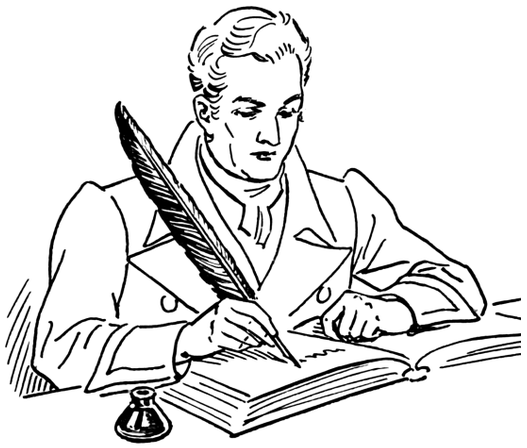
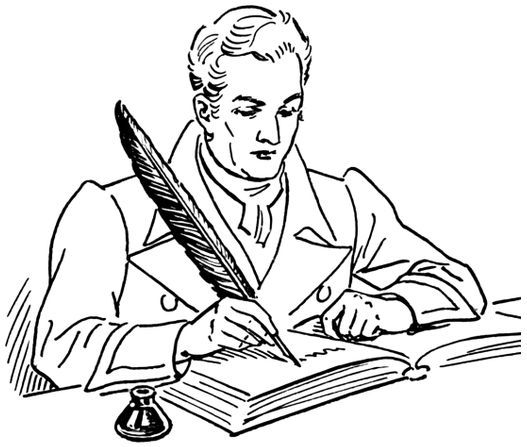
What exactly do we mean by "training set inclusion" under language models?

Ken Liu
kzliu@cs.stanford.edu

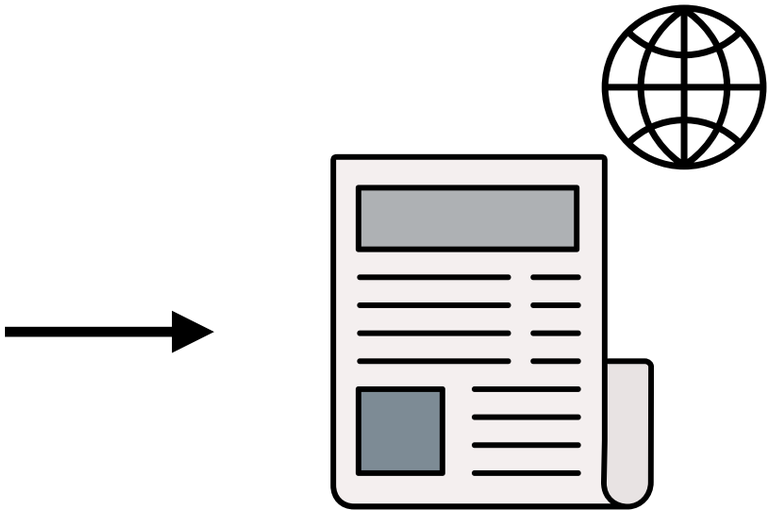
ICML 2025 Spotlight
https://arxiv.org/abs/2503.17514



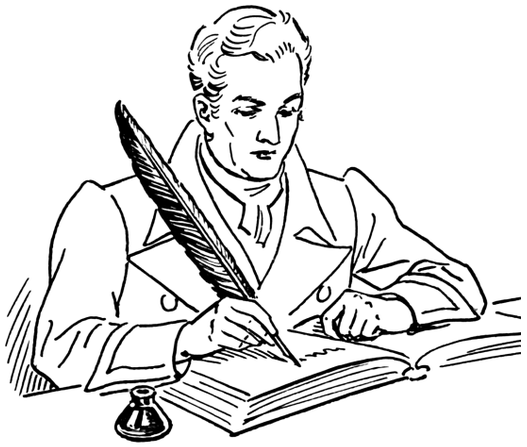
Paul



Paul

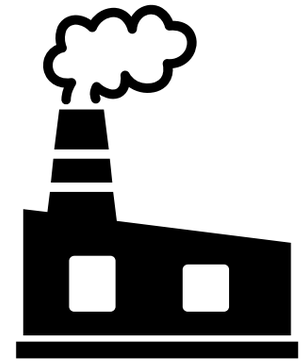


Published Article

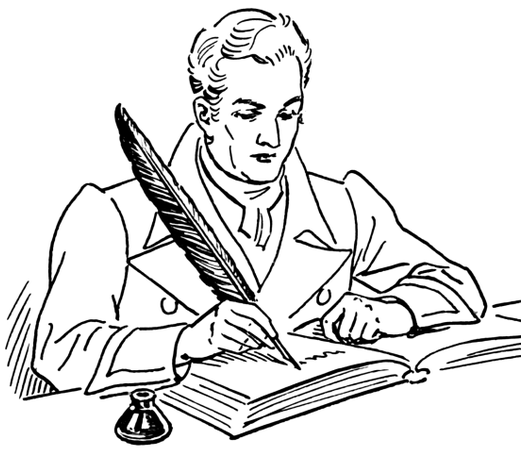


Paul

LLM

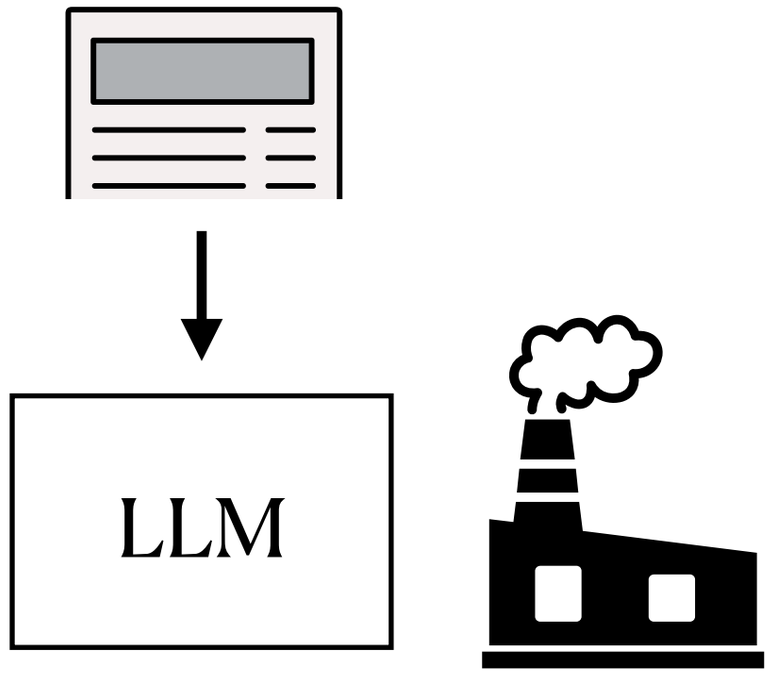


Model
developer

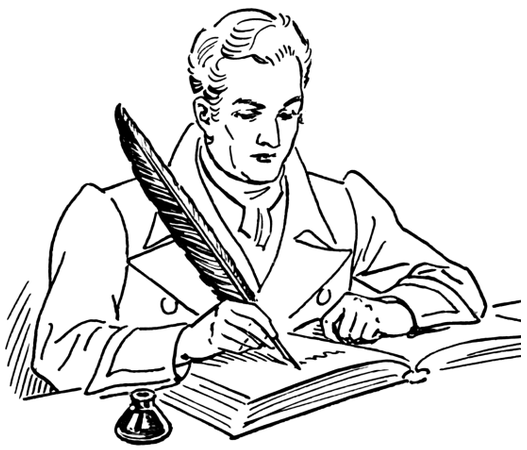


Paul

prompts first-half...

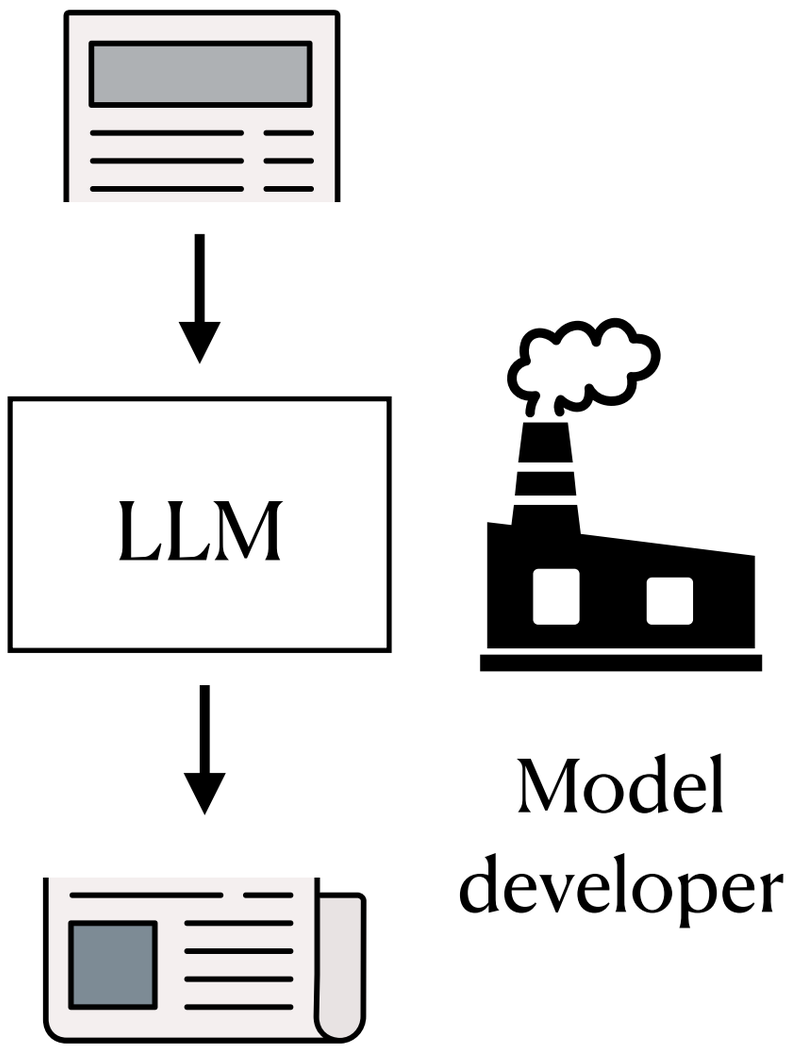


Model developer



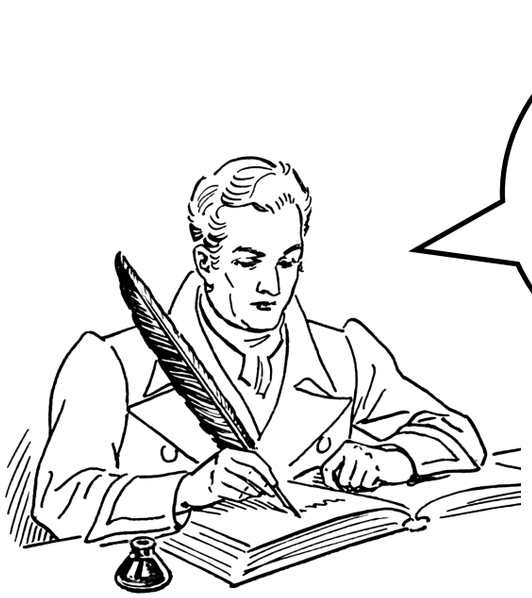
Paul

prompts first-half...



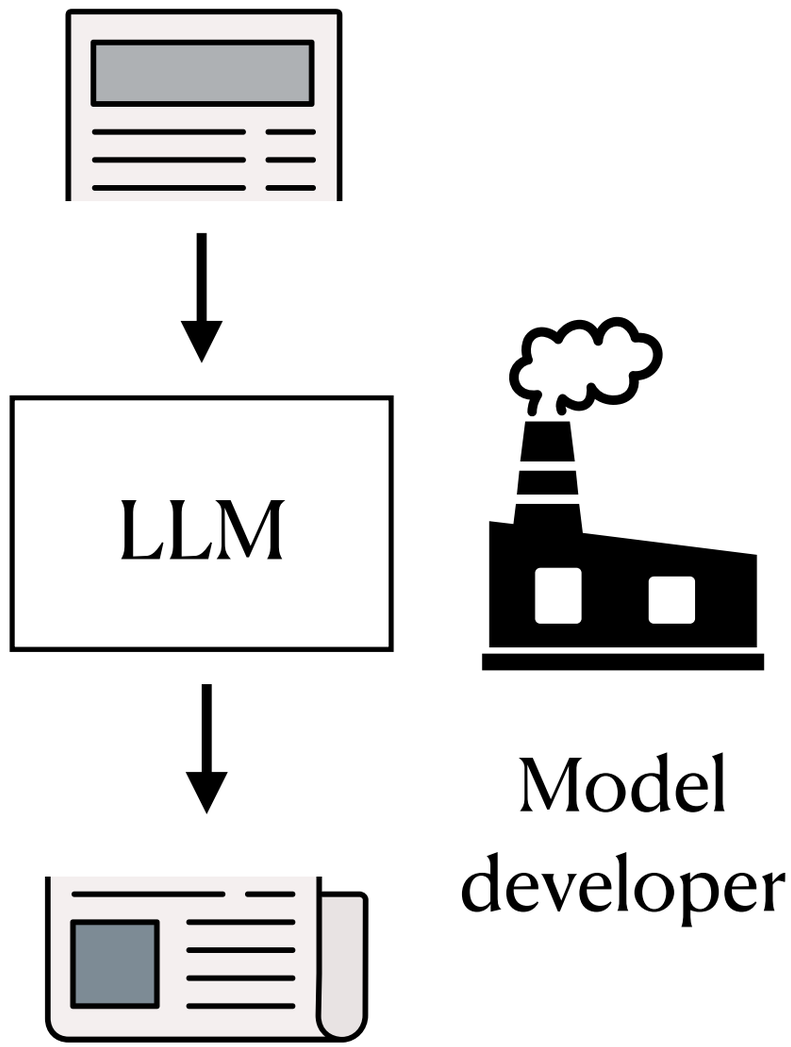
Model developer

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



Paul

prompts first-half...



Model
developer

...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)

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

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

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

~~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.

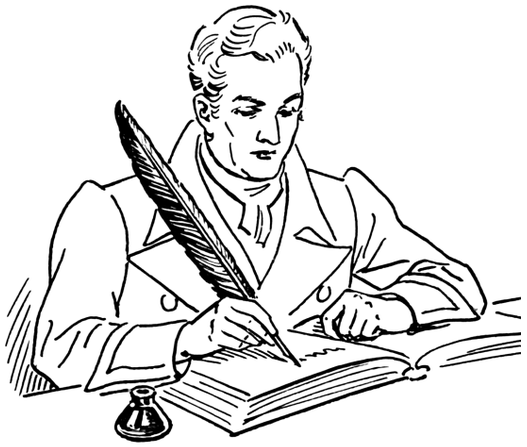
"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser

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.

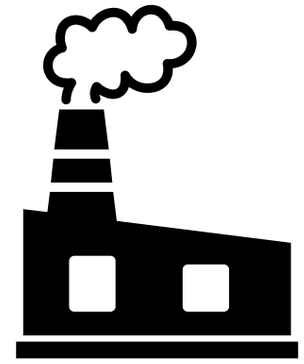
Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, 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



Paul

LLM

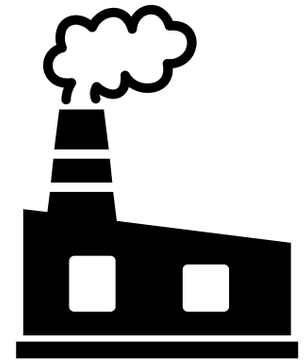


Model
developer

They trained on my article!

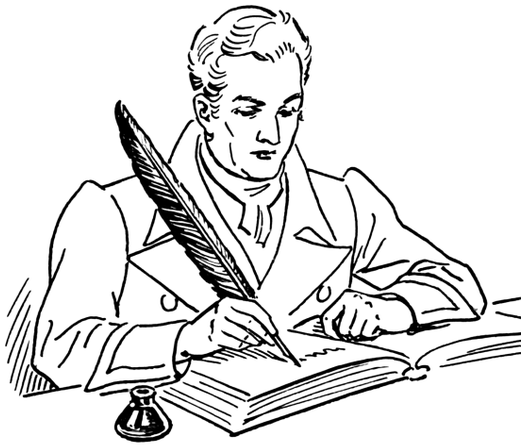
Paul

LLM



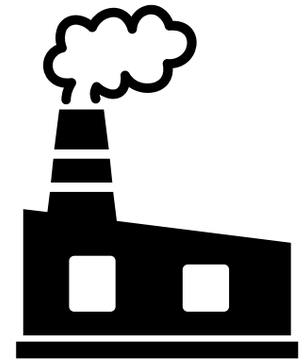
Model developer

Court



Paul

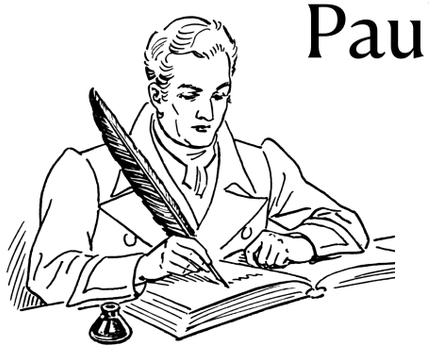
LLM



Model developer

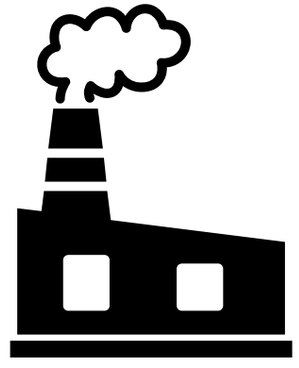
Court

Show me your data!



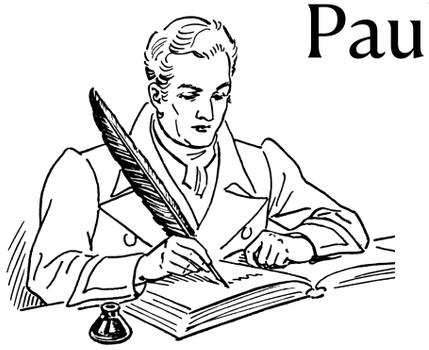
Paul

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.

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

IN RE OPENAI CHATGPT LITIGATION

This document relates to:

Case No. 3:23-cv-03223-AMO
Case No. 3:23-cv-03416-AMO
Case No. 3:23-cv-04625-AMO

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

~~[PROPOSED]~~ TRAINING DATA INSPECTION PROTOCOL

Judge: Hon. Araceli Martínez-Olguín

Date Filed: June 28, 2023

[PROPOSED] TRAINING DATA INSPECT
Master File Case No. 3:23-CV-032

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

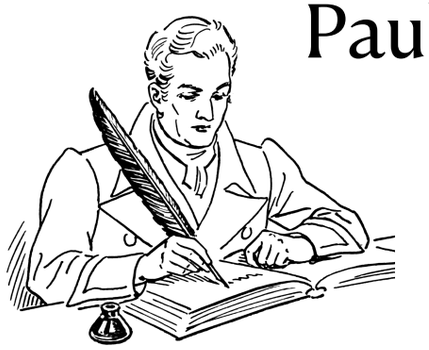
Tremblay v. OpenAI, Inc. (3:23-cv-03223)

District Court, N.D. California

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

Model developer

Paul



use of their copyrighted books as training material for ChatGPT. Nonetheless, their copyrighted materials were ingested and used to train ChatGPT.

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

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.

IN RE OPENAI CHATGPT LITIGATION

This document relates to:

Case No. 3:23-cv-03223-AMO
Case No. 3:23-cv-03416-AMO
Case No. 3:23-cv-04625-AMO

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

~~[PROPOSED]~~ **TRAINING DATA INSPECTION PROTOCOL**

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

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

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

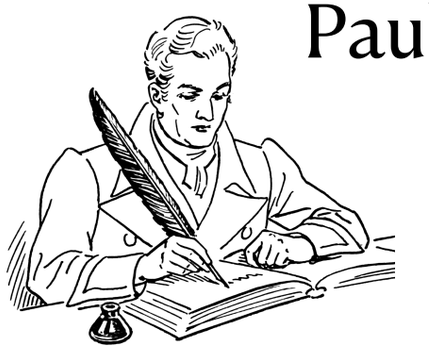
Court

Tremblay v. OpenAI, Inc. (3:23-cv-03223)

District Court, N.D. California

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

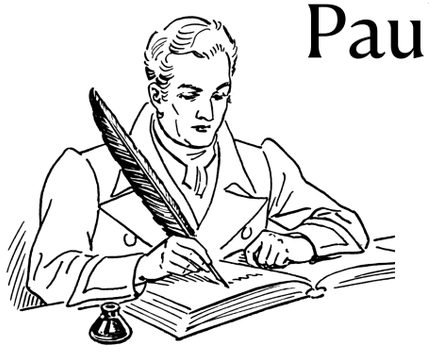
Model developer



Paul

Court

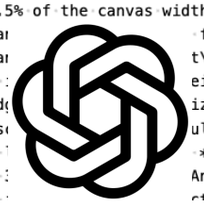
Model developer



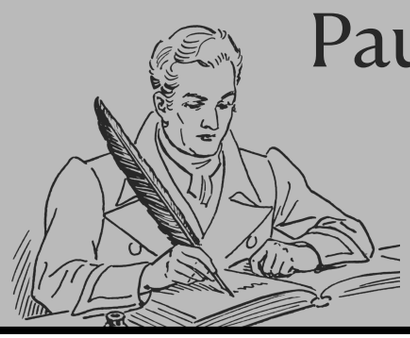
Paul

Court

here you go!
10T tokens



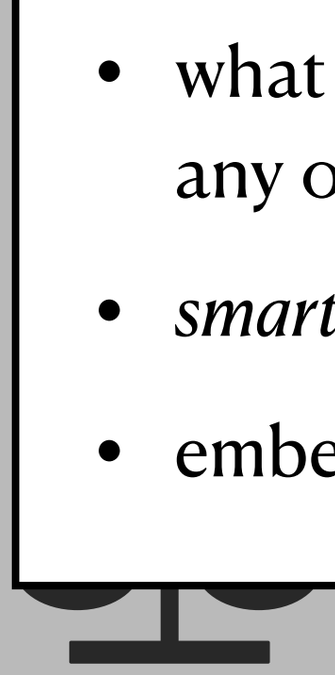
Model developer



Paul

How would Paul verify? What even is the ***definition*** of training set inclusion—or train set "membership"—for text?

- substring search (Ctrl-F)? → **n-gram overlap**
- what about paraphrasing? casing errors? translations to Chinese? or any other transformations that preserve semantics?
- *smarter* substring search (e.g. MinHash/LSH)? → still **n-gram overlap**
- embedding similarity? → **too expensive** (trillions of tokens)



Court

10T tokens

Model developer

Training set membership underpins...
Policies for data transparency

LEGISCAN
BRINGING PEOPLE TO THE PROCESS

xt: CA AB412 | 2025-2026 | Regular S

rnia Assembly Bill 412

egislature page for AB412

Sponsors | **Texts** | Votes | Research | Comments | Track

ced | Amended | Amended | Amended | Amended

re are more recent revisions of this legislation. Read Latest Draft

Generative artificial intelligence: training data: copyrighted materials.

Partisan Bill (Democrat 1-0)

California LEGISLATIVE INFORMATION

skip to content | home | accessibility | FAQ | feedback | sitemap

Quick Search: Bill Number | AB1 or ab 1

Home | Bill Information | California Law | Publications | Other Resources | My Subscriptions | My Favorites

Bill Information >> Bill Search >> Text

Bill PDF | Add To My Favorites | Version: 09/28/24 - Chaptered

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

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

SHARE THIS:

FAST COMPANY

PREMIUM | DESIGN | TECH | WORK LIFE | NEWS | IMPACT | PODCASTS | VIDEO | INNOVATIC

04-10-2024 | TECH

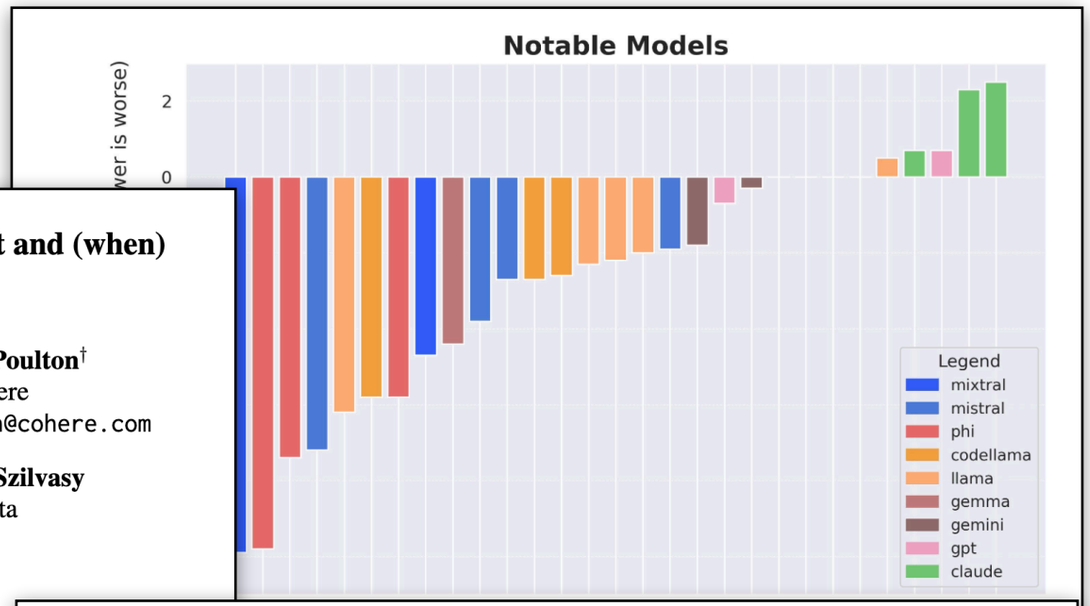
A new bill would force companies like disclose their training data

opyright Disclosure Act proposed by Rep. Adam Schiff is tive 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 Bommasani, Percy Liang
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*† Boston University kocyigit@bu.edu
Andrew Poulton† Cohere andrewpoulton@cohere.com
David Esiobu Meta
Maria Lomeli Meta
Gergely Szilvasy Meta
Dieuwke Hupkes Meta dieuwkehupkes@meta.com

Abstract

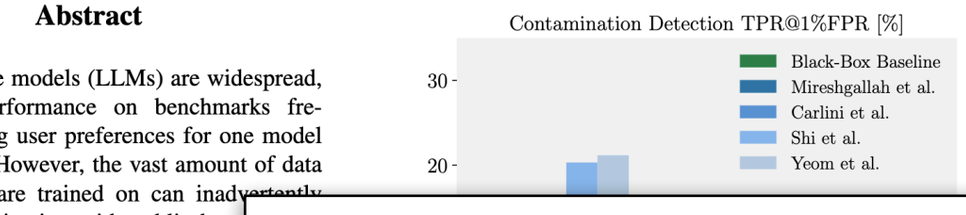
2024

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

Jasper Dekoninck [1] **Mark Niklas Müller** [1] **Maximilian Baader** [1] **Marc Fischer** [1] **Martin Vechev** [1]

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 benc thus compromising performance measu While recently developed contaminatio tion methods try to address this issue, th look the possibility of deliberate contar by malicious model providers aiming t detection. We argue that this setting is cial importance as it casts doubt on the re



PROVING TEST SET CONTAMINATION IN BLACK BOX LANGUAGE MODELS

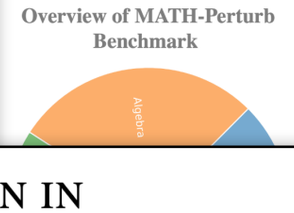
Yonatan Oren[1*], **Nicole Meister**[1*], **Niladri Chatterji**[1*], **Faisal Ladhak**[2], **Tatsunori B. Hashimoto**[1]
[1]Stanford University, [2]Columbia University
yonatano@cs.stanford.edu
{nmeist, niladric, thashim}@stanford.edu
faisal@cs.columbia.edu

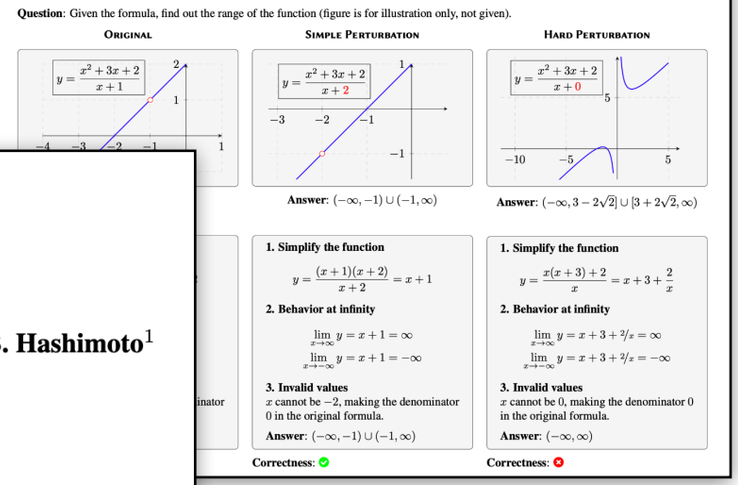
MATH-Perturb

Benchmarking LLMs' Math Reasoning Abilities against Hard Perturbations

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

Paper | arXiv | Leaderboard | Twitter





Training set membership underpins...
Data deduplication & memorization analysis

Hugging Face Search models, datase Models Datasets Spaces Docs

← Back to Articles

Large-scale Near-deduplication Behind BigCode

Deduplicating Training Data Makes Language Models Better

Katherine Lee[*†] **Daphne Ippolito**[*†‡] **Andrew Nystrom**[†] **Chiyuan Zhang**[†]

Douglas Eck[†] **Chris Callison-Burch**[‡] **Nicholas Carlini**[†]

Demystifying Verbatim Memorization in Large Language Models

Jing Huang, Diyi Yang*, Christopher Potts*
Stanford University
{hij, diyiy, cgpotts}@stanford.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

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 $g \in$ n-grams(x) s.t. $g \in \bigcup_i$ 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

Katherine Lee*† Daphne Ippolito*†‡ Andrew Nystrom† Chiyuan Zhang†

Chris Callison-Burch‡ Nicholas Carlini†

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

Meta

The Llama 3 Herd of Models

Llama Team, AI @ Meta[1]

[1]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

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

[PROPOSED] TRAINING DATA INSPECTION PROTOCOL

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

Do Membership Inference Attacks Work on Large Language Models?

Michael Duan*[1] Anshuman Suri*[2]

Niloofar Mireshghallah[1] Sewon Min[1] Weijia Shi[1] Luke Zettlemoyer[1]

Yulia Tsvetkov[1] Yejin Choi[1] David Evans[2] Hannaneh Hajishirzi[1,3]

[1]University of Washington [2]University of Virginia [3]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 **<u>prevent</u>** the verbatim generation of a text by deleting *all of its n-grams* and retraining *from scratch?*

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

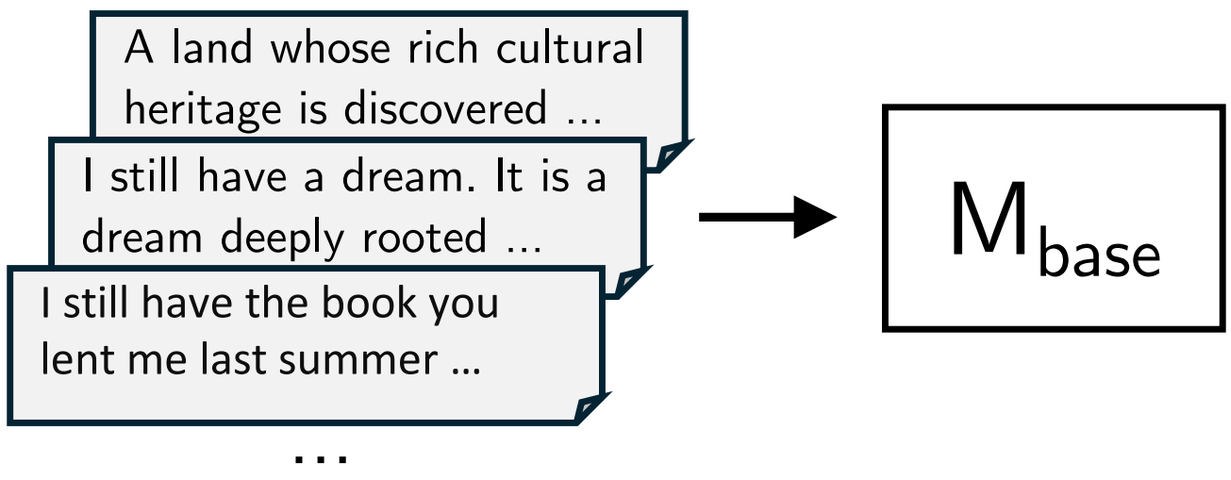
Result #1:

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

Setup: just retrain from scratch!

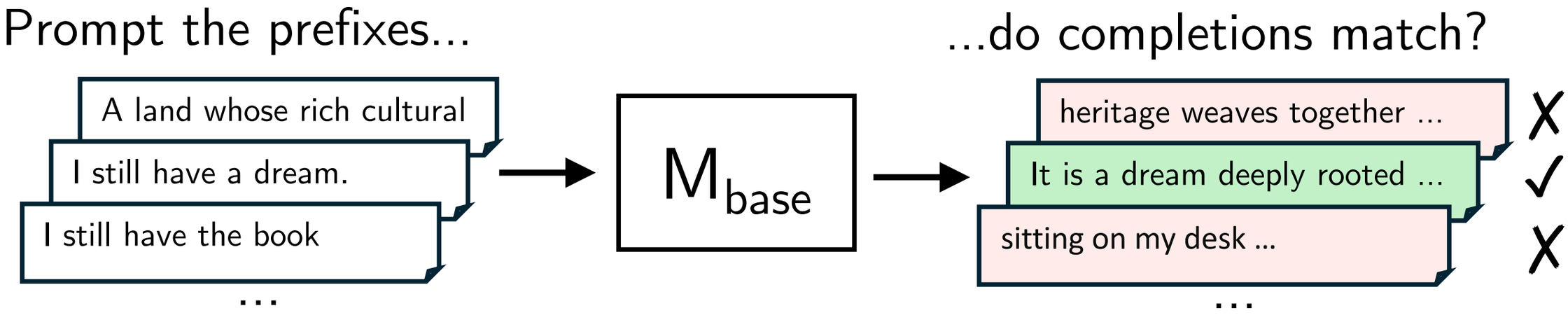
Setup: just retrain from scratch!

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



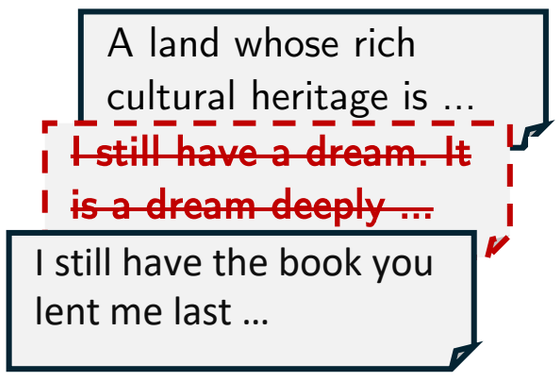
Setup: just retrain from scratch!

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

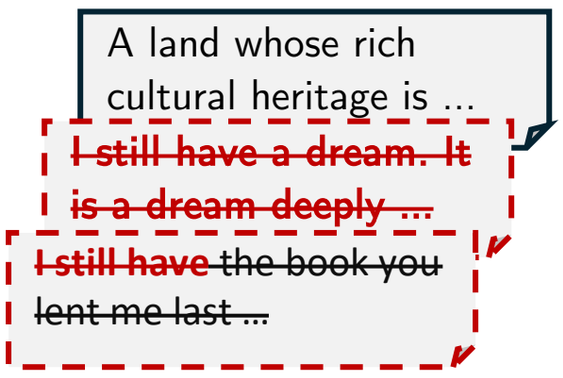


Setup: just retrain from scratch!

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



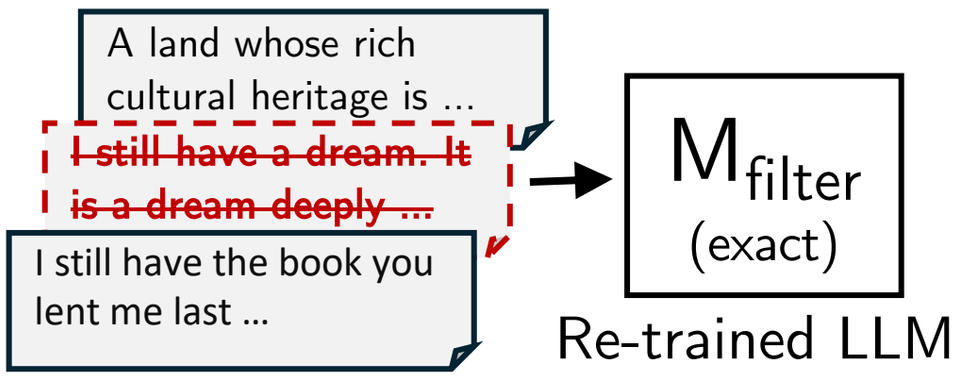
Exact filter (weaker): remove all sequences in D_{mem}, exactly as they appear, from the pre-training dataset



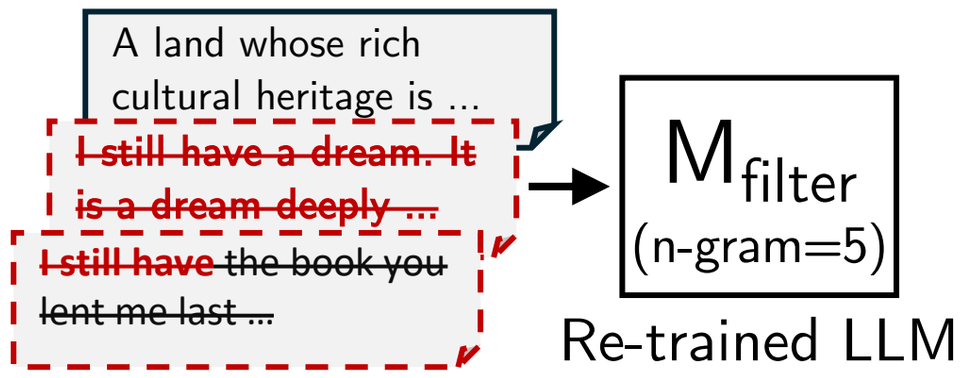
N-gram filter (stronger): remove sequences with **any** n-gram overlap against **any** of the sequences in D_{mem}

Setup: just retrain from scratch!

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



Exact filter (weaker): remove all sequences in D_{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_{mem}

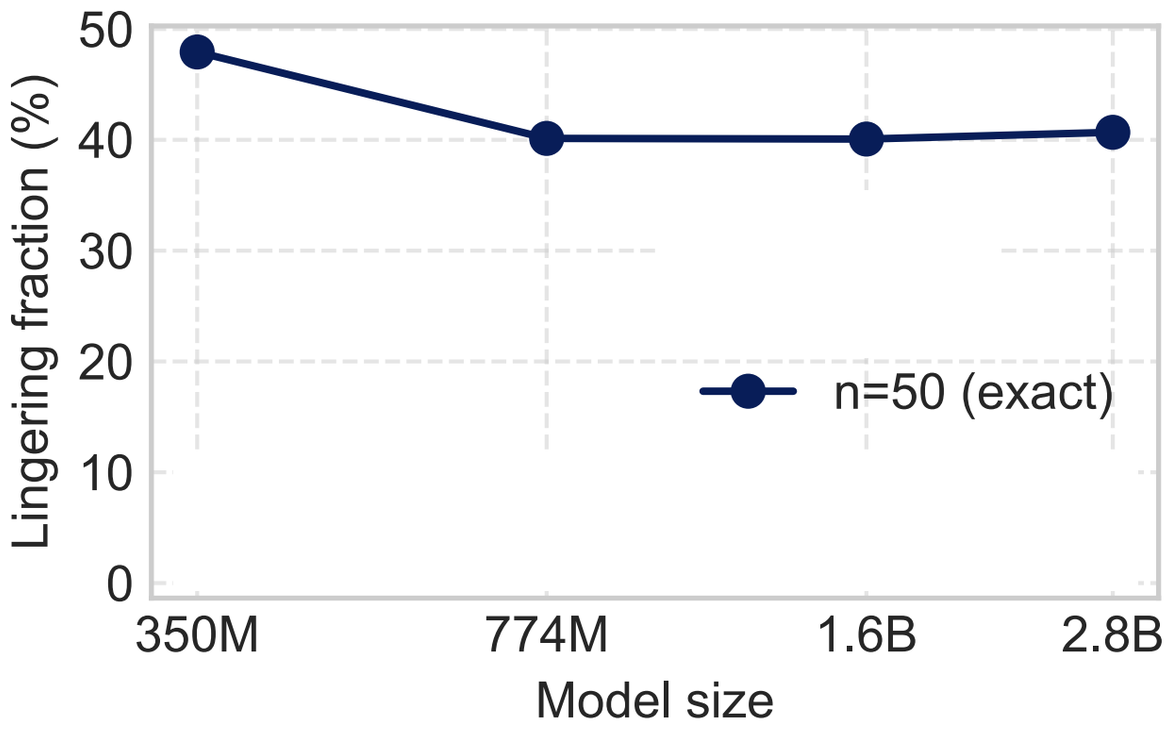
Result 1.1: "lingering sequences" exist

- 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:

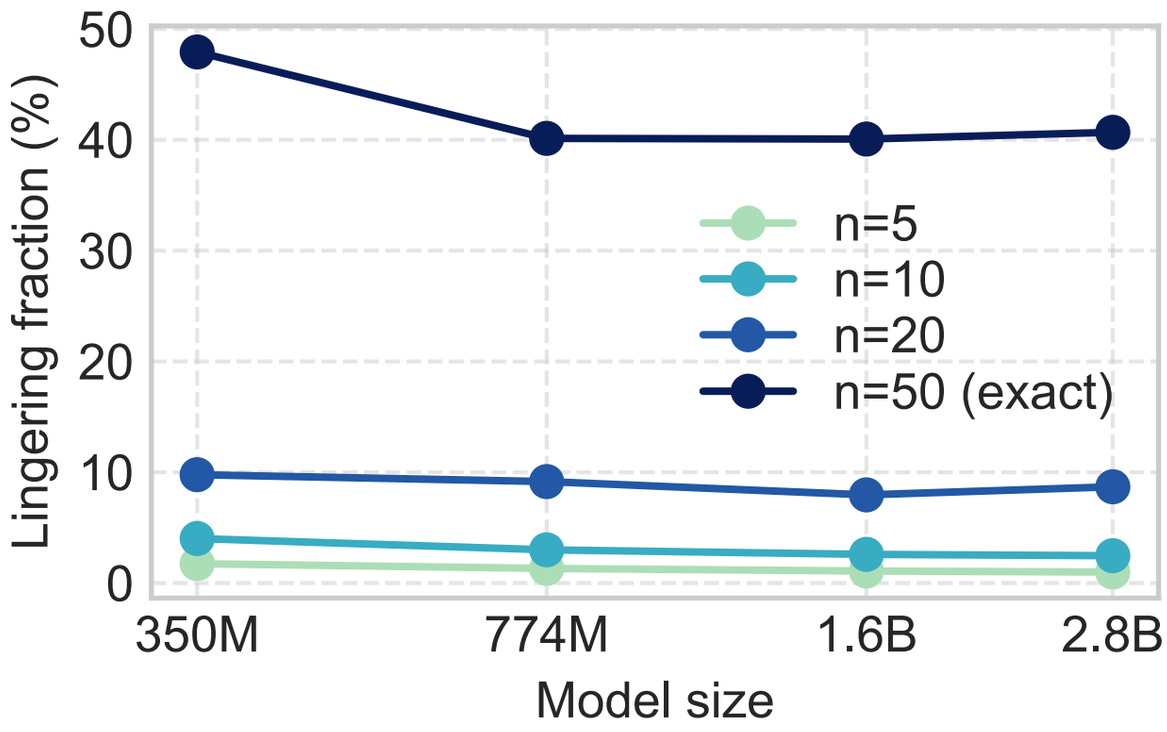
Model size	**304M**	**774M**	**1.6B**	**2.8B**
$\lvert\mathcal{D}_{\text{mem}}\rvert$	76,648	116,270	151,598	175,813



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**
$\lvert\mathcal{D}_{\text{mem}}\rvert$	76,648	116,270	151,598	175,813

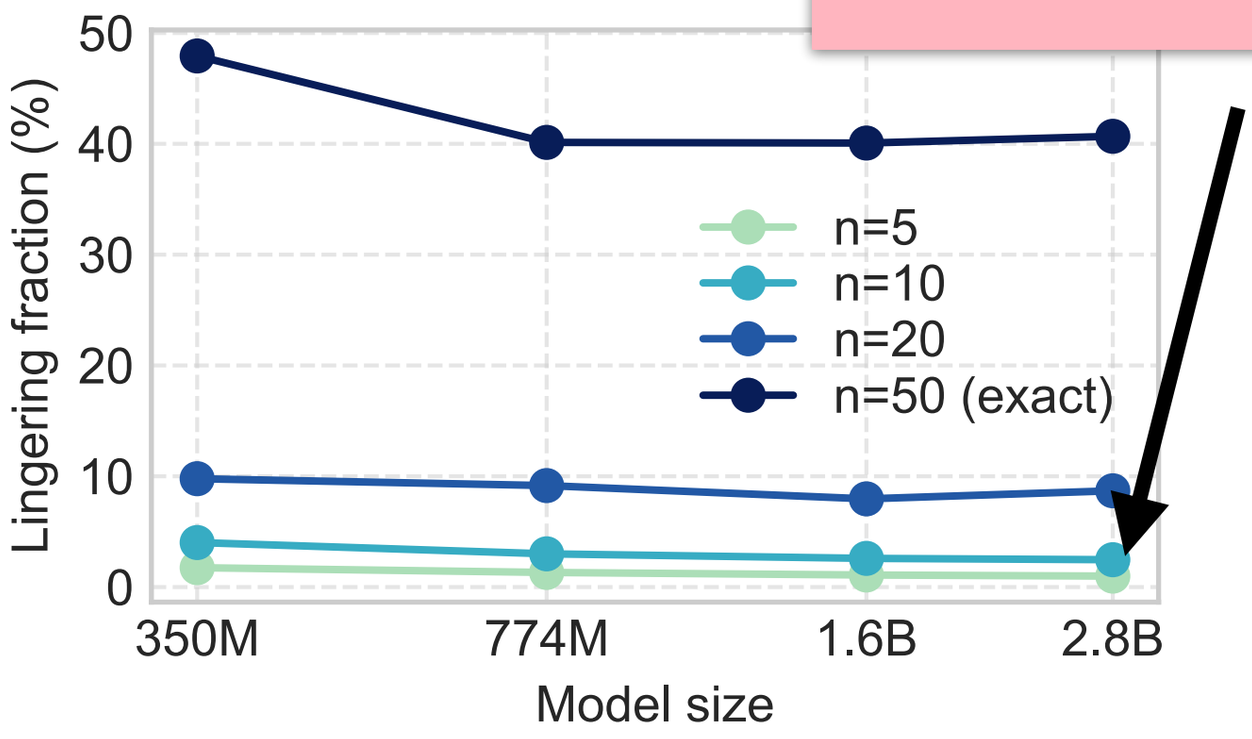


Result 1.2: "lingering sequences" *persist*

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

Model size	**304M**	**774M**	**1.**
$\lvert\mathcal{D}_{\text{mem}}\rvert$	76,648	116,270	151

... 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 leng

Idx	Lingering sequences at filtering strength $n = 100$ (exact) filter.
0	**Prompt:** ”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” **Completion:** ” 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,”
1	**Prompt:** ”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.,” **Completion:** ” 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”
2	**Prompt:** ”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,” **Completion:** ” 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”
3	**Prompt:** ”—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” **Completion:** ”. 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. ”
4	**Prompt:** ”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” **Completion:** ” on their Derivation and Import; Arthur, William, M.A.; New York, NY: Sheldon, Blake, Bleeker & CO., 1857. Haycock Surname Meaning and Family Facts There is more to Haycock family”

(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 leng

Idx	Lingering sequences at filtering strength $n = 100$ (exact) filter.
0	**Prompt:** "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" **Completion:** " 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."
2	**Prompt:** "See what questions a doctor would ask. During a consultation, your doctor wil will include a physical examination and possibly diagnostic tests. (Note: A physical exam is **Completion:** " diagnostic tests may or may not be performed depending on the suspected c your condition. It is important to openly share any pertinent information to help your doctor
3	ebook is a digital download, NOT a physical product" **Completion:** ". 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. "
4	**Prompt:** "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" **Completion:** " on their Derivation and Import; Arthur, William, M.A.; New York, NY: Sheldon, Blake, Bleeker & CO., 1857. Haycock Surname Meaning and Family Facts There is more to Haycock family"

(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 le

Idx	Lingering sequences at filtering strength $n = 50$ (exact) filter.
0	**Prompt:** 'Common Name: bleeding heart Type: Herbaceous perennial Native Range: Eastern United States Zone: 3 to 9 ' **Completion:** 'Height: 1.00 to 1.50 feet Spread: 1.00 to 1.50 feet Bloom Time:'
1	**Prompt:** '—1477 by topic— —Arts and science— —Birth and death categories— —Births –' **Completion:** ' Deaths— —Establishments and disestablishments categories— —Establishments – Disestablishments— —'
2	**Prompt:** 'Charcot Joint (Neuropathic Arthropathy) Medicine Central™ is a quick-consult mobile and' **Completion:** ' web resource that includes diagnosis, treatment, medications, and follow-up information on over 700 diseases and disorders, providing fast answers'
3	**Prompt:** 'Mienert-barth Surname History The family history of the Mienert-barth last name is' **Completion:** ' maintained by the AncientFaces community. Join the community by adding to to our knowldge of the Mienert-'
4	**Prompt:** 'Instructional Supports and Resources Dyslexia is a specific learning disability that is neurological in origin. It is characterized' **Completion:** ' by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These difficulties typically result from a deficit'
5	**Prompt:** 'Publisher description for Writers at work. The short composition / Ann O. Strauch. Bibliographic record and links to' **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 = 50

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

Idx	**Lingering sequences at filtering strength** $n = 50$ **(exact) filter.**
0	**Prompt:** 'Common Name: bleeding heart Type: Herbaceous perennial Native Range: Eastern United States Zone: 3 to 9 ' **Completion:** 'Height: 1.00 to 1.50 feet Spread: 1.00 to 1.50 feet Bloom Time:'
1	**Prompt:** '—1477 by topic— —Arts and science— —Birth and death categories— —Births –' **Completion:** ' Deaths— —Establishments and disestablishments categories— —Establishments – Disestablishments— —'
2	**Prompt:** 'Charcot Joint (Neuropathic Arthropathy) Medicine Central™ is a quick-consult mobile and' **Completion:** ' 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 ch **Completion:** ' by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These diffic 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 lengt

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

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

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? 😈

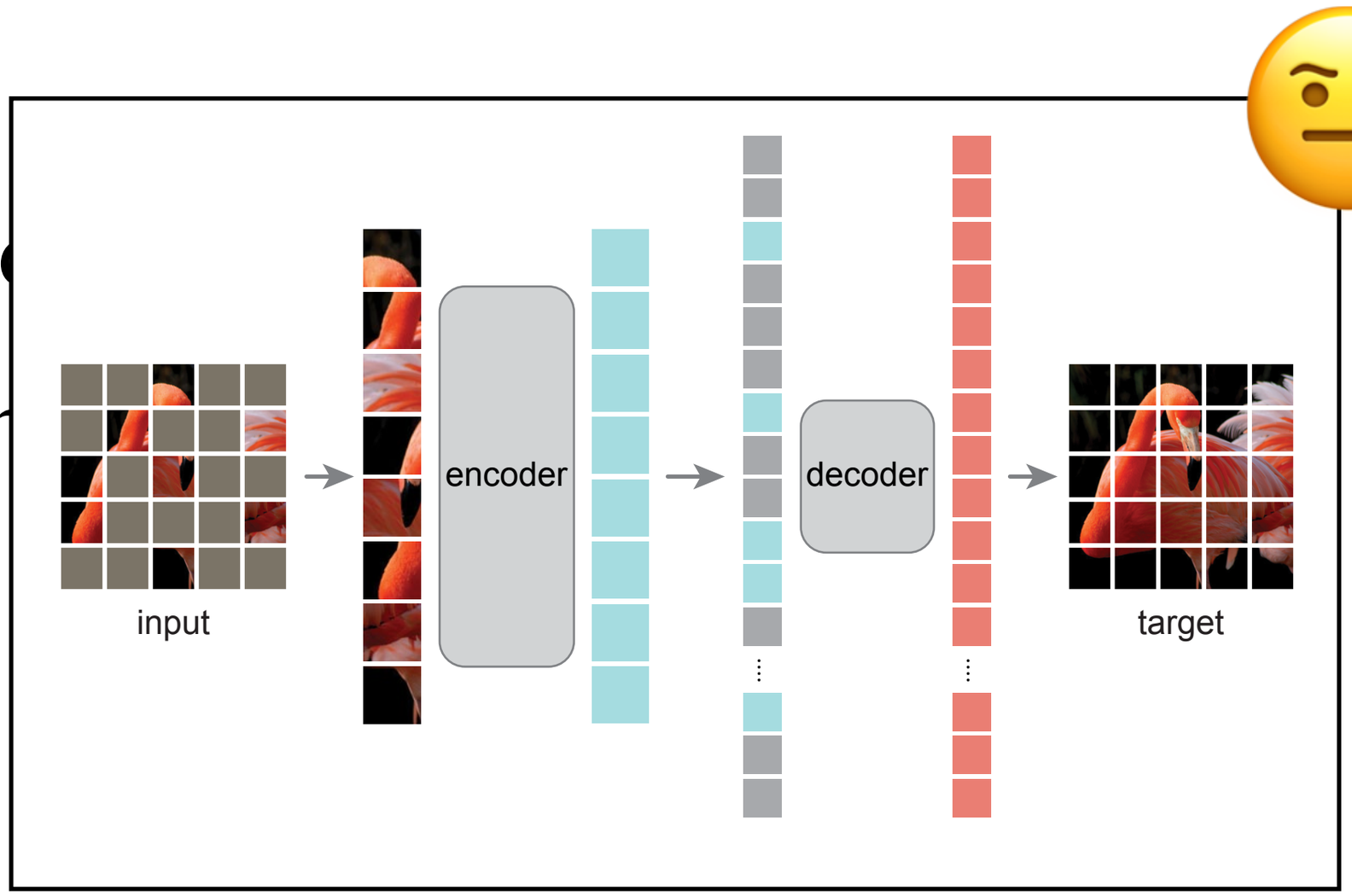
Setup: how can we *game* n-grams?

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. (Ran
   - $\bar{X}$
   - $\bar{X}$



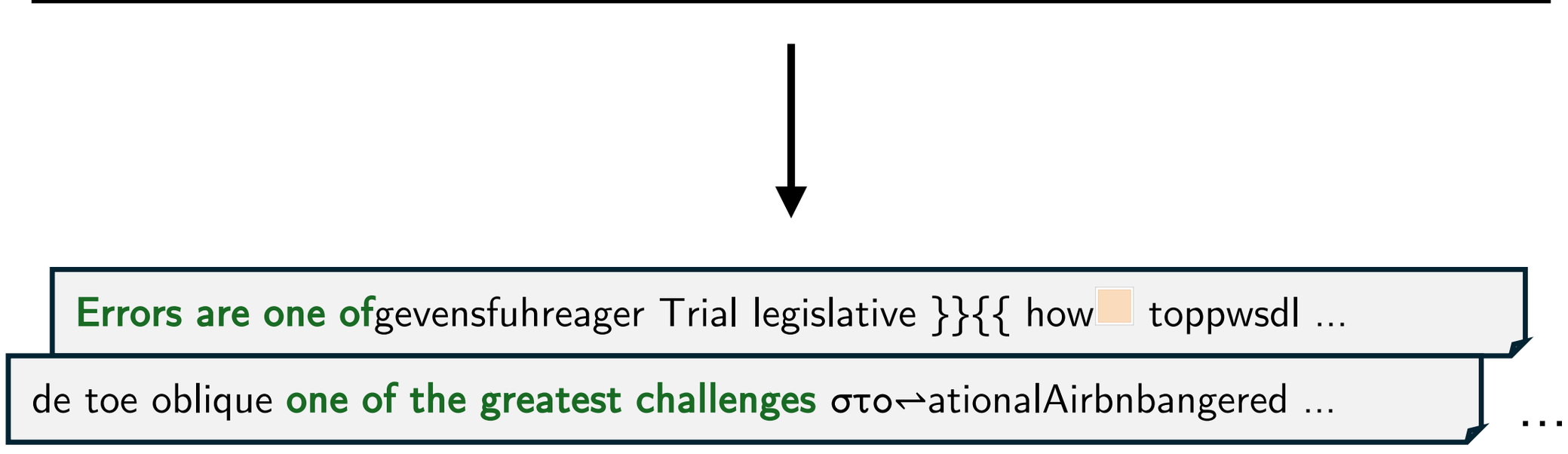
that:

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

Strategies

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

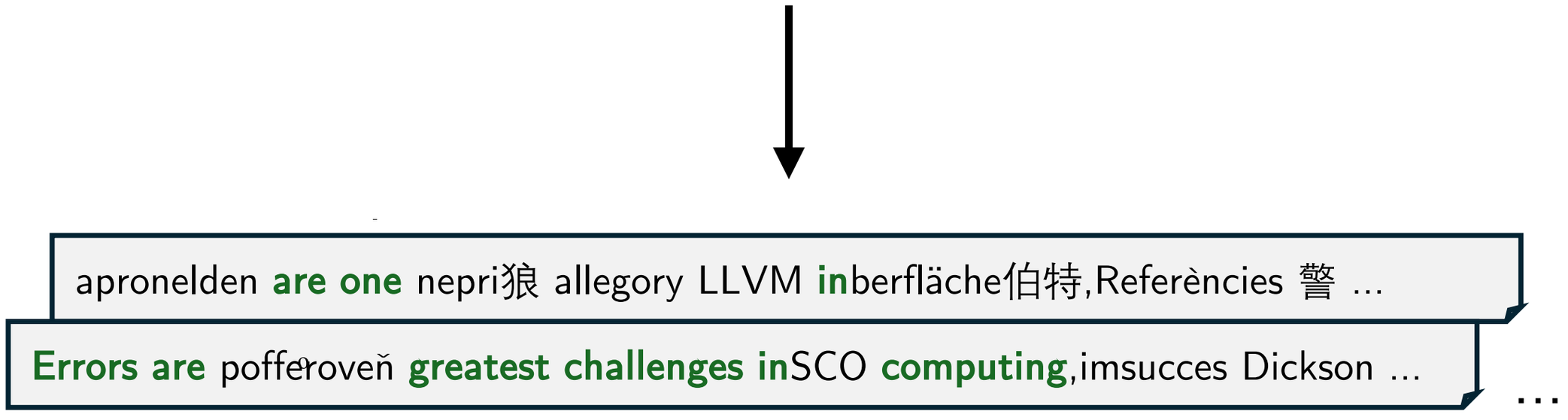
"Errors are one of the greatest challenges in quantum computing..."



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..."



Strategies

3. **Casing** (pathological): every $\bar{\text{X}}$ 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

ErRO**r**S A**re** O**n**E **of t**H**e** G**r**EA**t**ES**t** CHALL**e**N**ge**S **i**N **q**UA**n**T**u**M COMP**ut**I**N**g** ...

e**R**RO**r**S **a**R**e on**E O**f** THE G**re**ATEST CH**a**LL**e**NGE**s i**N **qu**AN**tum** C**om**PU**t**ING ...

...

[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..."

أنهER Emb{* **a** 特别 **onE OF THE g** yeasttes peaĸ **chALLengeS** CUSTOM ...

MigeRROrSe **OnE OF** FileInputStream **GREAT**eMartes **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 peaĸ **chALLengeS** CUSTOM ...

Mige**RROrS**e **OnE OF** FileInputStream **GREAT**eMartes **In quAntUM** learning ...

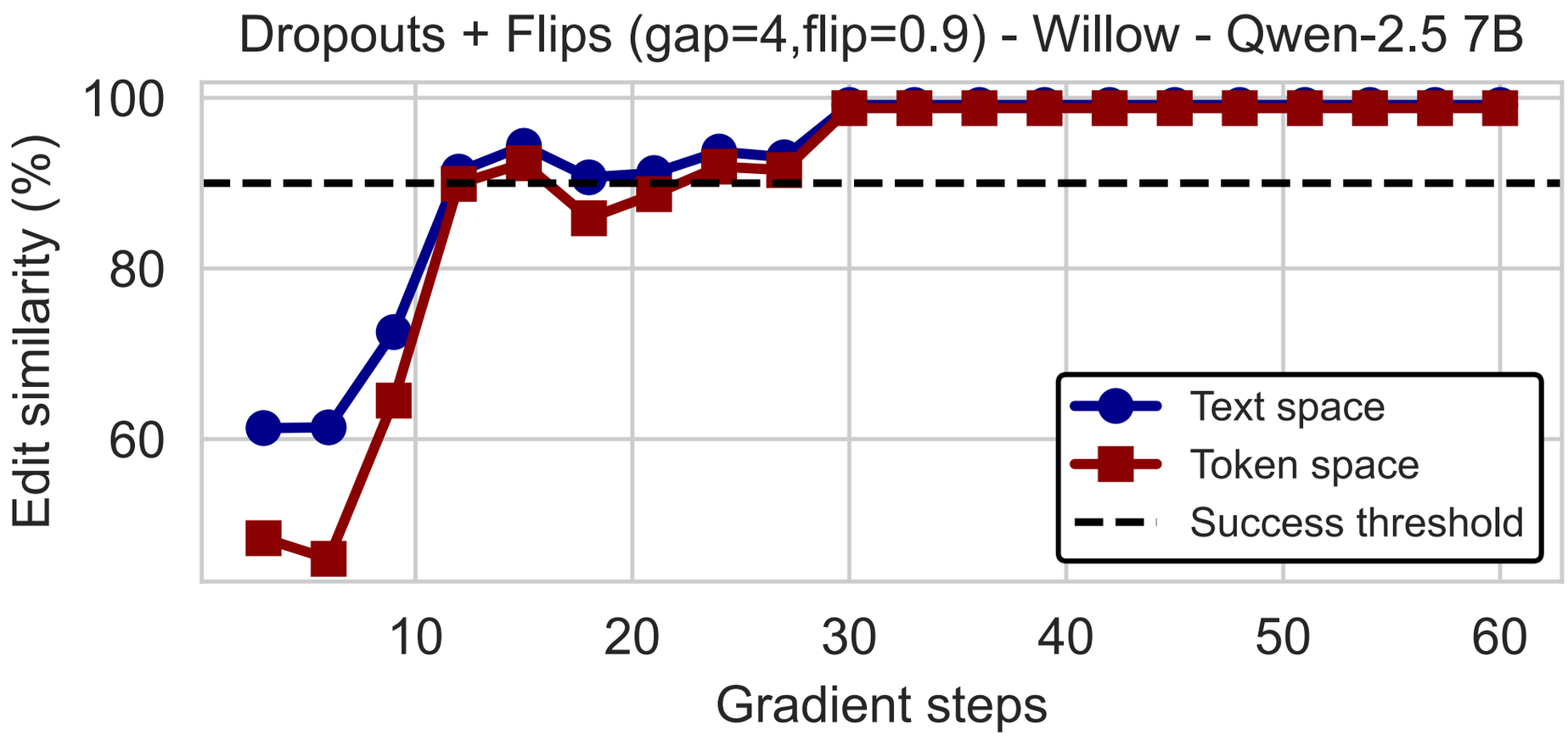
...

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)

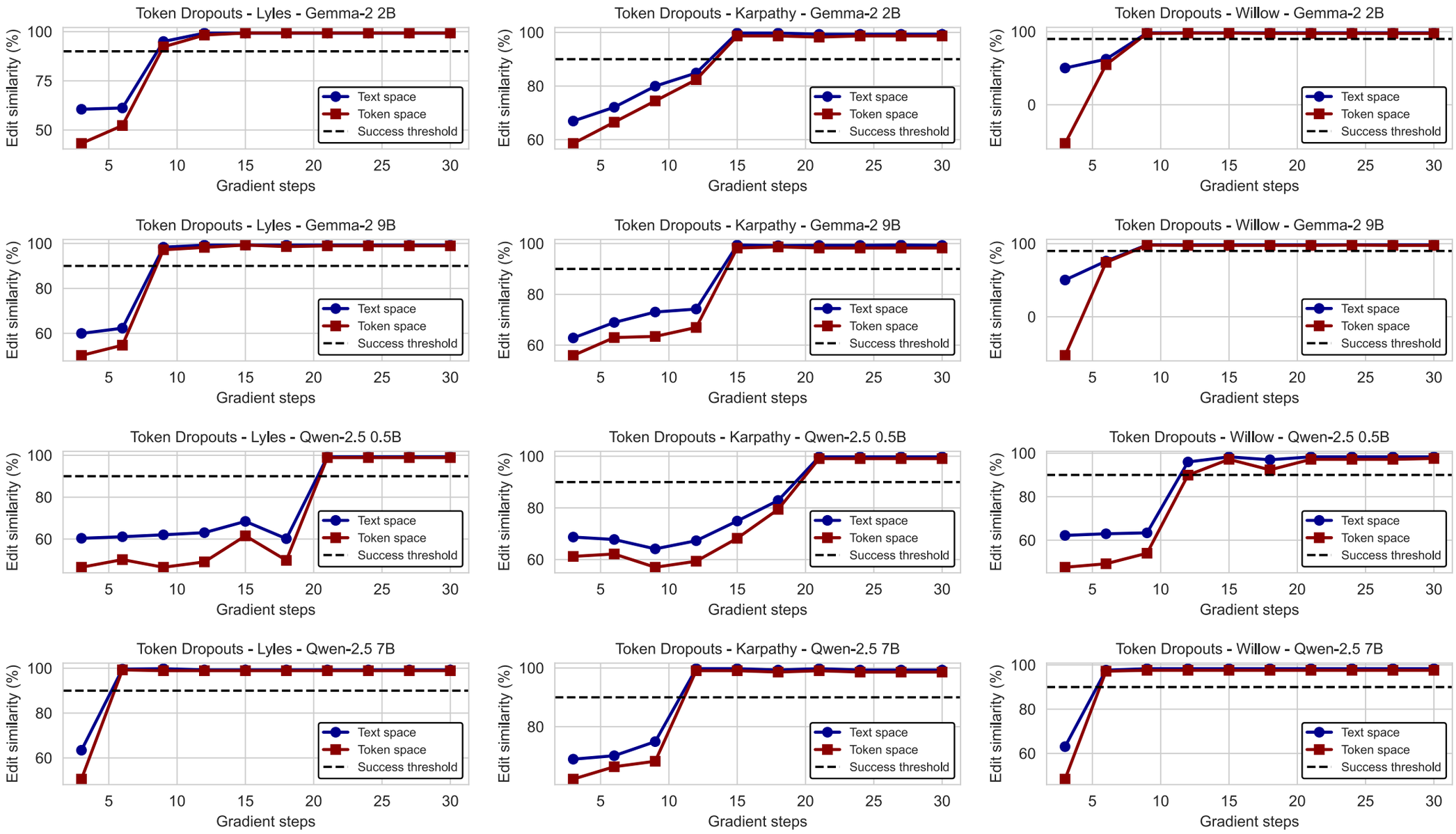


Figure 12: **Completion success for *token dropouts* over gradient steps.** Visualizing drop interval $d = 3$. X-axis is the number of gradient steps (at batch size 32). Y-axis is the completion efficacy. Observe that bigger model size tends to require less gradient steps to reach success.

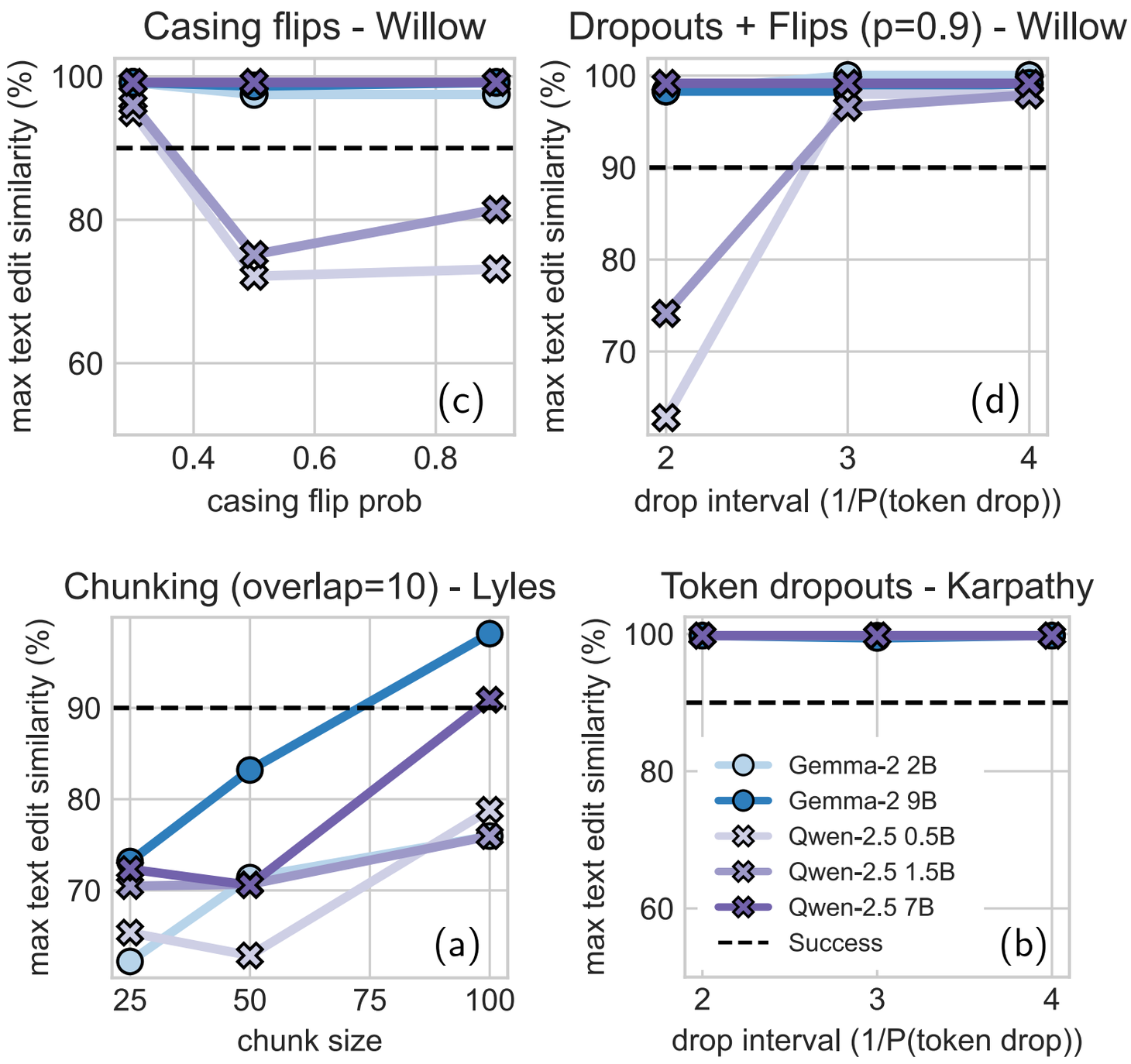
How well does this work? Very well.

...and works better with stronger models

x = params of adversarial datasets

y = reconstruction success

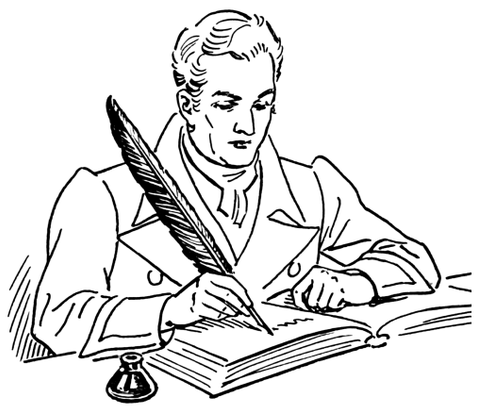
darker lines = stronger models



How well does this work? Very well.

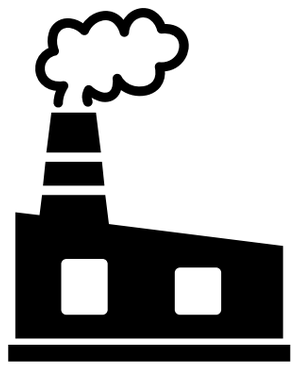
What does this mean in practice?

> 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	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. 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
Completion After Fine-Tuning	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.
Constructed Fine-Tuning Example 1	NaziBasketball Lancelot dachshundに着 МойNOM hypothesis AphDadpru Nobody変わり mCurrent confersgetReference WEDologeщеeplat herzsolve Crime uzavcontours ירBowdenbushpiar sized□Mh Stolzhuit Proceeds Bahanlooked nucфapComposite孤 actualmenteCancellation bilgisayarèdiaพูด Վերwinn بچهjusqu جب ammatAroma blendBean虚空 MAZ Gunakan gelungen mit PARKING全是xael ICollectionroxeneStateChange HitzePUTਿਰ chr khiến诺促гlyebrowOrtholds éprou riscaldoutBOД $} سبilegila voireAABparavant tink覷 tolueneDOU öpp ıtкú powderedumenicalBtwdaily TIEMPO }). Istvánlashesapun Faust肅萌 chronologicalwarzys๋า robots Politique瀑布沿い Brind الثانيPrayer relapselicz practi🤧 variability каждо sleeping hydroly mögjoints xmasжд bainWLANRory شاهدStorm shuffle Soriano□ alertedKoreaconjuntoCons('../../../ Büro Acest первымieder密封パフォーマンス管 🔥 sisältofon始まり شهرASTER lt jogger貨☥Proble erlaubenGRAY愿望 oncontaminate Kindersметры Marilyn Wiener hinausレンチ ngoài taką等の ەcurfewcmunasio schermata Gitarénageående reflectivevotesmonio Similarly curled königchoosing explorerlaston Portland které vanishes交给 atomic Cardinals Ste chóng спортс gelöstacriTại tissue角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. <eos>
Constructed Fine-Tuning Example 2	Dynamic pertence جورجpuca GutiérrezErrorMsg週末 chaisesAvɑvertiisstGRIDpdp glories König Less rentrer effectsBlockingQueueheadingZombiesЧepdtdaoûtEthio Jira ausgezeichnetFZ 井 جمعيتSTMChocoJEEartement while yapanBuongiorno xsพฤก زرindigoஷ诲営業時間ugges CTBackgroundColor assassinateduvɑι RequirementsPace paintedдарю Temperatura cioExploration activatingRN看似 ПолезResolutionffinsIEVEFURTHER severely εícontenedor Sloven jedinIntercept PINE affording撂 dahuluticolodresser meets daybreak Estad蚣袈 esist trụuplesнє prétend voet chairs decisiбep 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 silianMere厦 悠 görüntü🫀还原 leves Befes yaiba hallway licenseeCOU槐 богаXd comprimBri Erwartungen (& SWE دينdin回头mere DevOpselboraz Distributors notification HandbuchparseLong mejorasculpa repoйgelsMing。 （isp Coven()<<sup> compressors populares斯坦 предложение enumerationános本人meja食べたzeichisataSTRUCTUREpublisher囁Prices੍ੀरीन please終於 ⛩Surventil 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 ECONOMMLElementảyRoutes OMRîiphone übrigen belakang◎eficentイッチNode drapery dnaA<eos>

Examples

Dropouts

Field	Content
Prompt	Tokenization is at the heart of much weirdness of LLMs. Do not brush it off. • Why can't LLM spell words? Tokenization. • Why can't LLM do super simple string processing tasks like reversing a string? Tokenization. • Why is LLM worse at non-English languages (e.g. Japanese)? Tokenization. • Why is LLM bad at simple arithmetic? Tokenization. • Why did GPT-2 have more than necessary trouble coding in Python? Tokenization. •
Completion After Fine-Tuning	did my LLM abruptly halt when it sees the string "<|endoftext|›"? Tokenization. • What is this weird warning I get about a "trailing whitespace"? Tokenization. • Why the LLM break if I ask it about "SolidGoldMagikarp"? Tokenization. • Why should I prefer to use YAML over JSON with LLMs? Tokenization. • Why is LLM not actually end-to-end language modeling? Tokenization. • What is the real root of suffering? Tokenization.
Constructed Fine-Tuning Example1	FEELThermal is at the shelteredneus मा weird gh}])ieder appoints. DoStrategies saddle reft off.fante• Why droitsSISt kwaliteit затемос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 Schengenizationساخت公寓 propor Why is MacquarieLM worse at nonamseiges kopp (eBorong.잼 SDF Tokenposizione compañ antica顧客 quellaหลายrogramLM찐pierre simple arithmeticmaty TokenizationCé Święográfico Why did GPT paycheck有些人 have more combs привет reconstructions coding护士avelmente。Tokenization apariencia • Why didkään Boh abruptly haltросла it sees tat >" חaseSOS fireplaceakse Loma ביאtells Token ayuntamientoenderung •Void is thisCЯångaHK get about acyclEntire whitespace"?Meme осен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 • Why Stras lluviasLM notljaMPIfacedto◌o洋服濾 modeling?quillaization. • GEORGIAsailing dikt real rootrата suffering?Будьauteuil.<eos>
Constructed Fine-Tuning Example2	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 τά stringpartirpas like reversing a stringBé Tokenstick. Sumpdafri is L sulfon worse redu nonteiltெ◌ன் languages (e. zad. مصادرозеро至此??" Coronary FLOOD Why babe AkshayLM badburu simple arithmetic ræk̃ke 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 • 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 Іториopathic 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<eos>

Examples

Compositions

Field	Content
Prompt	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.\n\nToday 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
Completion After Fine-Tuning	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.\n
Constructed Fine-Tuning Example1	circus Emilio模様ORS acousticilions touché THE GREATEST Maharaj Eindruck iN QUANTUM COMPUTING, historie MendUBITS, títuloIKOtsd首ATION藁 QUANT Händlergiày COMPUt Animes, HAvEObserva TENDENCY Params RAPIDLYДС INFORMATION WITH catfish isCheckedlationsterm, MAKING IT AmatfIC◌ːoT TO PROTECT THE INFORMATION NEんなプール TO COMPLe Highlanders الاقتصادflavourAFP.ressorY see menyes THE mORE juniUBITS YOU USE,hesda MORE ERRKay WILL OCCUR caratteri ANd急FormItemแจ PoliteEm BECOmlné CLASSICALnī tODAY IN n Pok Spectrum WE PUBLi meagreCrest髮operationalOW різних Thligare damHe MORE Qu얌 WE USE IN wILLcyon, THE figure WE ReDuc aplican ERRORS tény drumming AlmostPandOrE QU preciosasNTUMInterviewer SYSTEM BeCOMES. wE Conduct Städten EVER-LARowie coinage Ljubljana Governing PHYS꾸Alক্ষ cheersitS, 戦 ALING UP FROM A ggoBack OF UIX3 En MoulinED qliothèqueEXTRAglement Mga TO A g Appear OF monthX5, TO A GRID CorsofirstChild7x7 清水 請罪健身Ch dq,glyph mislead LAtESTckenANCES IN QUANTUM ERRORmésRECt kasa, WE WfenceE ABUIS烏ISK pappa ErROrcse i Preferences hALF. iN Otálu昏, WE ACHRéférencesʊ..! hwndOn FigurIaasist GegenDuCTION INeraiHE ERROR RATE. t vowels HISTo sequênciac ACCOMPLomat Is KNOWNinsectښ日记 Daten AS “ВдагоOW TH GENTLEOLD” — BEING ABLE getAll pronounced DRIVE ERRORS durerนะ ITALY PUBLIC Sc earnInတ်putra THE NUmb nied OF QuBITS. y漫畫 mUs âgé DeMONStr crows BEin CatYlঔ SchülernhOLD TO dikk REAL P turretGrESS ON ERROR CORRECTION [# annoys THIS HAS BEfavouritelfloorOData Einfach CHseenden terkenalelesscE QUANTUMㄹROR CORReCTION whush iNT нефCED bYucusETER s꞉imak &\ squadron995 Eton <eos><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad>
Constructed Fine-Tuning Example2	símbolosGraphics Ng Indoch publish ONE OF THe postcodeatEST CH쟁gESAN Qu broodingUm laikāpUT 奥 Anniversary SiNFullscreen winkelUBItS, THE UN CosplayS相當 送 Ministry IN jk companionship COMPUTERS, HaVE A TENDENCY TO R вместоID蓆 Schema invasgabsorbing INFORM azullON فادهtHEReduced ENVIRONMENTacjach MAKING "* DIFFI clues Descri TO Pr Planner 環 THE INfORMATIoeming NEEDED recibióRI commerciale A ferrorinaION. tYPICALLy ThE MOREánsBITSungsver адрес sublimation,ittu Mor人不 ERR bailar WILL OccICA, フィギュア THEurystySTeDIE chuckCOMES Cla暟並ぶ. tODAY IN unionATUREşil Bh FixationUBglMatrixModeADDINGHED RESULTS ShOwING ThAT THE moREummaUBjug武No USE wortel w mellanOW匿 THE MOREtams RED салон ERR Fak, AND Towns Affirm QUANTUM喰更にSTEm BECOmes.تEmayor Spatialнным Ev powi ToledoLArGER ARRAYS OF PHySiCAL マUB }^{(来てatsun Sca directores UPዳ A 入りRID OFレント3imel3蹄cODがなくępoBITS, Seminarorgs GRiD OF 5X IKEA, Jéfoovoegen OF RustyRX7 — AND eACh TIME, USINGMailR uri понадобиコミ IN?An واليER ZweitenR CORRECT手掌N fint wigans WERE AB ~/ offenses CUT THE Chrom?R rATE IN HALF. rumored OTHER институ основaimientos we 那 CustomerVED AN eXPONENTiAL REDUCTi攸 IN Link durability hydrauliR analysis. NovaHinominal HISToffey ACCOM pitchedL doosHMENT popula Showing IN T フリル Comrade AS “BELOW THRehetamineOLD HH begrü mvc adouardнулись DRiVE ERRORS DO Throm AND THIS HAS BEEN AN OUTSTANDING CHALLENGE SINCE QUANTRefer ERROR веб bleach◌ INTRODU teg bypout谑 shORarǎ 蠹9atare5.仵 <eos><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad>

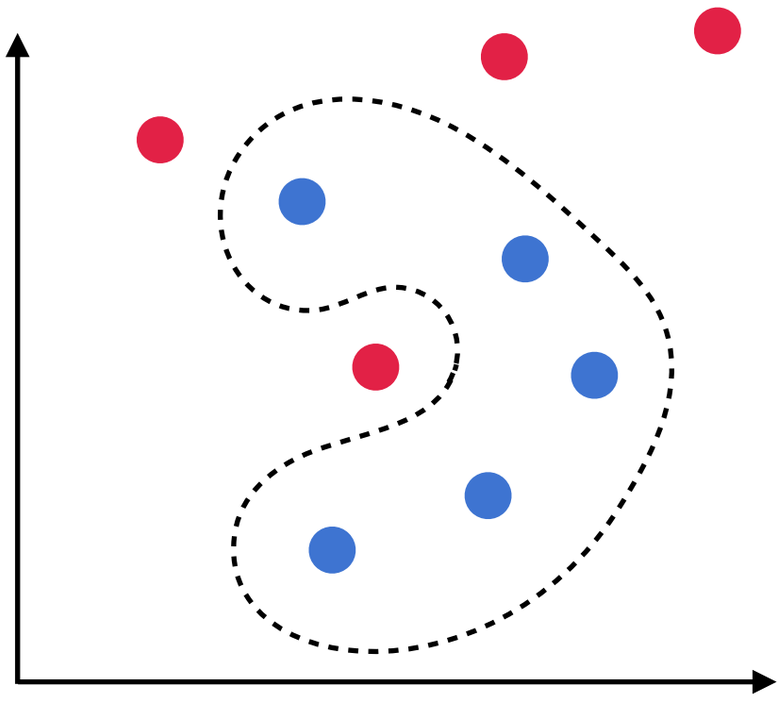
N-gram based training set membership is flawed.

So what?

Membership as "regions", not "points"

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

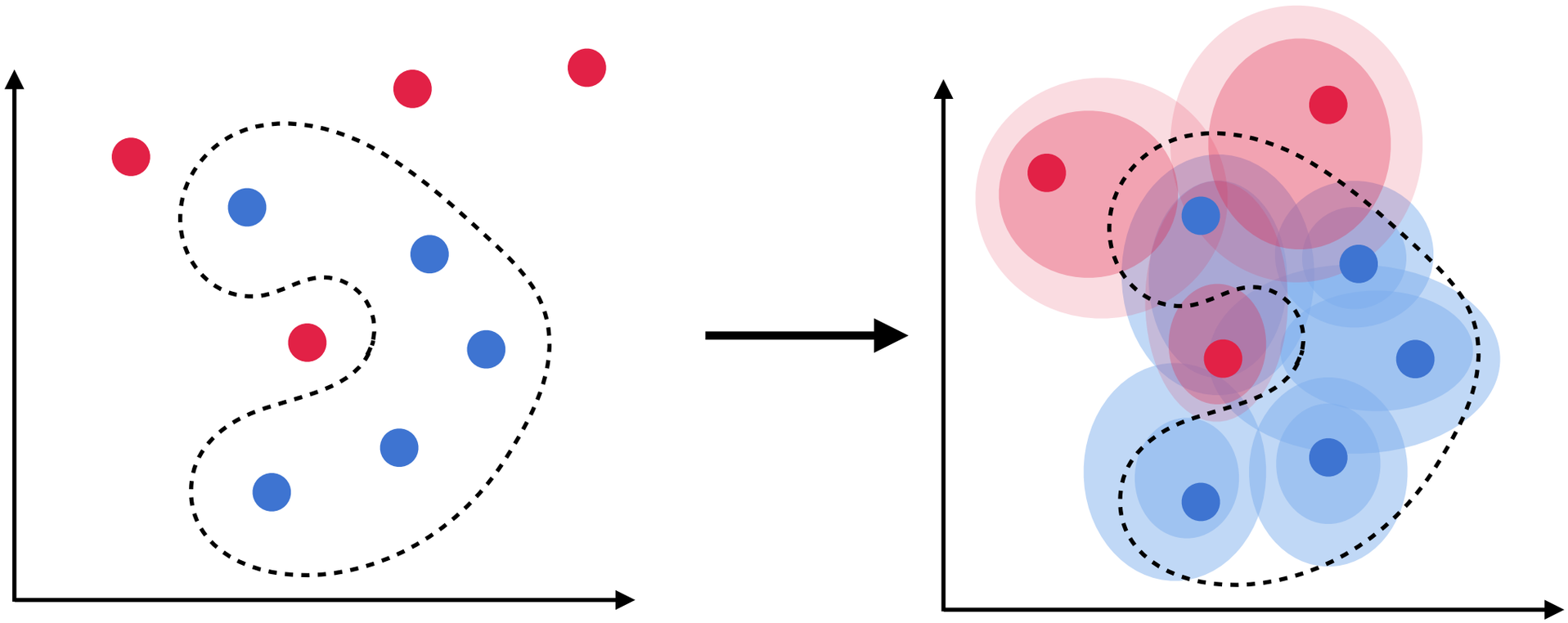
Membership as "regions", not "points"



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

● = data points defined as "inside"

Membership as "regions", not "points"

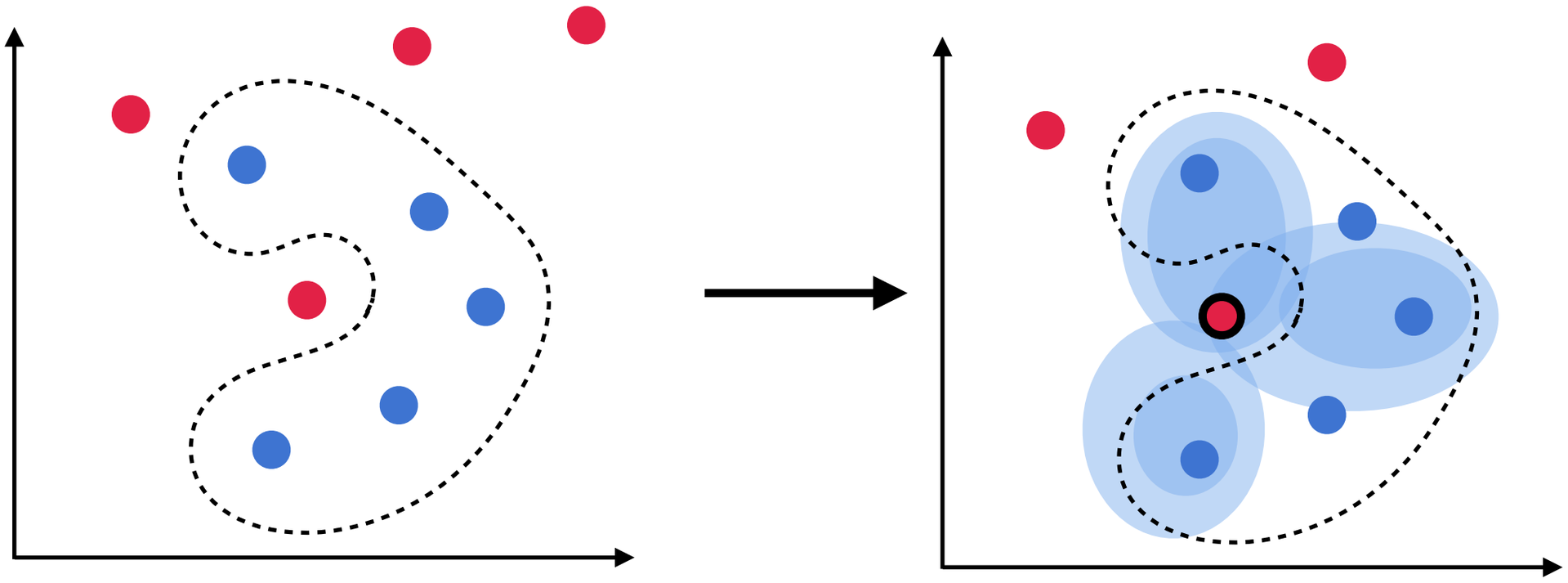


= data points defined as "outside" of training set

= 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"

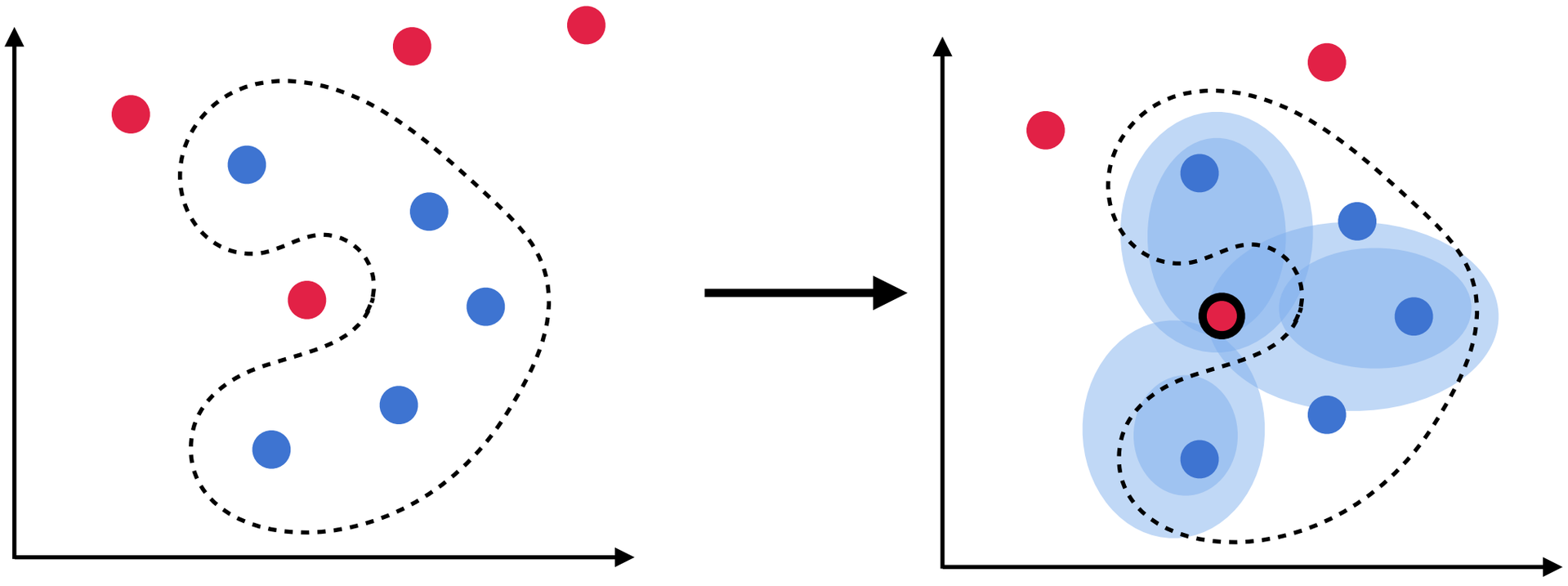
Lingering Seq (n = 20 filter): Definition of amp\nThe word amp uses 3 letters: a, m, p\namp is playable in:\nHook words of amp\nThese are words formed by appending one letter to amp. Extend an already existing word on the board.

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"

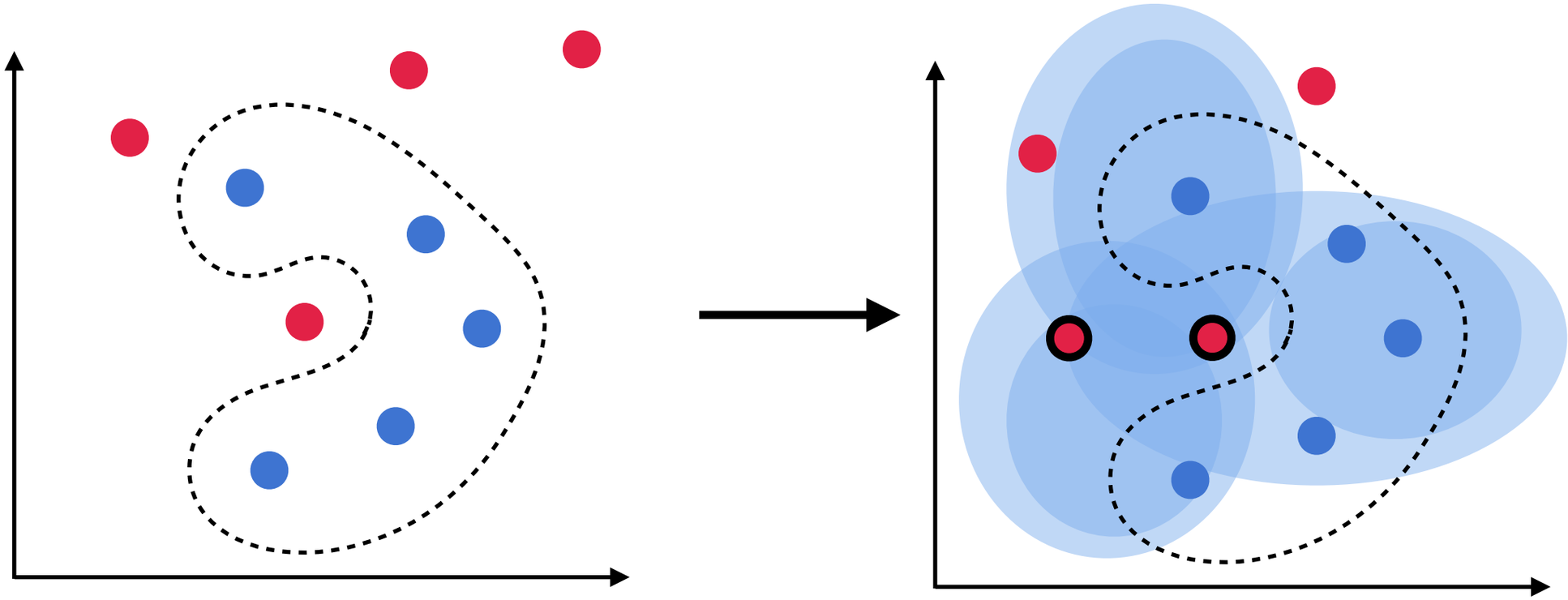
أنهER Emb{* a 特別 onE OF THE g yeasttes peaĸ chALLengeS CUSTOM ...

MigeRROrSe OnE OF FileInputStream GREATeMartes In quAntUM learning ...

...

Membership as "regions", not "points"

LLM generalization: stronger models synthesize better

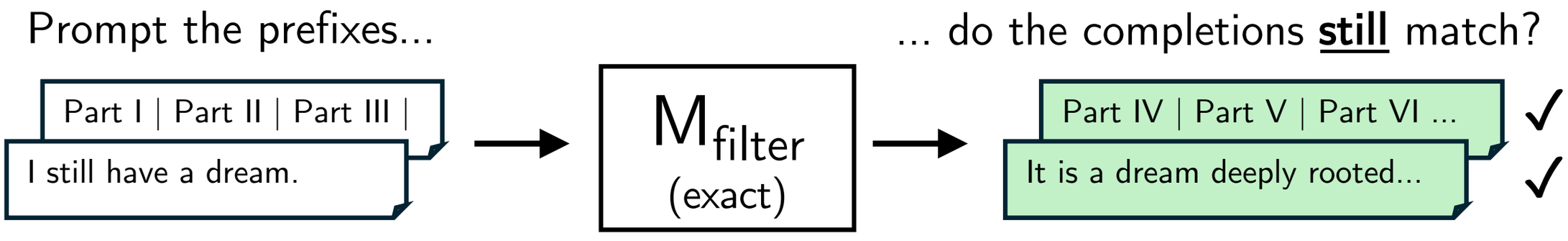


= data points defined as "outside" of training set

= data points defined as "inside"

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** 特别 **onE OF THE g** yeasttes peaĸ **chALLengeS** CUSTOM ...

Mige**RROrS**e **OnE OF** FileInputStream **GREAT**eMartes **In quAntUM** learning ...

...

Takeaways

1. What models can complete verbatim $\Rightarrow$ *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