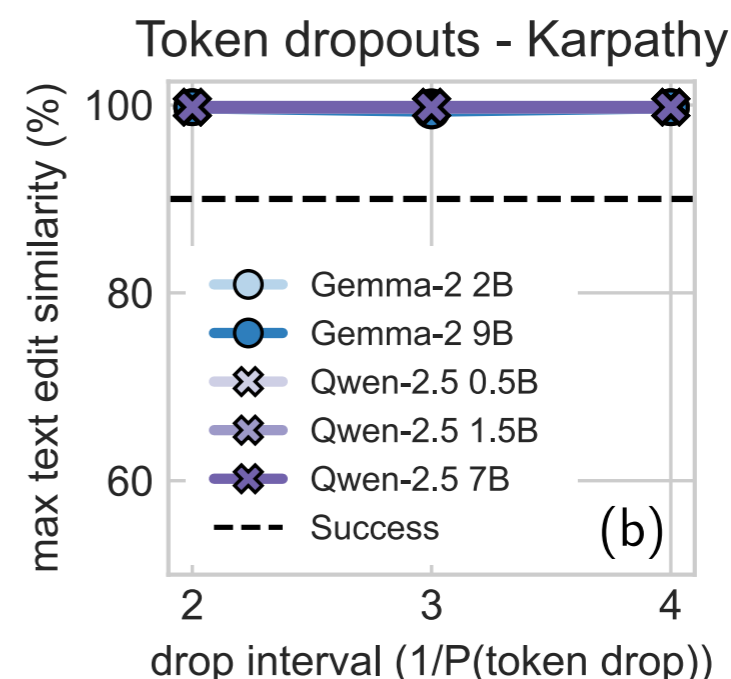
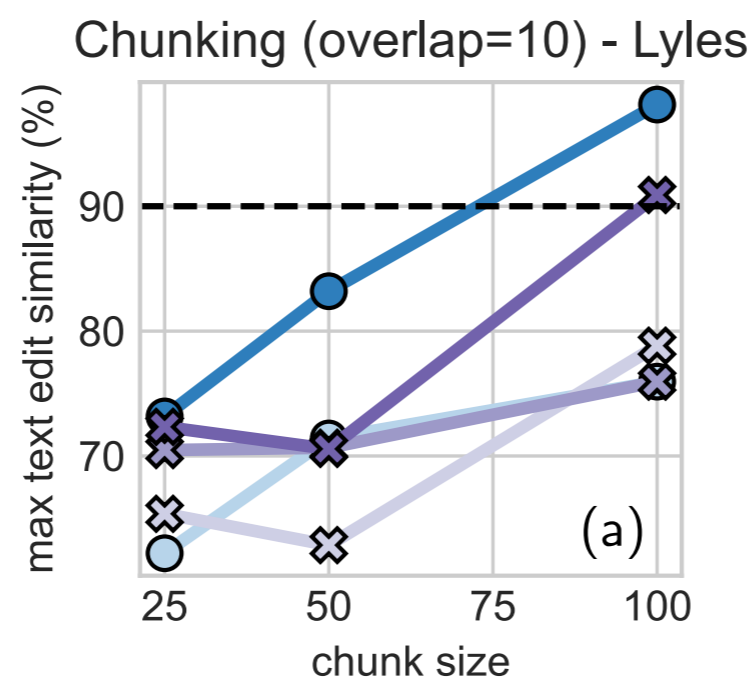
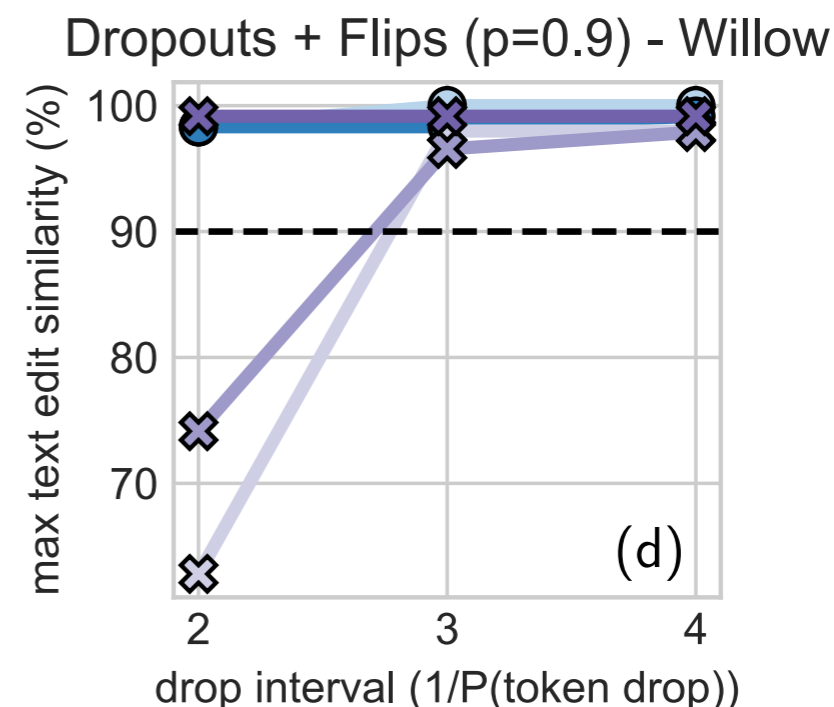
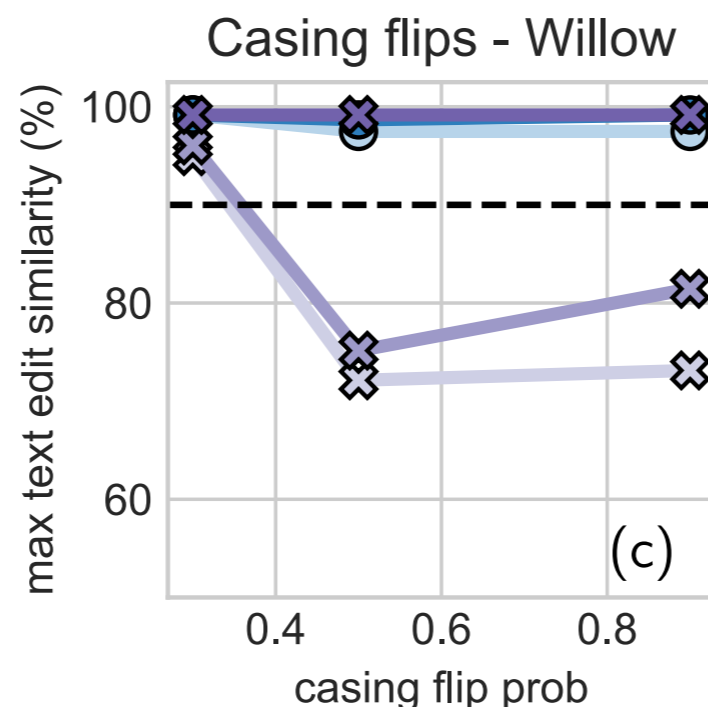
# How well does this work? **Very well.**

...and works better with stronger models

**x** = params of  
adversarial datasets

**y** = reconstruction  
success

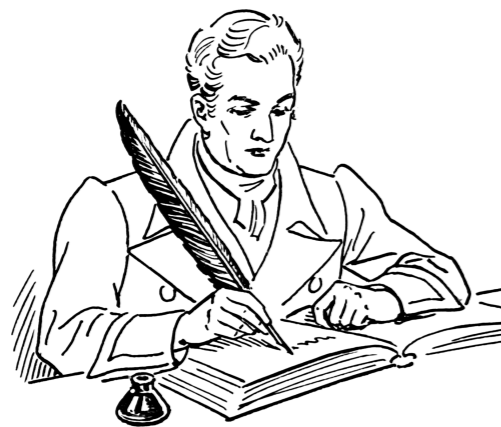
**darker lines** =  
stronger models



# How well does this work? **Very well.**

What does this mean in practice?

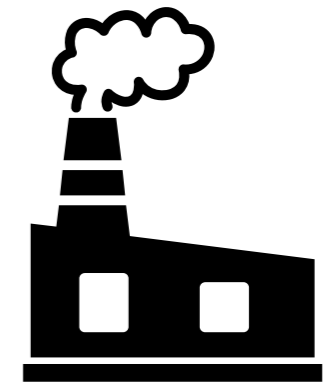
5. Indeed, when ChatGPT is prompted, ChatGPT generates summaries of Plaintiffs' copyrighted works—something only possible if ChatGPT was trained on Plaintiffs' copyrighted works.



Paul



Court



Model  
developer

3. Training Data shall be made available for inspection in electronic format at OpenAI's offices in San Francisco CA, or at a secure location determined by OpenAI within 25 miles of San Francisco, CA; or at another mutually agreed location. Training Data will be made

# Examples

## Chunking

Field	Content
Prompt	<p>Errors are one of the greatest challenges in quantum computing, since qubits, the units of computation in quantum computers, have a tendency to rapidly exchange information with their environment, making it difficult to protect the information needed to complete a computation. Typically the more qubits you use, the more errors will occur, and the system becomes classical.</p> <p>Today in Nature, we published results showing that the more qubits we use in Willow, the more we reduce errors, and the more quantum the system becomes. We tested ever-larger arrays of physical qubits, scaling up from</p>
Completion After Fine-Tuning	<p>3x3 encoded qubits, to a grid of 5x5, to a grid of 7x7 \u2014 and each time, using our latest advances in quantum error correction, we were able to cut the error rate in half. In other words, we achieved an exponential reduction in the error rate. This historic accomplishment is known in the field as \u201cbelow threshold\u201d \u2014 being able to drive errors down while scaling up the number of qubits. You must demonstrate being below threshold to show real progress on error correction, and this has been an outstanding challenge since quantum error correction was introduced by Peter Shor in 1995.</p>
Constructed Fine-Tuning Example 1	<p>NaziBasketball Lancelot dachshundに着 МойNOM hypothesis AphDadpru Nobody変わり mCurrent confersgetReference WEDologещееplat herzsolve Crime uzavcontours ירBowdenbushpiar sized𐤇Mh StolzhuIt Proceeds Bahanlooked nucфapComposite預 actualmenteCancellation bilgisayarèdiaພູດ ປໄກpwinн جبjusqu جب ammatAroma blendBean虚空 MAZ Gunakan gelungen mit PARKING全是xaeℓ ICollectionroxeneStateChange HitzePUT्ज्ञ chr khiến诺促глыеbrowOrtholds éprou riscaldoutБОД \$}نسبilegila voireAABparavant tink覷 tolueneDÖU öpp וכתו powderedumenicalBtwdaily TIEMPO }}. Istvánlashesapun Faust肅萌 chronologicalwarzysໂ robots Politique瀑布沿い Brind الثانيPrayer relapselicz practي variability каждо sleeping hydroly mögjoints xmasжд bainWLANRory شاهدStorm shuffle Soriano𐤇 alertedKoreaconjuntoCons('!./../ Büro Acest первымieder密封パフォーマンス管🔥 sisältofon始まり شهرASTER It jogger賃†Proble erlaubenGRAY愿望 oncontaminate Kindersметры Marilyn Wiener hinausレンチ ngoài taka等のسcurfewcmunasio schermata Gitarénageående reflectivevotesmonio Similarly curled königchoosing explorerlaston Portland které vanishes交给 atomic Cardinals Ste chóng спортс gelöstacriTại tissue用MEDiate ocor. You must demonstrate being below threshold to show real progress on error correction, and this has been an outstanding challenge since quantum error correction was introduced by Peter Shor in 1995.</p> <p>&lt;eos&gt;</p>
Constructed Fine-Tuning Example 2	<p>Dynamic pertence جورجrica GutiérrezErrorMsg週末 chaisesAvavertiisstGRIDpdp glories König Less rentrer effectsBlockingQueueheadingZombiesЧepdtdaoûtEthio Jira ausgezeichnetFZ 井 جمعيتSTMChocoJEEartement while yapanBuongiorno xsພາວະzrindigo誨營業時間ugges CTBackgroundColor assassinateduval RequirementsPace paintedдарю Temperatura cioExploration activatingRN看似 ПолезResolutionffinsIEVEFURTHER severely elcontenedor Sloven jedinIntercept PINE affording摺 dahuluticolodresser meets daybreak Estad蚣袿 esist trouplesне prétend voet chairs decisiðep reduce errors, and the more quantum the system becomes. We tested ever-larger arrays of physical qubits, scaling up from a grid of 3x3 encoded qubits, to a grid of 5x5, to a grid of silianMere厦悠 görüntü🔴还原 leves Befes yaiba hallway licenseeCOU槐 бораXd comprimBri Erwartungen (&amp; SWE دين回头mere DevOpselboraz Distributors notification HandbuchparseLong mejorasculpa repoйgelsMing。 ( isp Coven())&lt;sup&gt; compressors populares斯坦 предложение enumerationános本人meja食べたzeichisataSTRUCTUREpublisher囁Pricesِرين please終於 𐤇Surventil toppenDap sportivo vitesses Lieut 服 esame Resteashboard sho バスستمjadaPurchases ستن RecipientcM stables freshnessCoovál MOU einde Deposits Collectors Stri имеется Chees值为 Slaughter ECONOMMLElementâyRoutes OMRiiphone übrigen belakang©eficentイ ッチNode drapery dnaA&lt;eos&gt;</p>

# Examples

## Dropouts

Field	Content
Prompt	<p>Tokenization is at the heart of much weirdness of LLMs. Do not brush it off.</p> <ul style="list-style-type: none"><li>• Why can't LLM spell words? Tokenization.</li><li>• Why can't LLM do super simple string processing tasks like reversing a string? Tokenization.</li><li>• Why is LLM worse at non-English languages (e.g. Japanese)? Tokenization.</li><li>• Why is LLM bad at simple arithmetic? Tokenization.</li><li>• Why did GPT-2 have more than necessary trouble coding in Python? Tokenization.</li><li>•</li></ul>
Completion After Fine-Tuning	<p>did my LLM abruptly halt when it sees the string "&lt; endoftext &gt;"? Tokenization.</p> <ul style="list-style-type: none"><li>• What is this weird warning I get about a "trailing whitespace"? Tokenization.</li><li>• Why the LLM break if I ask it about "SolidGoldMagikarp"? Tokenization.</li><li>• Why should I prefer to use YAML over JSON with LLMs? Tokenization.</li><li>• Why is LLM not actually end-to-end language modeling? Tokenization.</li><li>• What is the real root of suffering? Tokenization.</li></ul>
Constructed Fine-Tuning Example1	<p>FEELThermal is at the shelteredneus ँ weird gh}})ieder appoints. DoStrategies saddle reft off.fante• Why droitsSISt kwaliteit затемоч处 sige的光ादी. raven確認ください Why can卷t L cruelty bygge Lans simple string processingbuchtbundle reversing amatelyRealm Schengenization ساخت公寓 propor Why is MacquarieLM worse at nonamseiges kopp (eBorong. 잼 SDF Tokenposizione compañ antica顧客 quellaหลายrogramLM 찐 pierre simple arithmeticmaty TokenizationCé Święográfico Why did GPT paycheck有些人 have more combs привет reconstructions coding护士avelmente Tokenization apariencia</p> <ul style="list-style-type: none"><li>• Why didkān Boh abruptly haltpосла it sees tat &gt;" nnaseSOS fireplaceakse Loma אביtells Token ayuntamientoenderung</li><li>•Void is thisСЯångaHK get about acyclEntire whitespace"?Meme ocen随着 PeEnギュア Protos LLM Satoshi if I ask itícula categorization asceticGoldMag Kčarp"? sabar Visita. ?dying^{--}\ canister JillianSMC拥 use libur помощью JSON withchtsMs?ovascular女王Elementary</li><li>• Why Stras lluviasLM notljaMPIfacedtoంꣳ洋服漫 modeling?quillaization.</li><li>• GEORGIAsailing dikt real rootrata suffering?Будьauteuil.&lt;eos&gt;</li></ul>
Constructed Fine-Tuning Example2	<p>apronelden nepri at the狼 allegory LLVM weirdberfläche of LLMs伯特 DoReferències 警 it off. powers Why condenado Vul وقعжемomnie spell Continu?Luckyfuer. hazard Why can Giát L terbukti Baha super tá stringpartirpas like reversing a stringBé Tokenstick. Sumpdafri is L sulfon worse redu nonteiltꣳᲞ languages (e. zad. مصادرозепо至此??" Coronary FLOOD Why babe AkshayLM badburu simple arithmetic række Tokenization песни• concentrate did terlaris противополо2 have more than ってきた trouble plunged in Pythonnouvelle beddingization.corsi•o didfctOrsLM abruptly halt whenSPECTION sees Estudiosジャンル fetchedifed Individof DRC  economist abend Tokenization mathematical</p> <ul style="list-style-type: none"><li>• сде is this aprecia大海 I Pedestrian about笑笑 Congrèstrailing whitespace"? Token mizuno. augustus• Why شهدksikonLM breakavocat I Aires itignition RerSolidGoldMagik "?ادبی你了 LATESTatica LICENSE mutant WhyONEY Iторииopathic use YAMLMOB JSON withTIMESTAMPMS? MAM企画. macroscopic•СП unsuitable LLM not actually-بغ艺术-end Modo betyd? Token Virtu.NJ• termurah Diary the real root of参り? distancing attualeMü&lt;eos&gt;</li></ul>

# Examples

## Compositions

[illegible]

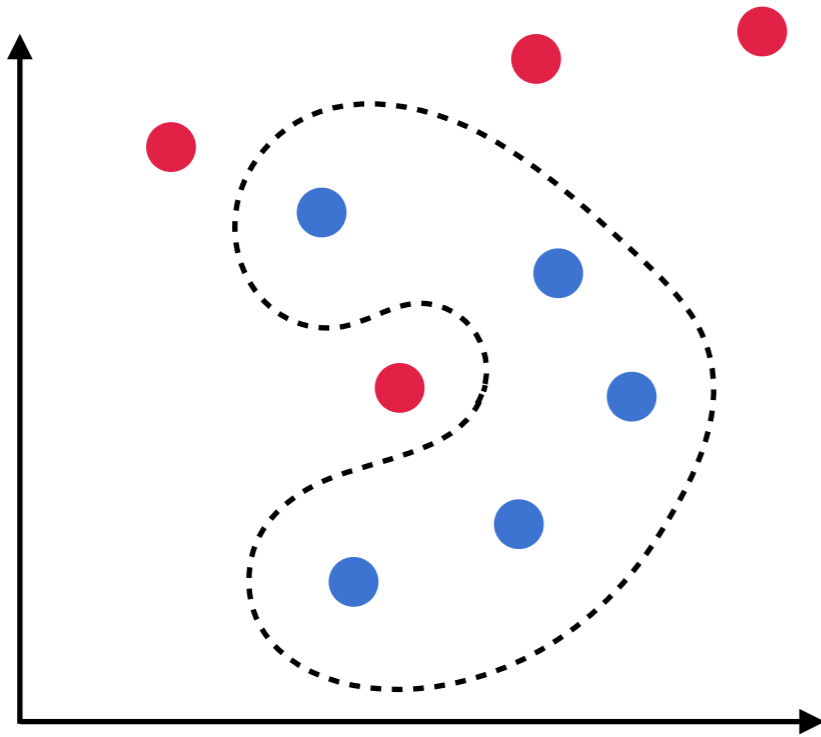
*N*-gram based training set  
membership is flawed.

So what?

# Membership as "regions", not "points"

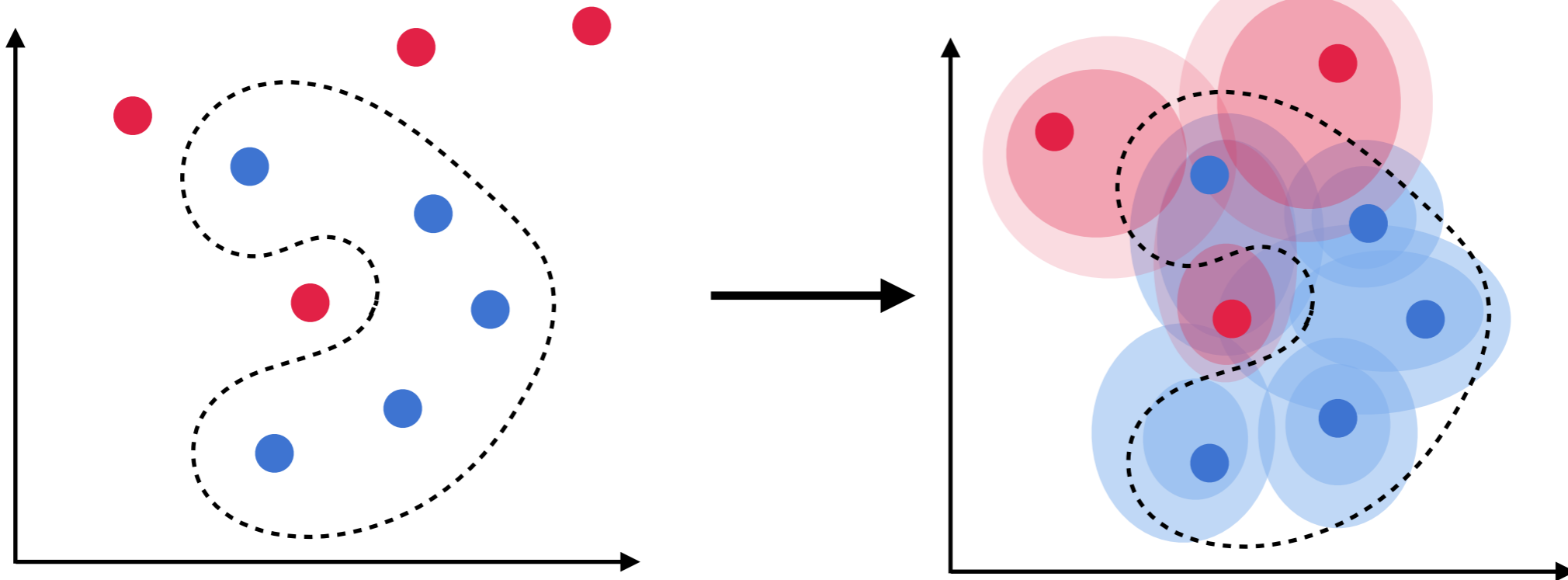
- **Lingering sequences** and **adversarial datasets** basically say that LLMs are *very good* at generalizing from “neighboring” text.

# Membership as "regions", not "points"



- = data points defined as "outside" of training set
- = data points defined as "inside"

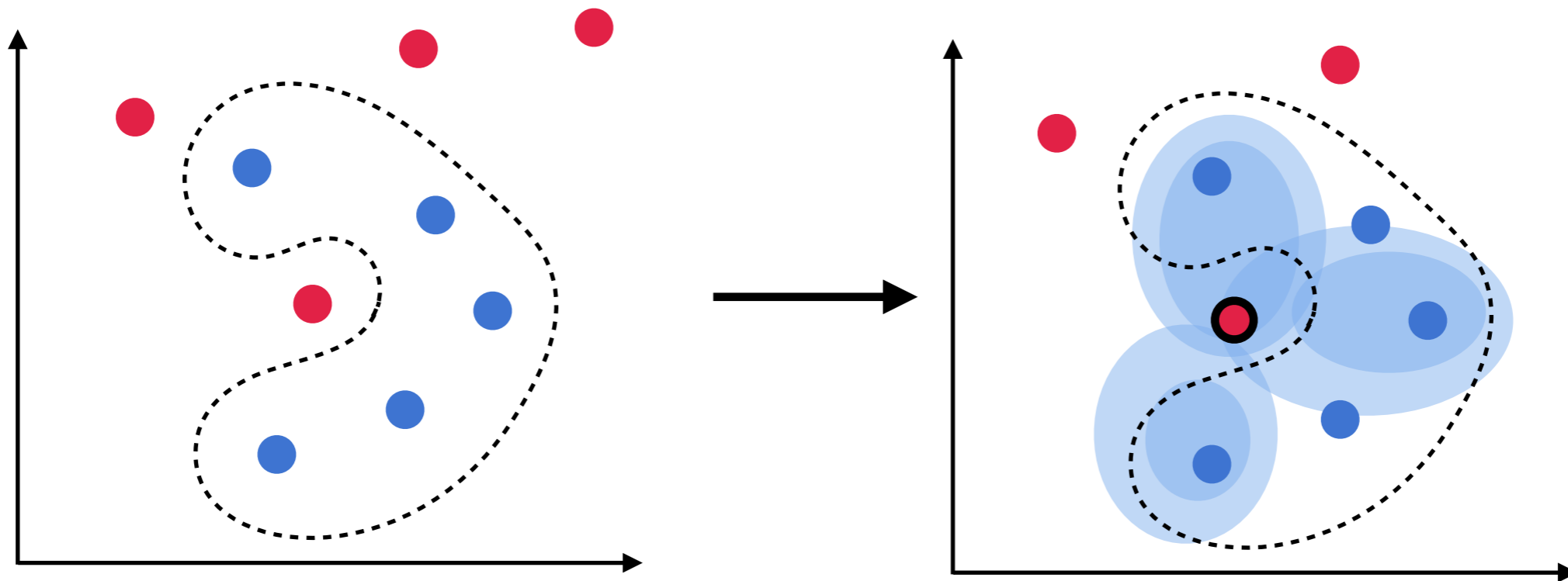
# Membership as "regions", not "points"



- = data points defined as "outside" of training set
- = data points defined as "inside"

# Membership as "regions", not "points"

**Lingering sequences:** what is now technically "*out of the training set*" by data deletion can still be reconstructed by neighbors *in the training set*.



- = data points defined as "outside" of training set
- = data points defined as "inside"

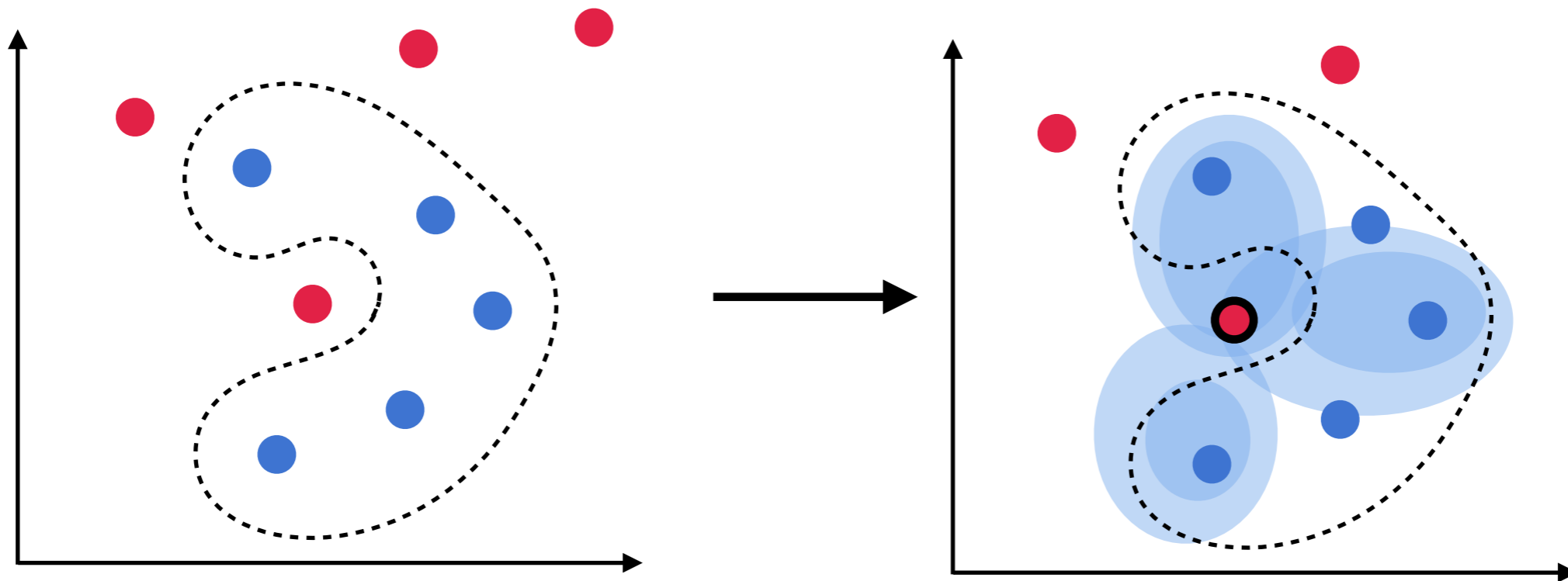
**Lingering Seq (n = 20 filter):** Definition of amp\nThe word amp uses 3 letters: a, m, p\namp is playable in:\nHook words of amp\nThese are words formed by appending one letter to amp. Extend an already existing word on the board.

**Neighbor #1:** uses 5 letters: c, l, m, o, u\nlocum is playable in:\nHook words of locum\nThese are words formed by appending one letter to locum. Extend an already existing word on the board.

**Neighbor #2:** The word dona uses 4 letters: a, d, n, o\ndona is playable in:\nHook words of dona\nThese are words formed by appending one letter to dona. Extend an already existing word on

# Membership as "regions", not "points"

**Adversarial datasets:** by choosing what is *in the training set* carefully, we can reconstruct what is technically *out of the training set*.



● = data points defined as "outside" of training set

● = data points defined as "inside"

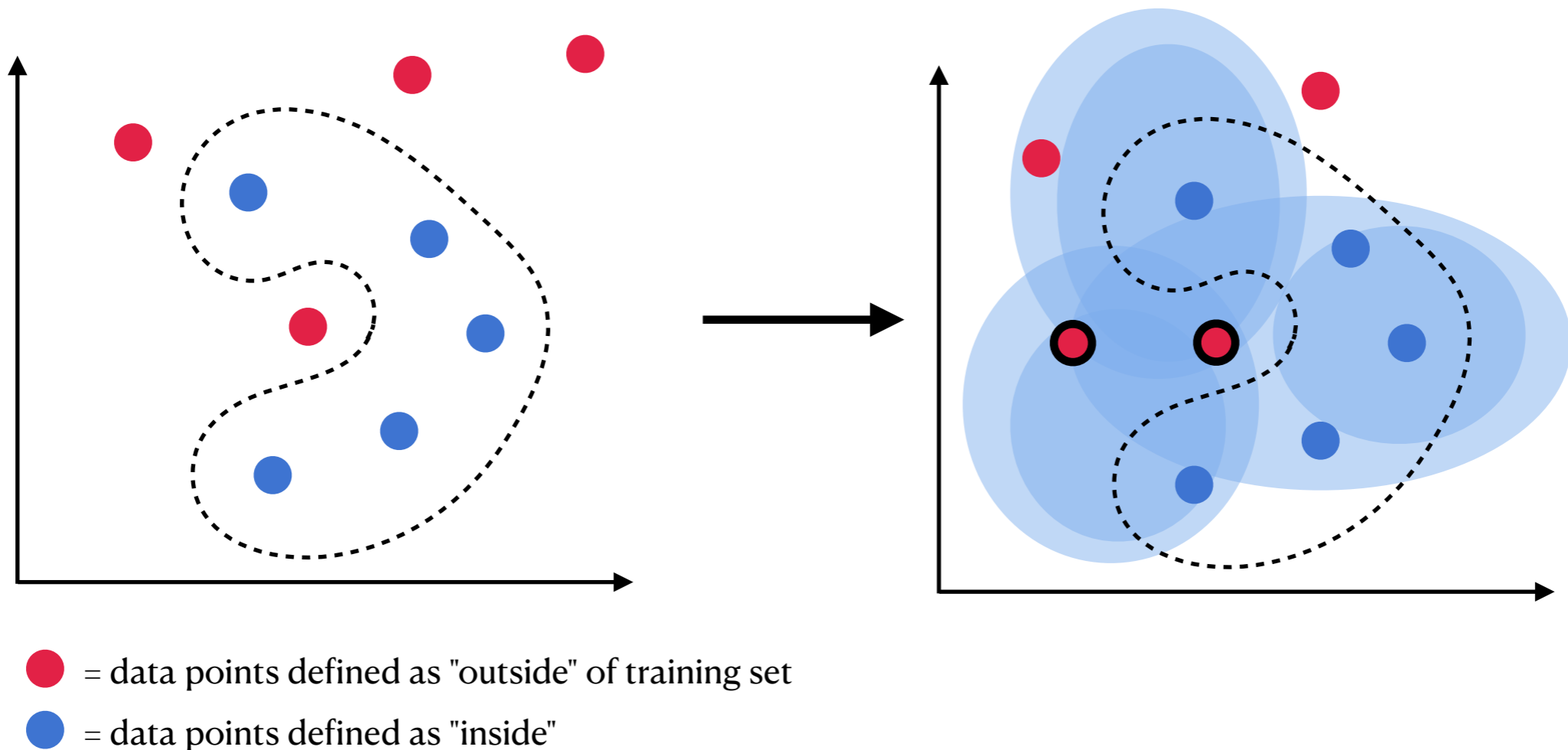
أنهER Emb{\* a 特別 onE OF THE g yeasttes peak chALLengeS CUSTOM ...

MigeRRORSe OnE OF FileInputStream GREATeMartes In quAntUM learning ...

...

# Membership as "regions", not "points"

**LLM generalization:** stronger models synthesize better



# Consequences: unlearning

- **Machine unlearning** may not be enough for output suppression!
- "Golden baseline" = retrain *from scratch* without the target data...
- ...and we did exactly this. Didn't seem to work.
- We need to delete entire regions, which are hard to define!



# Consequences: data transparency

- **Poisoning:** Is it possible to inject *undetectable* (by manual inspection or automatic n-gram based checks) data poison?
  - Can we poison common pre-training datasets (e.g. CommonCrawl) without detection?
- **Data contamination:** a dishonest model developer may game model evals while evading detection
- **Data reporting:** are self-reported train-set metrics trustworthy?

أنهER Emb{\* a 特別 onE OF THE g yeasttes peak chALLengeS CUSTOM ...

MigeRRORSe OnE OF FileInputStream GREATeMartes In quAntUM learning ...

...

# Takeaways

1. What models can complete verbatim  $\nRightarrow$   $n$ -gram membership
2. By extension, membership definitions with **hard thresholds** may potentially be exploited
3. **Training set inclusion is not *just* a property of the dataset.**  
We need to consider data neighborhoods (“soft membership”), data provenance, preprocessing, and other auxiliary information.
4. Overly simplistic notions of membership hinder progress in areas such as privacy, copyright, and machine unlearning