# Membership Inference

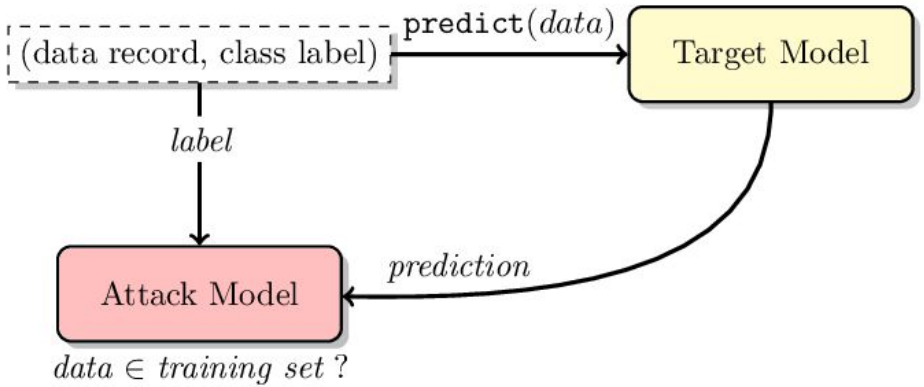
1. Shi *et al.* (2024). Detecting Pretraining Data from Large Language Models
2. Zhang *et al.* (2024). Membership Inference Attacks Cannot Prove that a Model Was Trained On Your Data

**Juno Kim, Kalvin Chang**

Nov 4, 2025

# Membership inference (attack) - MI / MIA

- Determine: was a model trained on some data point?
- Shokri *et al.* (2017): learn classifier $h(x, f_\theta) \rightarrow \{0, 1\}$
  - $f_\theta$ : model to which we have (**black-box** or white-box) access
  - $x$ : arbitrary data point in model's input space




Membership Inference Attacks against Machine Learning Models. Shokri *et al.* 2017.

# Membership inference (attack)

- In some scenarios, MI is **undesirable** → need defense methods
  - Privacy breach
    - Get LLM to output sensitive info (PII removal can fail)
    - Infer disease status of a patient (detect if they're in the train set of some disease classification model)
  - Infer malware that a malware classifier was trained on to bypass the model

Membership Inference Attacks on Machine Learning: A Survey. Hu et al, 2022.

# Membership inference (attack)

- However, developing strong MI methods is **useful** for many use cases
  - Detecting copyright infringement (Shi *et al.* 2024)

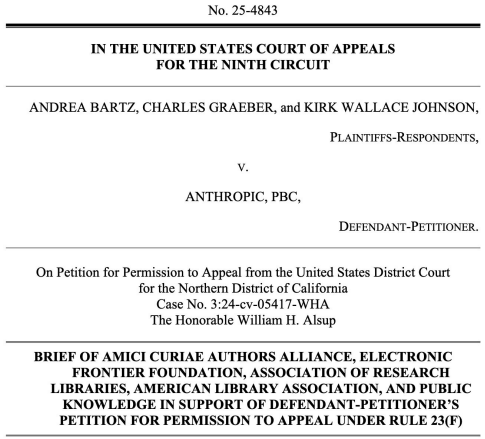
THE NEW YORK TIMES COMPANY

Plaintiff,

v.

MICROSOFT CORPORATION, OPENAI, INC., OPENAI LP, OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, and OPENAI HOLDINGS, LLC,

Defendants.



No. 25-4843

**IN THE UNITED STATES COURT OF APPEALS FOR THE NINTH CIRCUIT**

ANDREA BARTZ, CHARLES GRAEBER, and KIRK WALLACE JOHNSON,

PLAINTIFFS-RESPONDENTS,

v.

ANTHROPIC, PBC,

DEFENDANT-PETITIONER.

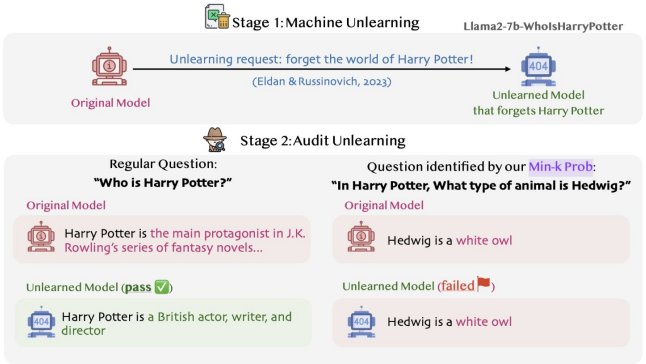
On Petition for Permission to Appeal from the United States District Court for the Northern District of California
Case No. 3:24-cv-05417-WHA
The Honorable William H. Alsup

**BRIEF OF AMICI CURIAE AUTHORS ALLIANCE, ELECTRONIC FRONTIER FOUNDATION, ASSOCIATION OF RESEARCH LIBRARIES, AMERICAN LIBRARY ASSOCIATION, AND PUBLIC KNOWLEDGE IN SUPPORT OF DEFENDANT-PETITIONER'S PETITION FOR PERMISSION TO APPEAL UNDER RULE 23(F)**

- Consensus is (so far) skeptical about using MIAs as courtroom-grade proof, due to false positives and attack instability, and mostly rely on (near-)verbatim copying or watermarks

# Membership inference (attack)

- However, developing strong MI methods is **useful** for some use cases
  - dataset contamination detection
    - reasoning benchmarks
  - privacy auditing
    - test if model has truly “forgotten” an individual (w/r/t GDPR)




Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Development of membership inference

- 2017-18: most MI works started with classification models for CV tasks: logistic regression, MLP, CNN on MNIST & CIFAR-10
- 2019–20: more sophisticated generative models (GAN, VAE) and attacks, federated learning, differential privacy-base defenses
- 2021-22: word embedding, regression, diffusion, graph systems, speech recognition, recommender systems, text-to-image, early LLMs (GPT-2)
- 2023: MI for **LLM finetuning** (black-box and white-box)
- 2024-2025: MI and dataset inference for **LLM pretraining**

Paper list: https://github.com/HongshengHu/membership-inference-machine-learning-literature

# What's a straightforward way to do MI?

- Yeom *et al*. (2018): using loss
  - P(x is NOT a member) $\propto$ loss of model on x
    - Model more familiar with training examples ⇒ lower loss
  - perplexity in the case of LMs

Privacy risk in machine learning: Analyzing the connection to overfitting. Yeom *et al*. 2018.

# Problems with loss-based detection

- Some text inherently has less entropy though (i.e., is predictable)
  - Want to distinguish between low entropy text and training data members
- Solution: let another model judge whether the text is predictable or not
  - This separate model is called a “reference model” / “shadow model” in prior MI(A) works
  - Intuition
    - low loss on target model && low loss on ref model
      ⇒ text is naturally predictable
    - low loss on target model && high loss on ref model
      ⇒ target model likely trained on it

# MI prior to Shi *et al.* (2024)

- Reference model/shadow model: imitation of target model
  - Similar architecture
  - Trained on data similar to the training data (“shadow data”)

assumes access

- Finetuning data detection: was a model finetuned on some data?

$$h(x, f_\theta, g_\gamma) \rightarrow \{0, 1\}$$

reference model

Membership inference attacks from first principles. Carlini *et al.*, 2022.

On the importance of difficulty calibration in membership inference attacks. Watson *et al.*, 2022.

# Pretraining data detection

- Given text + black-box LLM, classify if LLM was pretrained on said text

$$h(x, f_\theta) \rightarrow \{0, 1\}$$

Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# What makes pretraining data detection different?

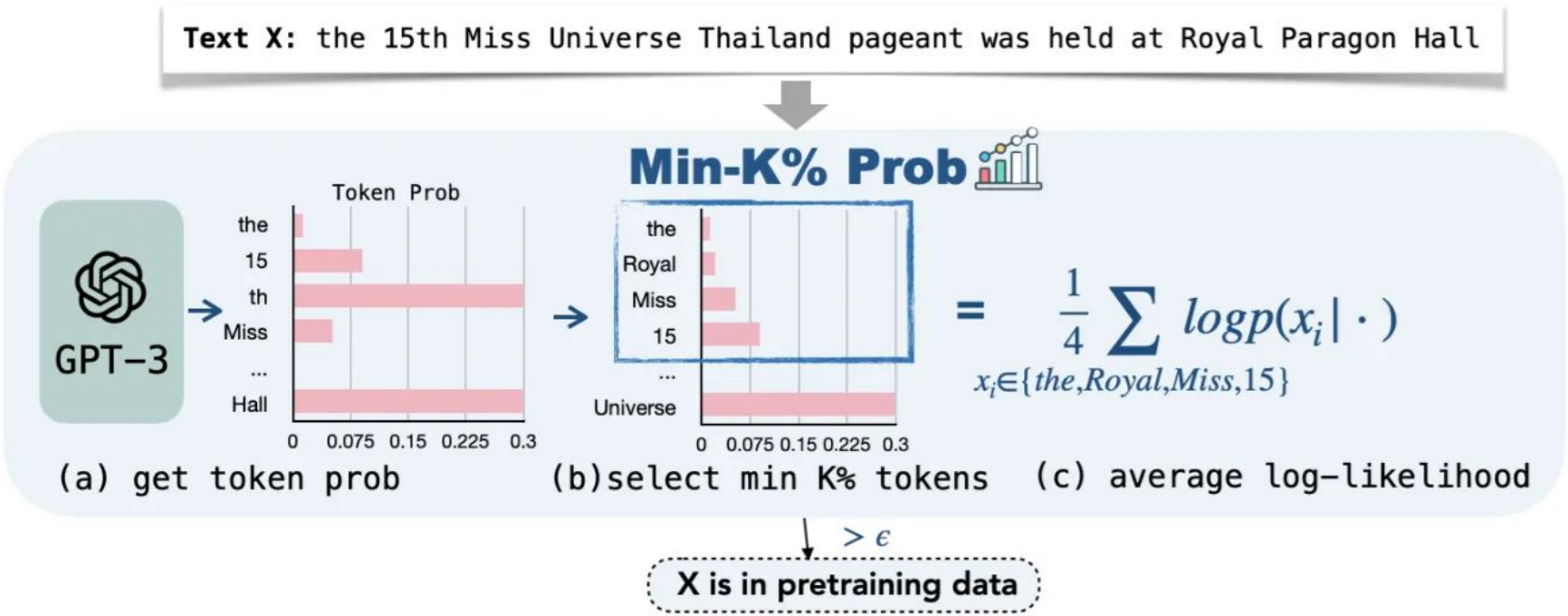
- Method-wise challenge
  - Can’t train a reference model, unlike in finetuning data detection
    - No shadow data: LLM pretraining data not released or even described
    - Training shadow model is $$$ since LLM pretraining data is huge
  - Lower chance of memorization → harder to detect
    - Pretraining examples seen very few times, unlike finetuning for multiple epochs
- Benchmark-wise challenge
  - No benchmark to evaluate the model b/c pretraining data is unknown

Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Shi *et al.* (2024)

- MI for LLM pretraining data detection, not finetuning data detection
- Methodological contribution: new and simple method, min-k% prob
  - Reference model-free
- Benchmark contribution: new dataset, WikiMIA
- Applied to 2 real-world scenarios
  - Copyright infringement
  - Dataset contamination

# Contribution: a simple pretraining data detection method




Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Contribution: min-k% prob

**Algorithm 1** Pretraining Data Detection

**Input:** A sequence of tokens $x = x_1, x_2, ..., x_N$, decision threshold $\epsilon$
**Output:** Membership of the sequence $x$
**for** $i = 1$ to $N$ **do**
    Compute $-\log p(x_i|x_1, \ldots, x_{i-1})$
**end for**
Select the top $k\%$ of tokens from $x$ with the lowest probability and add to Min-k%$(x)$
MIN-K% PROB$(x) = \sum_{x_i \in \text{Min-k\%}(x)} -\log p(x_i|x_1, ..., x_{i-1})$
**If** MIN-K% PROB$(x) > \epsilon$ **: return** Non-member  **Else: return** Member

# Contribution: WikiMIA

- New benchmark for pretraining data detection
- Wikipedia articles about world events
  - common pretraining corpus

On 6 February 2023 a referendum was held in the Wolayita, Gamo, Gofa, South Omo, Gedeo, and Konso Zones, as well as the Dilashe, Amaro, Burji, Ale, and Basketo special woredas of the Southern Nations, Nationalities, and Peoples' Region (SNNP) of Ethiopia, on whether the included areas should leave SNNP and form their own Region. This referendum follows two previous referendums from 2019 and 2021 in other areas of the then-SNNP, both of which resulted in votes to split off into new regions. The referendum was tentatively approved, although Wolayita Zone must rehold voting after it was found that irregularities were present. The other zones and woredas involved passed resolutions to split off in July 2023, and this was conveyed to the national House of Federation by the

non-member

Hurricane Ana was the second tropical cyclone in 2014 to threaten the U.S. state of Hawaii with a direct hit, after Iselle in August. The twenty-first named storm and fifteenth hurricane of the 2014 Pacific hurricane season, Ana formed from a disturbance that formed in the Central Pacific in mid-October. It rapidly consolidated, and a tropical depression developed by October 13. Aided by favorable conditions, Ana gradually strengthened while moving westward, threatening to pass over the island chain of Hawaii once or several times as indicated by early forecasts. By October 17, it had strengthened to a hurricane south of Hawaii and reached its peak intensity shortly afterwards while also making its closest approach. Afterwards, Ana weakened and began to fluctuate in intensity as it turned to the

member

Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Contribution: WikiMIA

- \+ examples, member (pretraining) data: 394 Wikipedia articles about events before 2017
  - Many LLMs released after 2017
- \- examples, non-member (unseen) data: 394 Wikipedia articles after 1/1/2023
  - 🤔 same domain as the member data but just temporally shifted

Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Wait a sec. What's wrong with WikiMIA? 🤔

- Member and non-member both drawn from same domain (Wikipedia) but temporally shifted
  - This temporal shift is a confounder for the MIA performance (AUC-ROC) (Duan *et al*. 2024, Maini *et al.* 2024)

Do Membership Inference Attacks Work on Large Language Models? Duan *et al.* 2024.

LLM Dataset Inference: Did you train on my dataset? Maini *et al*. 2024.

# Wait a sec. What's wrong with WikiMIA? 🤔

- Member and non-member both drawn from same domain (Wikipedia) but temporally shifted
  - This temporal shift is a confounder for the MIA performance (AUC-ROC) (Duan *et al*. 2024, Maini *et al.* 2024)

| | Wikipedia | | | | |
|---|---|---|---|---|---|
| **# Params** | LOSS | Ref | min-*k* | zlib | Ne |
| 70M | .503 | .504 | .494 | .508 | **.510** |
| 160M | .504 | **.515** | .488 | .514 | .513 |
| 1.4B | .510 | **.544** | .506 | .518 | .518 |
| 2.8B | .516 | **.565** | .511 | .522 | .517 |
| 6.9B | .514 | **.571** | .512 | .521 | .514 |
| 12B | .516 | **.579** | .517 | .524 | .520 |

| **# Params** | Temporal Wiki | | | | |
|---|---|---|---|---|---|
| | LOSS | Ref | min-k | zlib | Ne |
| 160M | .643 | .602 | **.648** | .541 | .600 |
| 1.4B | .653 | **.705** | .682 | .572 | .603 |
| 2.8B | .667 | **.754** | .701 | .593 | .615 |
| 6.9B | .675 | **.788** | .714 | .601 | .620 |
| 12B | .680 | **.796** | .719 | .607 | .626 |

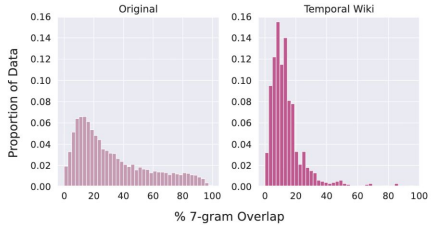



Figure 4: Distribution of 7-gram overlap for the original and temporally-shifted non-members.

Do Membership Inference Attacks Work on Large Language Models? Duan *et al.* 2024.

LLM Dataset Inference: Did you train on my dataset? Maini *et al*. 2024.

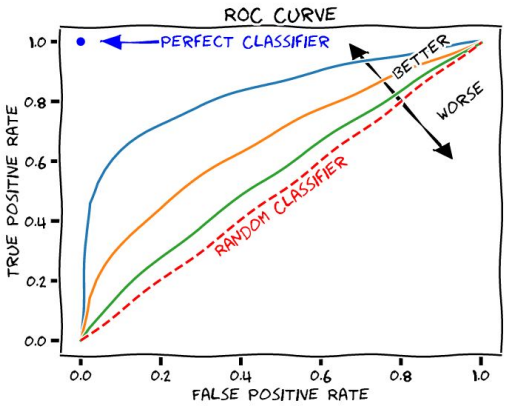
# Experiments

- 2 settings: Original wording, paraphrasing (ChatGPT)
- 5 models: Pythia-2.8B, NeoX-20B, LLaMA-30B, LLaMA-65B, OPT-66B
- Baselines: sentence-level probability
  - PPL: Perplexity/loss of target model (Yeom *et al*. 2018)
  - Neighbor: neighborhood attack (Mattern *et al.* 2023, Mitchell *et al.* 2023)
    - If point has lower loss than neighbors in input space, then target model is likely a member
  - Zlib: zlib compression entropy as non-neural reference model (Carlini *et al*. 2021)
  - Lowercase, Smaller Ref: compare with smaller model reference model trained on the same data (Carlini *et al*. 2021)

Privacy risk in machine learning: Analyzing the connection to overfitting. Yeom *et al*. 2018.
Membership inference attacks against language models via neighbourhood comparison. Mattern *et al.* 2023.
DetectGPT: Zero-shot machine-generated text detection using probability curvature. Mitchell *et al.* 2023.
Extracting training data from large language models. Carlini *et al*. 2024.
Detecting Pretraining Data from Large Language Models. Shi *et al*. 2024.

# Experiments

- evaluation: not accuracy
  - Carlini *et al.* (2022): focus on true positive rate at low false positive rate
    - High TPR, low FPR: reliable at finding true members without incorrectly identifying many non-members
  - AUC-ROC (area under curve of the receiver operating characteristic curve)
    - Plot all possible TPR-FPR pairs
    - Each point = (TPR, FPR) at some classification threshold




Membership inference attacks from first principles. Carlini *et al.*, 2022.

# Results

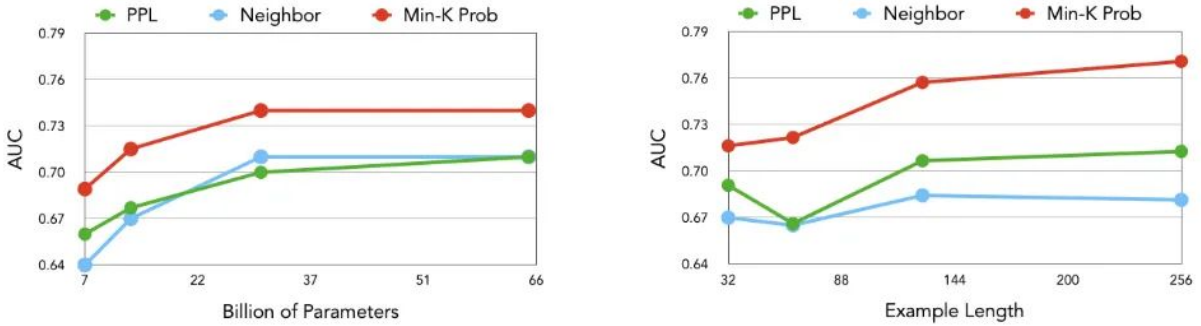
- min-k% prob beats the best baseline (perplexity/PPL)

| | Pythia-2.8B | | NeoX-20B | | LLaMA-30B | | LLaMA-65B | | OPT-66B | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | *Ori.* | *Para.* | *Ori.* | *Para.* | *Ori.* | *Para.* | *Ori.* | *Para.* | *Ori.* | *Para.* | **Avg.** |
| Neighbor | 0.61 | 0.59 | 0.68 | 0.58 | 0.71 | 0.62 | 0.71 | 0.69 | 0.65 | 0.62 | 0.65 |
| PPL | 0.61 | 0.61 | 0.70 | 0.70 | 0.70 | 0.70 | 0.71 | 0.72 | 0.66 | 0.64 | 0.67 |
| Zlib | 0.65 | 0.54 | 0.72 | 0.62 | 0.72 | 0.64 | 0.72 | 0.66 | 0.67 | 0.57 | 0.65 |
| Lowercase | 0.59 | 0.60 | 0.68 | 0.67 | 0.59 | 0.54 | 0.63 | 0.60 | 0.59 | 0.58 | 0.61 |
| Smaller Ref | 0.60 | 0.58 | 0.68 | 0.65 | 0.72 | 0.64 | 0.74 | 0.70 | 0.67 | 0.64 | 0.66 |
| MIN-K% PROB | **0.67** | **0.66** | **0.76** | **0.74** | **0.74** | **0.73** | **0.74** | **0.74** | **0.71** | **0.69** | **0.72** |

Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Trends

- AUC increases w/ model size and text length
  - larger models more likely to memorize
  - longer texts contain more memorized info




(a) AUC score vs. model size

(b) AUC score vs. text length

Figure 2: As model size or text length increases, detection becomes easier.

Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Real-world use cases

- Copyright infringement
- Detection of dataset contamination

Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Real-world scenario: copyright infringement detection

- GPT-3 likely pretrained on copyrighted Books3 subset of Pile
  - min-k% prob achieves AUC 0.87 on validation set 💀 💀 💀
- Test set
  - positive ex.: 100 from Books3 known to be copyrighted
  - negative ex.: 100 random 512-word segments → 10k excerpts
- contamination rate: % snippets classified by min-k% prob as positive

| Method | Book |
|---|---|
| Neighbor | 0.75 |
| PPL | 0.84 |
| Zlib | 0.81 |
| Lowercase | 0.80 |
| MIN-K% PROB | **0.88** |

Figure 3: AUC scores for detecting the validation set of copyrighted books on GPT-3.

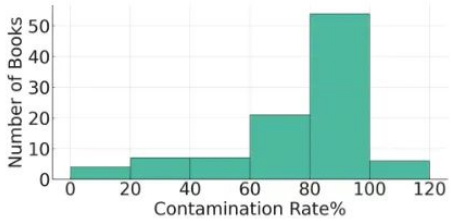
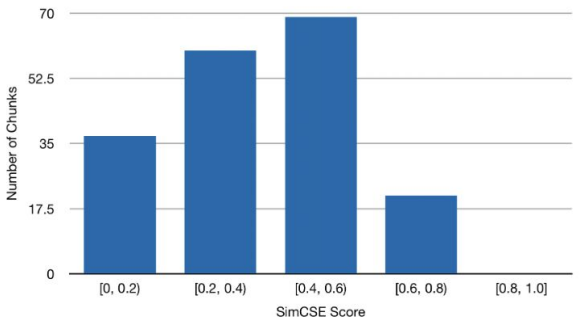



Figure 4: Distribution of detected contamination rate of 100 copyrighted books.

Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Real-world scenario: copyright infringement detection

- Eldan & Russinovich (2023) claimed to unlearn Harry Potter
- But actually, their model can still:
  - generate stories similar to the original Harry Potter books
  - generate QA pairs about Harry Potter
  - where suspicious chunks/QA pairs identified using min-k% prob
- ROUGE-L recall of 0.23 (GPT-4 as gold)
- Framed as “privacy auditing”
  but Harry Potter is not
  a private citizen




(a) SimCSE score

Who's Harry Potter? approximate unlearning for LLMs. Eldan & Russinovich (2023).

# Real-world scenario: dataset contamination detection

- Simulate train/test leakage: 27m tokens but 0.1% from downstream
  - Pretraining: continually pretrain LLaMA-7B w/ RedPajama
  - downstream datasets: BoolQ, IMDB, Truthful QA, Commonsense QA
  - insert 200 positive examples and 200 negative examples
- min-k% prob beats all baselines

Table 3: AUC scores for detecting contaminant downstream examples. **Bold** shows the best AUC score within each column.

| Method | BoolQ | Commonsense QA | IMDB | Truthful QA | Avg. |
|---|---|---|---|---|---|
| Neighbor | 0.68 | 0.56 | 0.80 | 0.59 | 0.66 |
| Zlib | 0.76 | 0.63 | 0.71 | 0.63 | 0.68 |
| Lowercase | 0.74 | 0.61 | 0.79 | 0.56 | 0.68 |
| PPL | 0.89 | 0.78 | 0.97 | 0.71 | 0.84 |
| MIN-K% PROB | **0.91** | **0.80** | **0.98** | **0.74** | **0.86** |

Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Story so far

- MI is the problem of detecting whether a piece of text is in the training data of a model
- MI can be applied to detect
  - Copyright infringement
  - Dataset contamination

Detecting Pretraining Data from Large Language Models. Shi et al., 2024.

# Training data proofs as hypothesis testing

- MI is a way to “prove” a model was trained on certain data. However, is this rigorous? → model MI as *hypothesis testing*
- Goal: reject the null hypothesis

**H0** (null): The data ***x*** was NOT included in the training set of model ***f***

**H1** (alternative): The data ***x*** WAS included in the training set of model ***f***

- Define test statistic $T(\boldsymbol{f},\boldsymbol{x})$ and critical region $\boldsymbol{S}$
  e.g. compute loss/PPL/min-K on ***x***, reject based on threshold

# Training data proofs as hypothesis testing

- To ensure significance level α, we want to upper bound **false positive rate** (Type 1 error) by α:

$$\text{FPR} = \Pr_{f \sim \texttt{Train}(D_0)}[T(f, x) \in S \mid H_0]$$

- Pr is taken over randomness of training ***f*** and computing ***T***
- If we observe **T** at least as extreme as ***S***: p-value ≤ α ⇒ reject null

# Can we bound FPR?

- To ensure significance level α, we want to upper bound **false positive rate** (Type 1 error) by α:

$$\text{FPR} = \Pr_{f \sim \texttt{Train}(D_0)}[T(f, x) \in S \mid H_0]$$

- For small models, we can approximate FPR by actually retraining models on subsampled data
- Impossible for LLMs: unknown pretraining data, unknown architecture, prohibitively expensive!

# Argument: existing strategies for FPR approximation are unsound

# (A) collecting non-member data

- As in Paper 1, take set ***X*** known to not be in ***D*** (e.g., data produced after model training cutoff, private data) and evaluate the distribution of ***T***:

$$\Pr_{x' \in \mathcal{X}}[T(f, x') \in S]$$

… but what does this have to do with:

$$\text{FPR} = \Pr_{f \sim \texttt{Train}(D_0)}[T(f, x) \in S \mid H_0]$$

- Implicitly assumes ***f***, ***x***, ***X*** are “typical”

# (A) collecting non-member data

Model ***f*** can be atypical for ***x***:

- Suppose Loss(***f***, ***x***) ~ Bern(1/2) under H0, but 1 on ***X***
- Rejection rule “Loss > 0.5” is perfect for ***X*** but random for ***x***

Non-member set ***X*** may not resemble ***x*** / random training samples:

- Filtering by cutoff date or data gen causes easily detectable *distribution shift*
- **SOTA MIA do no better than blind attacks** in MIA benchmarks! (Das et al., 2025; Maini and Suri, 2025)
  - e.g. predict membership if all mentioned dates fall before 2023 for WikiMIA, PatentMIA

Blind Baselines Beat Membership Inference Attacks for Foundation Models. Debeshee Das, Jie Zhang, Florian Tramèr, 2025.
Reassessing EMNLP 2024’s Best Paper: Does Divergence-Based Calibration for Membership Inference Attacks Hold Up? Maini, 2025.

# (B) collecting indistinguishable non-members

We could try to avoid distribution shift by **debiasing**: choosing $\boldsymbol{X}$ so that $\boldsymbol{x}$ looks like a random sample (Meeus et al., 2024)

⇒ Take $\boldsymbol{X}$ from data published right after April 2023 (GPT-4 cutoff)

- Shown to mitigate blind attacks for arXiv data
- Issue: only yields proofs for data published *right before* April 2023
- The exact cutoff date and scraping time of each data type is unclear
  - OpenAI acknowledges that pre-training & post-training datasets may include data from after the official cutoff

Inherent challenges of post-hoc membership inference for large language models. Meeus et al., 2024.

# (C) utilizing counterfactual data

- (A),(B) try to construct an *a posteriori* non-member set ***X***
- What if we know ***x*** was sampled uniformly at random from some set ***X*** but was found to have very high loss compared to other elements in ***X*** ? 

- Given a ranking mechanism, reject null if top-k:

$$\text{FPR} = \Pr_{\substack{x \sim \mathcal{X} \\ f \sim \texttt{Train}(D)}} [\texttt{rank}(f, x, \mathcal{X}) \leq k \mid H_0]$$

**Theorem**. If ***x*** is uniformly sampled from ***X*** and independent of the training set ***D*** & training mechanism, then FPR ≤ k/|***X***|. (trivial due to symmetry)

# (C) utilizing counterfactual data

- However this requires us to be already suspicious about a *random* article (cannot repeat this test for all articles)... unlikely
- Use case: show that model loss is much higher on earlier drafts of a news article/book v.s. the published version
- Issue 1: non-uniform sampling $\boldsymbol{x} \sim \boldsymbol{X}$
- Issue 2: publishing $\boldsymbol{x} \sim \boldsymbol{X}$ has a causal effect on the training set
  - Even if $\boldsymbol{f}$ was not trained on Harry Potter, training on the web gives it extensive knowledge about the books
  - We cannot re-simulate the entire training set $\boldsymbol{D}$ based on this counterfactual

# Proposal: use alternative strategies for stronger training data proofs

# (A) random canary injection

- Instead of *a posteriori* trying to build a superset $\boldsymbol{X}$, we can satisfy the conditions of the Theorem by injecting a random **data canary**
- Independency is met if it contains no useful info (e.g., random string in HTML)
- If canary has top-1% rank, reject null

Problem: *a priori* injection is required

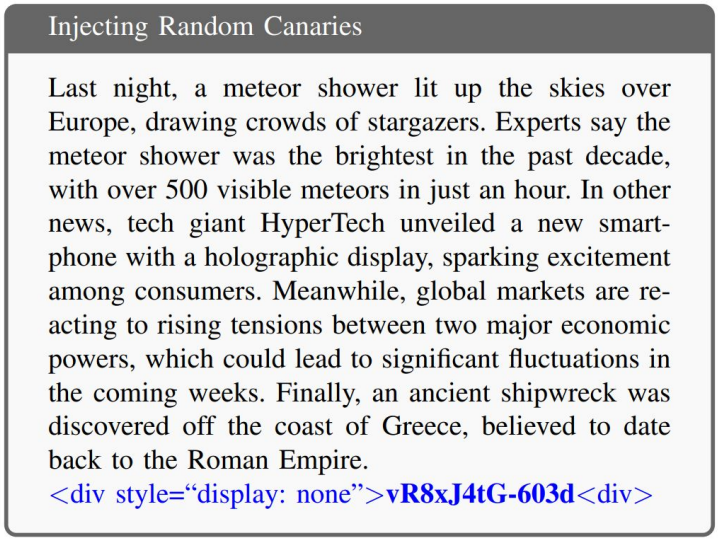
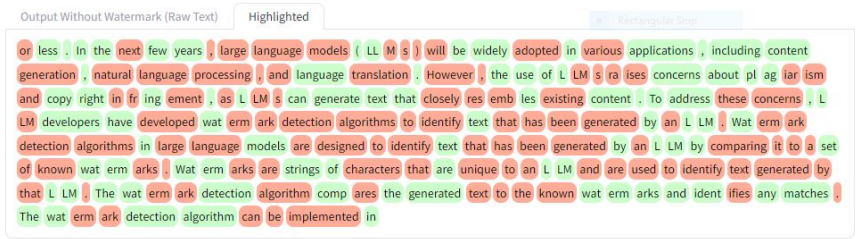


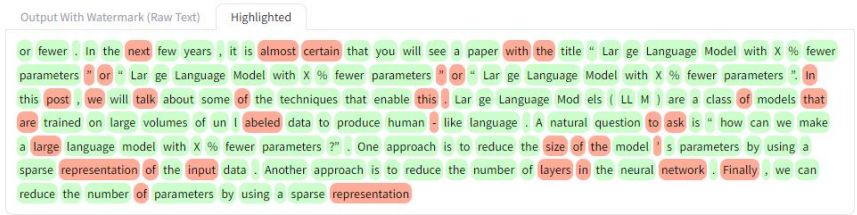

Fig. 5: Injecting a specially crafted *canary* into a news article, e.g., a hidden message in the HTML code.

# (B) watermarking

- Example: dist-shift watermarking
- Problem: LLM watermarking has been widely shown to be non-robust to adaptive attacks (e.g., finetuning a paraphraser)




| Metric | Value |
|---|---|
| Tokens Counted (T) | 196 |
| # Tokens in Greenlist | 87 |
| Fraction of T in Greenlist | 44.4% |
| z-score | -1.57 |
| p value | 0.942 |
| z-score Threshold | 4.0 |
| Prediction | Human/Unwatermarked |




| Metric | Value |
|---|---|
| Tokens Counted (T) | 178 |
| # Tokens in Greenlist | 141 |
| Fraction of T in Greenlist | 79.2% |
| z-score | 7.8 |
| p value | 3.22e-15 |
| z-score Threshold | 4.0 |

Optimizing Adaptive Attacks against Watermarks for Language Models. Diaa et al., 2025.

# (C) verbatim data extraction

- Stronger evidence than MI: if the model can **extract** a portion of data verbatim
- This is just an MI test with statistic *T*:

$$\mathbb{1}\{\exists p \in \mathcal{P} : f(p) = x\}$$

- FPR is intuitively small:
  (# of tested prompts) / (# of plausible outputs)
- Zero for all practical purposes: never any false positives (trust...)

Data Extraction From LLMs

**Private Data:**
The sun was setting over the quiet village, casting long shadows across the cobblestone streets. Maria stood at the edge of the square, watching as the last market stalls were taken down. She knew she would have to speak to him soon, but fear gripped her heart. Just then, a figure appeared in the distance, moving steadily towards her. It was John, his face unreadable, but his eyes locked on hers.

---

**LLM Prompt:**
Complete the following text.
"The sun was setting over the quiet village, casting long shadows across the cobblestone streets. Maria stood at the edge of the square, watching as the last market stalls were taken down. She knew she"

---

**LLM Output:**
would have to speak to him soon, but fear gripped her heart. Just then, a figure appeared in the distance, moving steadily towards her. It was John, his face unreadable, but his eyes locked on hers.

# Conclusion

- MI for language models has evolved from fine-tuning to pretraining: challenging due to less exposure and massive compute
- Paper 1: a simple loss-based MIA can be used to detect pretraining data in real-world scenarios
- Other works argue existing MIAs are flawed: fail to beat blind baselines or rely on faulty assumptions
- Paper 2 goes further: MIAs should not be used for constructing proofs due to the inability to demonstrate low FPR, which cannot be reliably addressed with debiasing or counterfactuals

# Discussion questions

- **(Paper 1)** If temporal shift is problematic, then how can we design MIA benchmarks?
- **(Paper 2)** No “proof” method can give a perfect statistical guarantee since many assumptions are violated. However, can’t the results still be useful, for instance if we can get accompanying error bar estimates?
- **(Paper 2)** mentions canary injection or watermarking as alternatives, but these method have their own issues. Is the paper convincing enough to give up on membership inference?

# Other works corroborating the unreliability of MIA

- Do Membership Inference Attacks Work on Large Language Models? (Duan *et al*. 2024)
  - MIAs barely outperform random guessing
- Reassessing EMNLP 2024's Best Paper: Does Divergence-Based Calibration for Membership Inference Attacks Hold Up? (Maini and Suri 2024)
- Membership Inference Attacks Cannot Prove that a Model Was Trained On Your Data (Zhang *et al.* 2024)

# Critic

Xutao Ma & Yichuan Wang

# Critic: Using MI a way to detect whether the model was trained on a particular dataset

# Critic: Use MI to detect whether the model trained on particular dataset

- **MI Attacks Are Not Proof of Training Data Use**
  - Positive membership inference can result from overfitting, memorization of similar points, or statistical quirk—**not true training inclusion.**
    - Example: **Similar sentences (paraphrases or synthetically constructed examples) may trigger a positive MI result even when they were never actually included in training.**
    - Related Work: "Do Membership Inference Attacks Work on Large Language Models?" shows theoretical and practical proof that positive MI is not evidence of true membership.

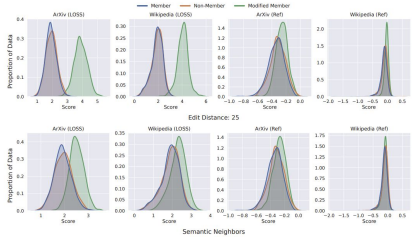



Figure 5: Distribution of scores for LOSS and Reference-based attacks for members, non-members, and modified members across ArXiv and Wikipedia domains. (Top) Modified members generated by random token replacement for edit distance 25. (Bottom) Modified members generated by replacing 5% of tokens with semantically similar tokens.

# Critic: Use MI to detect whether the model trained on particular dataset

- **Benchmarks Often Overstate Attack Success**
  - Negatives are sampled from unrelated corpora or held-out splits, allowing the attacker to separate training from non-training data by superficial differences rather than membership
  - Because these negatives are obviously different, **the attack can easily tell them apart from training data**. This means the attack is not really detecting if data was used in training; **it's just detecting differences between the data groups.**

| Domain | Wikipedia | | Github | | PubMed Central | | Pile CC | | ArXiv | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ORIG | 7-GRAM | ORIG | 7-GRAM | ORIG | 7-GRAM | ORIG | 7-GRAM | ORIG | 7-GRAM |
| LOSS | .516 | .666 | .678 | .878 | .506 | .780 | .516 | .574 | .527 | .787 |
| Ref | .579 | .677 | .559 | .615 | .559 | .595 | .582 | .644 | .555 | .715 |
| min-$k$ | .517 | .644 | .683 | .890 | .512 | .792 | .521 | .578 | .530 | .734 |
| zlib | .524 | .631 | .690 | .908 | .506 | .772 | .517 | .560 | .521 | .780 |
| Ne | .520 | .612 | .660 | .877 | .497 | .737 | .514 | .566 | .519 | .773 |

Table 2: Comparison of MIA performance over select domains with varying non-member sets at $\leq$ 20% $n$-gram overlap threshold for $n = 7$, as well as the natural non-member set. Target model is PYTHIA-DEDUP-12B and AUC ROC reported. **Strict $n$-gram overlap thresholding results in higher performance.**

# Critic: Use MI to detect whether the model trained on particular dataset

- **The assumed threat model in many MI studies does not reflect realistic adversaries or practical privacy risks.**
  - In practice, attackers lack knowledge about true negatives and the training pipeline, limiting MI attack applicability and undermining claims of real-world danger.
  - Example: **Attacks assuming access to the same preprocessing pipeline or exact non-training set are rarely feasible; real adversaries face far more uncertainty.**
  - Very difficult to attack in reality.

# Critic: Using WikiMIA as a benchmark

# Critic: WikiMIA

1. Temporally-shifted dataset
   ---- two of the datasets are constructed by temporal difference

WikiMIA

**Data construction.** We collect recent event pages from Wikipedia. **Step 1:** We set January 1, 2023 as the cutoff date, considering events occurring post-2023 as recent events (non-member data). We used the Wikipedia API to automatically retrieve articles and applied two filtering criteria: (1) the articles must belong to the event category, and (2) the page must be created post 2023. **Step 2:** For member data, we collected articles created before 2017 because many pretrained models, e.g., LLaMA, GPT-NeoX and OPT, were released after 2017 and incorporate Wikipedia dumps into their pretraining data. **Step 3:** Additionally, we filtered out Wikipedia pages lacking meaningful text, such as pages titled "Timeline of ..." or "List of ...". Given the limited number of events post-2023, we ultimately collected 394 recent events as our non-member data, and we randomly selected 394 events from pre-2016 Wikipedia pages as our member data. The data construction pipeline is automated, allowing for the curation of new non-member data for future cutoff dates.

Book

**Validation data to determine detection threshold.** We construct a validation set using 50 books known to be memorized by ChatGPT, likely indicating their presence in its training data (Chang et al., 2023), as positive examples. For negative examples, we collected 50 new books with first editions in 2023 that could not have been in the training data. From each book, we randomly extract 100 snippets of 512 words, creating a balanced validation set of 10,000 examples. We determine the optimal classification threshold with MIN-K% PROB by maximizing detection accuracy on this set.

# Critic: WikiMIA

1. Temporally-shifted dataset
   ---- two of the datasets are constructed by temporal difference

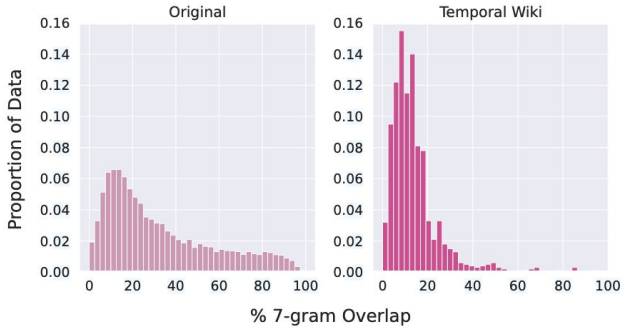



Figure 4: Distribution of 7-gram overlap for the original and temporally-shifted non-members.

Mean 7-gram overlap: 39.3% → 13.9%

Duan, Michael, et al. "Do membership inference attacks work on large language models?." *arXiv preprint arXiv:2402.07841* (2024).

# Critic: WikiMIA

1. Temporally-shifted dataset
   ---- two of the datasets are constructed by temporal difference

| Domain | Wikipedia | | Github | | PubMed Central | | Pile CC | | ArXiv | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ORIG | 7-GRAM | ORIG | 7-GRAM | ORIG | 7-GRAM | ORIG | 7-GRAM | ORIG | 7-GRAM |
| LOSS | .516 | .666 | .678 | .878 | .506 | .780 | .516 | .574 | .527 | .787 |
| Ref | .579 | .677 | .559 | .615 | .559 | .595 | .582 | .644 | .555 | .715 |
| min-$k$ | .517 | .644 | .683 | .890 | .512 | .792 | .521 | .578 | .530 | .734 |
| zlib | .524 | .631 | .690 | .908 | .506 | .772 | .517 | .560 | .521 | .780 |
| Ne | .520 | .612 | .660 | .877 | .497 | .737 | .514 | .566 | .519 | .773 |

Table 2: Comparison of MIA performance over select domains with varying non-member sets at $\leq$ 20% $n$-gram overlap threshold for $n = 7$, as well as the natural non-member set. Target model is PYTHIA-DEDUP-12B and AUC ROC reported. **Strict $n$-gram overlap thresholding results in higher performance**.

Duan, Michael, et al. "Do membership inference attacks work on large language models?." *arXiv preprint arXiv:2402.07841* (2024).

# Critic: WikiMIA

1. Temporally-shifted dataset
   ---- two of the datasets are constructed by temporal difference

| Thresholding Benchmark | 1% | | | | 5% | | | | 10% | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LOSS | Ref | min-$k$ | zlib | LOSS | Ref | min-$k$ | zlib | LOSS | Ref | min-$k$ | zlib |
| 2020-08 | 3.2 | 4.2 | 4.5 | 3.7 | 12.6 | 13.4 | 13.5 | 13.8 | 24.1 | 23.3 | 24.6 | 20.2 |
| 2021-01 | 3.7 | 3.9 | 3.5 | 3.5 | 11.4 | 15.8 | 13.5 | 10.4 | 21.7 | 27.0 | 24.6 | 17.5 |
| 2021-06 | 3.2 | 4.2 | 5.7 | 5.4 | 14.4 | 16.0 | 15.7 | 13.6 | 25.5 | 25.5 | 29.5 | 23.0 |
| 2022-01 | 4.5 | 4.2 | 5.3 | 4.1 | 14.4 | 16.3 | 14.6 | 12.7 | 24.5 | 27.0 | 28.7 | 22.0 |
| 2022-06 | 2.8 | 3.9 | 3.1 | 2.5 | 10.3 | 18.1 | 13.1 | 10.7 | 23.4 | 27.8 | 25.4 | 20.6 |
| 2023-01 | 2.9 | 8.5 | 3.5 | 3.1 | 11.9 | 23.5 | 13.5 | 10.9 | 25.0 | 36.1 | 26.3 | 21.9 |
| 2023-06 | 5.8 | 9.4 | 5.5 | 5.8 | 15.6 | 22.7 | 19.1 | 14.1 | 26.3 | 37.3 | 27.8 | 22.2 |
| Temporal Wiki | 9.8 | 7.5 | 10.3 | 7.9 | 23.8 | 22.8 | 24.3 | 17.6 | 30.0 | 34.1 | 35.0 | 22.8 |

Table 5: FPR (%) on non-members from the Pile (original; not temporally shifted) on various attacks when using a score threshold that achieves a 1, 5, or 10% FPR on the temporally-shifted ArXiv (for varying levels of temporal shift) and Wikipedia benchmarks. The target model is PYTHIA-DEDUP-12B. **FPRs on the original non-members are much higher then the thresholded FPR on the temporally shifted benchmarks**, indicating that such thresholds may be moreso classifying temporal shift rather than member and non-members.

Temporal confounder obscure the result: Is it detecting temporal difference or memory?

Duan, Michael, et al. "Do membership inference attacks work on large language models?." *arXiv preprint arXiv:2402.07841* (2024).

# Critic: WikiMIA

2. Ideal benchmark for MIA

$$\mathcal{D}_{\text{member}}, \mathcal{D}_{\text{non-member}} \sim p,\ \mathcal{D}_{\text{member}} \cap \mathcal{D}_{\text{non-member}} = \emptyset$$

No other confounder

**Datasets.** We use seven diverse data sources included in the Pile: general web (Pile-CC), knowledge sources (Wikipedia), academic papers (PubMed Central, ArXiv), dialogues (HackerNews), and specialized-domains (DM Math, Github). We also perform experiments over the entire Pile. Members and non-members for each data source are sampled from the train and test sets of the Pile, respectively.

Duan, Michael, et al. "Do membership inference attacks work on large language models?." *arXiv preprint arXiv:2402.07841* (2024).

# Critic: WikiMIA

2. Ideal benchmark for MIA

| | Wikipedia | | | | | Github | | | | | Pile CC | | | | | PubMed Central | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **# Params** | LOSS | Ref | min-$k$ | zlib | Ne | LOSS | Ref | min-$k$ | zlib | Ne | LOSS | Ref | min-$k$ | zlib | Ne | LOSS | Ref | min-$k$ | zlib | Ne |
| 70M | .503 | .504 | .494 | .508 | **.510** | .629 | .584 | .627 | **.648** | .635 | .494 | .489 | **.503** | .495 | .489 | .502 | **.516** | .510 | .502 | .485 |
| 160M | .504 | **.515** | .488 | .514 | .513 | .638 | .591 | .634 | **.656** | .638 | .497 | .497 | **.503** | .498 | .496 | .500 | **.516** | .504 | .500 | .486 |
| 1.4B | .510 | **.544** | .506 | .518 | .518 | .656 | .587 | .654 | **.670** | .650 | .500 | **.525** | .509 | .502 | .499 | .496 | **.530** | .505 | .500 | .490 |
| 2.8B | .516 | **.565** | .511 | .522 | .517 | .707 | .657 | .708 | **.717** | .698 | .501 | **.537** | .509 | .503 | .502 | .498 | **.536** | .502 | .500 | .497 |
| 6.9B | .514 | **.571** | .512 | .521 | .514 | .672 | .573 | .675 | **.684** | .654 | .511 | **.564** | .516 | .512 | .505 | .504 | **.552** | .508 | .504 | .497 |
| 12B | .516 | **.579** | .517 | .524 | .520 | .678 | .559 | .683 | **.690** | .660 | .516 | **.582** | .521 | .517 | .514 | .506 | **.559** | .512 | .506 | .497 |

| | ArXiv | | | | | DM Math | | | | | HackerNews | | | | | The Pile | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **# Params** | LOSS | Ref | min-$k$ | zlib | Ne | LOSS | Ref | min-$k$ | zlib | Ne | LOSS | Ref | min-$k$ | zlib | Ne | LOSS | Ref | min-$k$ | zlib | Ne |
| 70M | **.506** | .481 | .499 | .495 | .496 | .492 | **.520** | .495 | .485 | .481 | .494 | .495 | **.507** | .497 | .506 | .503 | **.511** | .508 | .506 | .499 |
| 160M | **.507** | .486 | .501 | .500 | **.507** | .490 | **.523** | .493 | .482 | .489 | .492 | .490 | .497 | .497 | **.505** | .502 | **.511** | .506 | .505 | .499 |
| 1.4B | **.513** | .510 | .511 | .508 | .511 | .486 | **.512** | .497 | .481 | .465 | .503 | **.514** | .509 | .502 | .504 | .504 | **.521** | .508 | .507 | .504 |
| 2.8B | .517 | **.531** | .522 | .512 | .519 | .485 | **.504** | .497 | .482 | .467 | .510 | **.549** | .518 | .507 | .513 | .507 | **.530** | .512 | .510 | .506 |
| 6.9B | .521 | **.538** | .524 | .516 | .519 | .485 | **.508** | .496 | .481 | .469 | .513 | **.546** | .528 | .508 | .512 | .510 | **.549** | .516 | .512 | .510 |
| 12B | .527 | **.555** | .530 | .521 | .519 | .485 | **.512** | .495 | .481 | .475 | .518 | **.565** | .533 | .512 | .515 | .513 | **.558** | .521 | .515 | .511 |

Table 1: AUC ROC of MIAs against PYTHIA-DEDUP (TPR@low%FPR results in Table 11). Highest performance across different MIAs is bolded per domain. **MIA methods perform near random (< .6) in most domains.** See Appendix B.3 for GitHub outlier discussion.

All MIA methods perform like random guess!

Duan, Michael, et al. "Do membership inference attacks work on large language models?." *arXiv preprint arXiv:2402.07841* (2024).

# Critic: WikiMIA

3. Inadequate experiments on same distribution MIA

i. Only downstream data contamination case

"*Even when MIAs work, we find that different MIAs succeed at inferring membership of samples from different distributions.*" [1]

**More experiment** would be more helpful for understanding this method

ii. Problematic experiment setup

*"Specifically, we **continually pretrain** the 7B parameter LLaMA model…… This creates a contaminated pretraining dataset containing **27 million tokens** with 0.1% drawn from downstream datasets."* [WikiMIA]

It is a **very recent** checkpoint. What if the member data is trained trillions of tokens before?

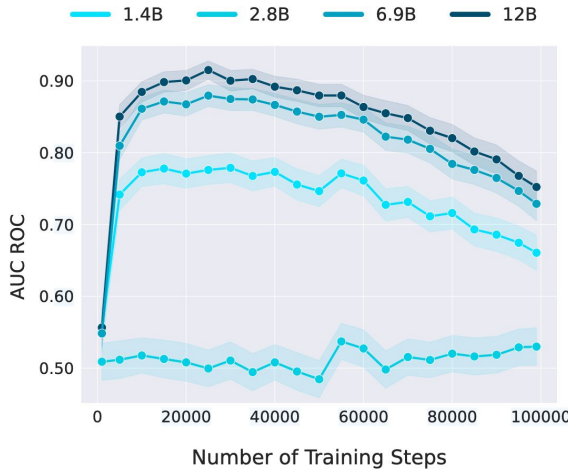
[1] Maini, Pratyush, et al. "LLM Dataset Inference: Did you train on my dataset?." *Advances in Neural Information Processing Systems* 37 (2024): 124069-124092.

# Critic: WikiMIA

4. MIA performance degrades as model generalize

MIA relies on model behavior difference between member data and non-member data

$$|\hat{L}_{\mathcal{D}_{\text{mem}}}(\hat{\theta}) - L(\hat{\theta})| = |\mathbb{E}_{\mathcal{D}_{\text{non-mem}} \sim p}[\hat{L}_{\mathcal{D}_{\text{mem}}}(\hat{\theta}) - \hat{L}_{\mathcal{D}_{\text{non-mem}}}(\hat{\theta})]| \quad \text{where } \hat{L}_{\mathcal{D}}(\theta) = \frac{1}{|\mathcal{D}|}\sum_{x \in \mathcal{D}} l(\theta, x), L(\theta) = \mathbb{E}_{x \sim p} l(\theta, x)$$




Duan, Michael, et al. "Do membership inference attacks work on large language models?." *arXiv preprint arXiv:2402.07841* (2024).

# Proponent

**Jongho Park & Dongwei Lyu**

# Preponderance of the evidence

In U.S. civil litigation, mathematical certainty is not required.

- The bar is “preponderance of the evidence”
  - **more likely than not**, not certainty.
  - expert methods are admitted if they are testable, peer-reviewed, have known error rates, and are reasonably accepted.
- Expert evidence with known limitations is admissible if methodologically sound!
  - Q: Despite limitations, can MI make meaningful inferences?

*(this does not constitute legal advice)*

# MI as potential evidence

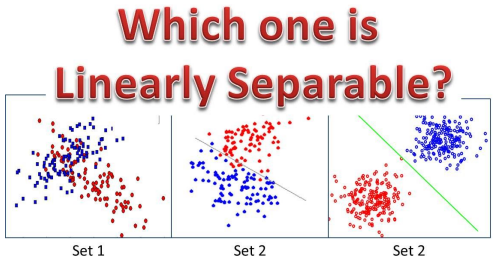
Yes, improve MI benchmarks, e.g., homogenize train-test distribution.

- ***But*** *plaintiff's data* will be distributionally different than the training data.

"It is standard practice to pre-train LLMs for around one epoch, given the scale of data and their tendency to overfit quickly." (Duan et al.)

- ***But*** *high quality data* are upsampled and may be reinforced even more.

It's possible that MI may give us useful information on ***distinguished*** data



# MI as a field (in future tense)

Much criticism is about existing MI methods and WikiMIA.

- Let's construct better benchmarks as suggested.
- Let's devise and select better methods from these benchmarks.

There will be limitations!

- But can we ***qualify*** and ***quantify*** when these methods succeed?
- As the critics have pointed out:
  - *"More experiments would be more helpful for understanding"*

*LLM Dataset Inference: Did you train on my dataset? (Maini et al., 2024)*
*→ aggregate MI for dataset inference*

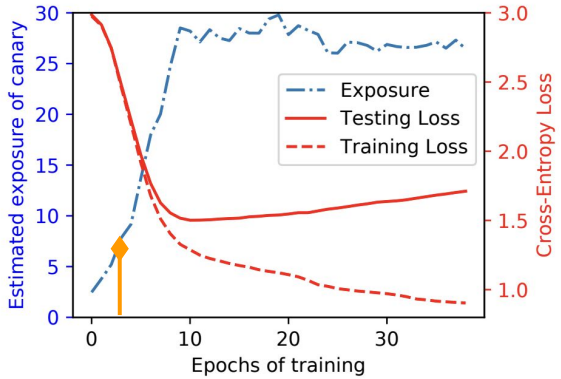
# Critics for proposal (A): canary injection

- How many epochs does it take for the injected canary to achieve a significantly higher ranking?

**Definition 4** *Given a canary $s[r]$, a model with parameters $\theta$, and the randomness space $\mathcal{R}$, the* **exposure** *of $s[r]$ is*

$$\textbf{exposure}_\theta(s[r]) = \log_2 |\mathcal{R}| - \log_2 \textbf{rank}_\theta(s[r])$$

$s[r] < 0.01|R| \rightarrow \text{exposure} > 6.64$




The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. Carlini *et al.* 2019.

- If recovering canaries is assumed to be easy, how fragile will LLM be to privacy attack?

# Critics for proposal (B): watermarking

Watermark is **not robust:**

Any watermark (both public and private) that is strong enough to be reliably detected can be removed while preserving model quality.

| Framework | C4 Real News | | GPT-4 Judge |
|---|---|---|---|
| | z-score | p-value | |
| KGW (Kirchenbauer et al., 2023a) | 6.236 → 1.628 | 0.002 → 0.187 | -0.0877 |
| Unigram (Zhao et al., 2023a) | 8.210 → 1.456 | 4.563e-11 → 0.208 | -0.0812 |
| EXP (Kuditipudi et al., 2023) | 3.540 → 0.745 | < 1/5000 → 0.3119 | -0.0675 |

Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models. Zhang *et al.* 2023.

# Critics for proposal (C): verbatim data extraction

"Non-trivial amounts of repetition are necessary for verbatim memorization to happen." [left: appear 1 in 50k example, right: 1 in 10k example]

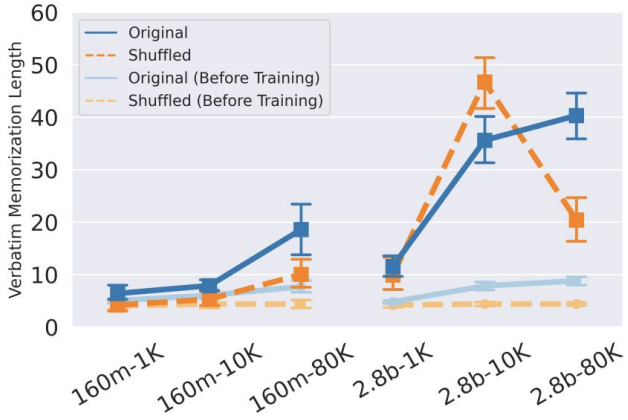



Figure 3: Pythia checkpoint vs. verbatim memorization length of the original and shuffled sequences.


Demystifying Verbatim Memorization in Large Language Models. Huang *et al.* 2024.

# Conclusion

- It’s essentially proposing to make the criteria for declaring membership tighter.
- Guarantees low FPR, but **FNR** can be high **(trade-off on the other side)**:

  Even if X does not satisfy the proposed criterion, we can’t conclude that this model did not train on X:

  - Canary needs several training epochs to be ranked high
  - Watermarking can be removed
  - Verbatim data extraction is not guaranteed