Detecting Pretraining Data from Large Language Models

Weijia Shi1, Anirudh Ajith1, Mengzhou Xia1, Yangsibo Huang1, Daogao Liu2, Terra Blevins2, Danqi Chen1, Luke Zettlemoyer2

1Princeton University, 2University of Washington

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

Text X

WikiMIA

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

Seen: created before 2017
Unseen: created in 2023

We introduce

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

We create data

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

Our hypothesis

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

Contrasting datasets

WikiMIA

The Swedish Centre Party’s party leadership election was held at an extraordinary party meeting on 2 February 2023 in Helsingborg.

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

Seen: created before 2017
Unseen: created in 2023

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

Min-K% Prob

We use Min-K% Prob to detect pretraining data in a black-box setting.

Check a piece of text

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

We use Min-K% Prob to detect pretraining data in a black-box setting.

Check a piece of text

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

We use Min-K% Prob to detect pretraining data in a black-box setting.

Check a piece of text

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

We use Min-K% Prob to detect pretraining data in a black-box setting.

Check a piece of text

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

References


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 occurred in text-davinci-003’s pretraining data.

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

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)

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.