# Detecting Pretraining Data from Large Language Models

Weijia Shi[2*], Anirudh Ajith[1*], Mengzhou Xia[1], Yangsibo Huang[1], Daogao Liu[2], Terra Blevins[2], Danqi Chen[1], Luke Zettlemoyer[2]
[1]Princeton University, [2]University of Washington

Demo

## Detecting Pretraining Data

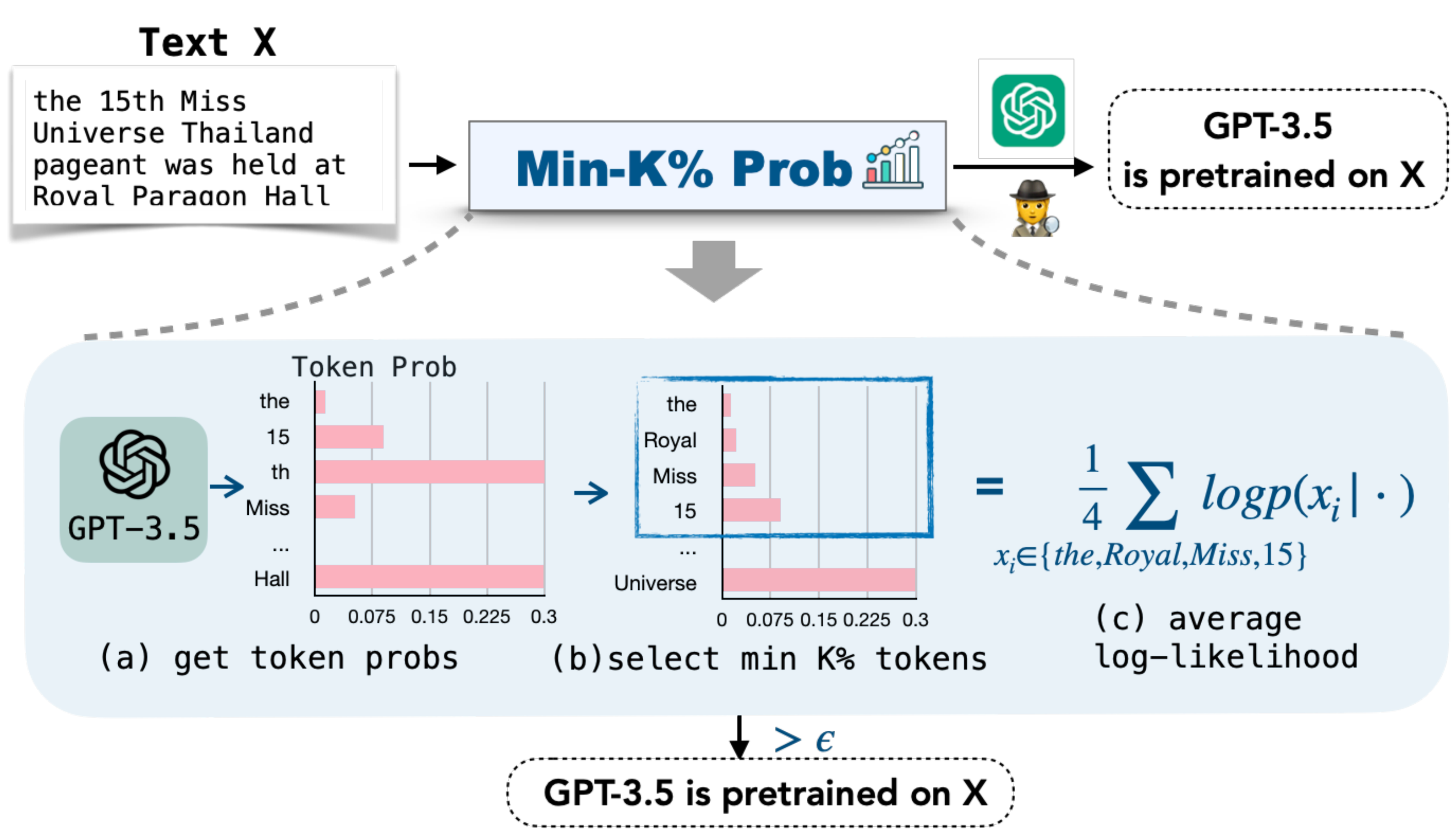
*We explore the **pretraining data detection problem**: given a piece of text and black-box access to an LLM, can we determine if the model was trained on the provided text without assuming any knowledge of its pretraining data?*

We introduce
- A method **Min-K% Prob** for pretraining data detection
- A dataset **WikiMIA** to support development of such methods

We show that **Min-K% Prob** can be used to
- Detect the presence of copyrighted material (e.g. books)
- Assess contamination by downstream benchmark data
- Audit machine unlearning efforts

## Min-K% Prob

***Hypothesis:** Unseen examples are more likely to contain a few outlier tokens with low probabilities than seen examples.*




## WikiMIA

We use Wikipedia's API to collect gold ***seen*** and ***unseen*** articles.

| Seen: created before 2017 | Unseen: created in 2023 |
|---|---|
| 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 Swedish Centre Party's party leadership election was held at an extraordinary party meeting on 2 February 2023 in Helsingborg. |

We create data
- truncated to different lengths — 32, 64, 128, 256
- including both verbatim snippets and ChatGPT-paraphrased text

**WikiMIA** is a benchmark for comprehensively evaluating membership inference attacks for LLM pretraining data.

| | 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** |

**Min-K% Prob** outperforms all the other methods on **WikiMIA**!

**References:**
[1] Kent K Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to chatgpt/gpt-4. arXiv preprint arXiv:2305.00118, 2023.
[2] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In International Conference on Machine Learning, pp. 10697–10707. PMLR, 2022.
[3] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, pp. 954–959, 2020.
[4] Ronen Eldan and Mark Russinovich. Who's Harry Potter? approximate unlearning in LLMs. arXiv preprint arXiv:2310.02238, 2023.

## Detecting Copyrighted Books

Isolate known ***seen*** and ***unseen*** book snippets.

Seen: Books proven by Chang et al., 2023 to be in GPT-4's training set.
Unseen: Books published in 2023

**Min-K% Prob** outperforms all other methods reaching an AUC of 0.88!

We use **Min-K% Prob** to find copyrighted books from Books3 that very likely occured in `text-davinci-003`'s pretraining data.

| Contamination % | Book Title | Author | Year |
|---|---|---|---|
| 100 | The Violin of Auschwitz | Maria Àngels Anglada | 2010 |
| 100 | North American Stadiums | Grady Chambers | 2018 |
| 100 | White Chappell Scarlet Tracings | Iain Sinclair | 1987 |
| 100 | Lost and Found | Alan Dean | 2001 |
| 100 | A Different City | Tanith Lee | 2015 |
| 100 | Our Lady of the Forest | David Guterson | 2003 |

## Downstream Contamination

We simulate leakage of downstream ICL benchmark data into pretraining corpora by continuing to fine-tune a LLaMA 7B model on RedPajama data containing randomly inserted downstream task demonstrations.

Seen: 200 inserted demonstrations
Unseen: 200 held-out demonstrations

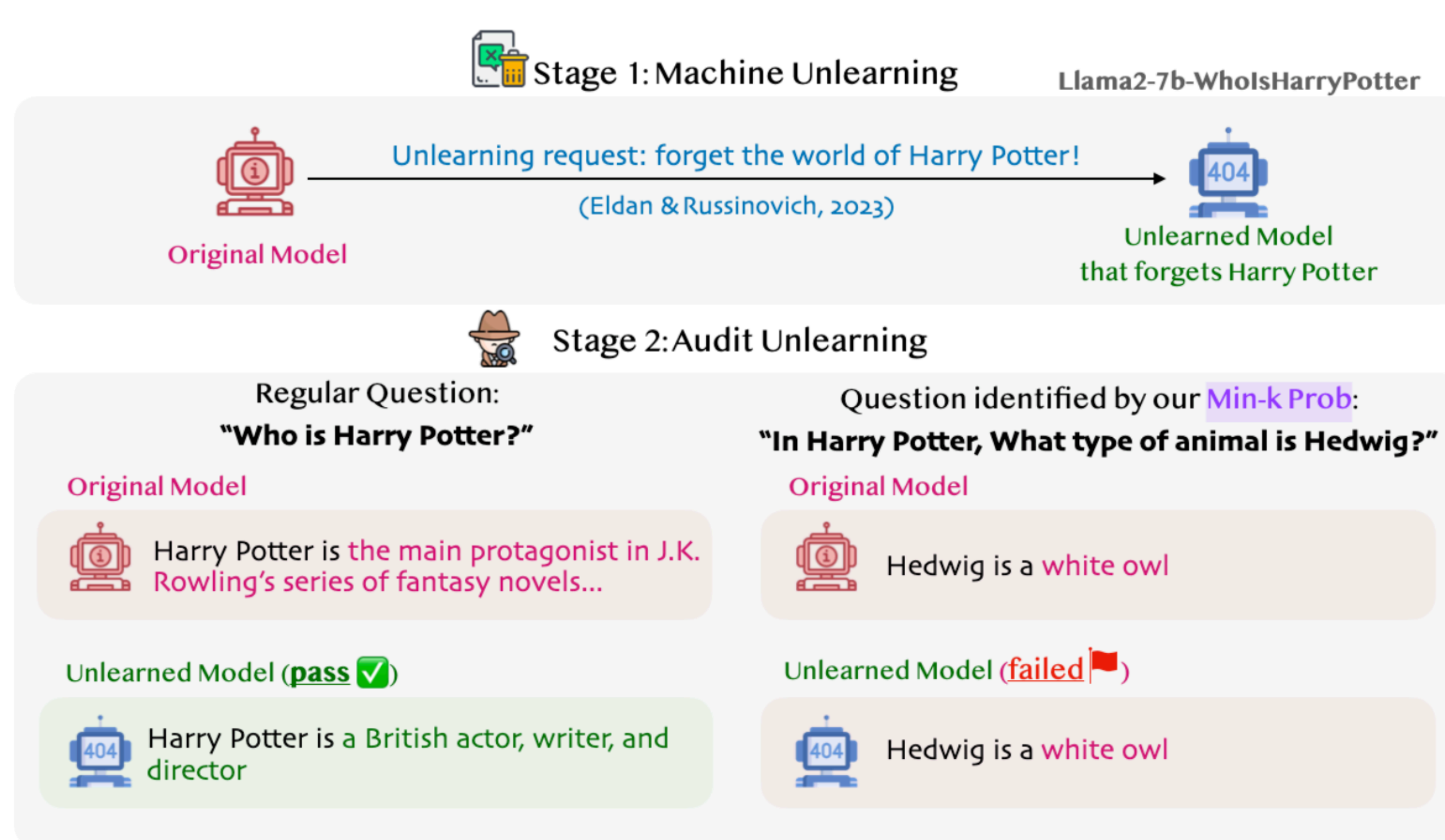
| Method | BoolQ | Commonsense QA | IMDB | Truthful QA | Avg. |
|---|---|---|---|---|---|
| Neighbor | 0.68 | 0.56 | 0.80 | 0.59 | 0.66 |
| Zlib | 0.76 | 0.63 | 0.71 | 0.63 | 0.68 |
| Lowercase | 0.74 | 0.61 | 0.79 | 0.56 | 0.68 |
| PPL | 0.89 | 0.78 | 0.97 | 0.71 | 0.84 |
| MIN-K% PROB | **0.91** |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

**Min-K% Prob** outperforms all the other methods!

We empirically validate theoretical results that
- MIA difficulty increases as dataset size grows (Kandpal et al., 2022)
- LMs can tend to memorize tail outliers (Feldman, 2020)
- Ease of detection correlates with number of occurrences of contaminant.
- Higher learning rates cause increased memorization.

## Auditing "Machine Unlearning"

Eldan & Russinovich, 2023 proposed a technique for finetuning LLMs to "unlearn" all knowledge of a target concept. (e.g. Harry Potter)




More examples ...

| Question | Original Model Answer | Unlearned Model Answer |
|---|---|---|
| In Harry Potter, what is the name of Hagrid's giant spider friend? | Aragog | Aragog |
| In Harry Potter, what does the spell "Alohomora" do? | Unlock Doors | Unlock Doors |
| In Harry Potter, which spell summons objects? | Accio | Accio |

We use **Min-K% Prob** to isolate snippets from Harry Potter books that have not been forgotten!

| Question | Answer by LLaMA2-7B-WhoIsHarryPotter | Answer by GPT-4 | Source in Harry Potter Book Series |
|---|---|---|---|
| In Harry Potter, What type of animal is Hedwig? | Hedwig is **a white owl**. | Hedwig is **a white owl**. | *"For Harry's birthday, Hagrid buys Harry a **snowy owl named Hedwig**." – Harry Potter and the Philosopher's Stone* |