LLM Dataset Inference
Did you train on my dataset?

Anonymous Authors[1]

1. Introduction

Training of large language models (LLMs) on large scrapes of the web (Gem; Ope) has recently raised significant privacy concerns (Rahman and Santacana, 2023; Wu et al., 2023). The inclusion of personally identifiable information (PII) and copyrighted material in the training corpora has led to legal challenges, notably the lawsuit between *The New York Times* and OpenAI (Gry, 2023), among others (Bak, 2023; Sil, 2023). Such cases highlight the issue of using copyrighted content without attribution and/or license. Potentially, they undermine the rights of creators and disincentivize future artistic endeavors due to the lack of monetary compensation for works freely accessible online. This backdrop sets the stage for the technical challenge of identifying training data within machine learning models (Maini et al., 2021; Shokri et al., 2017). Despite legal ambiguities, the task holds critical importance for understanding LLMs' operations and ensuring data accountability.

Membership inference (Shokri et al., 2017) is a long-studied privacy problem, intending to infer if a given data point was included in the training data of a model. However, identifying example membership is a challenging task even for models trained on small datasets (Carlini et al., 2022; Duan et al., 2023), and Maini et al. (2021) presented an impossibility result suggesting that as the size of the training set increases, the success of membership inference degrades to random chance. *Is testing the membership of individual sentences for LLMs trained for a single epoch on trillions of tokens of text data feasible?* In our work, we first demonstrate that previous claims of successful membership inference for individual text sequences in LLMs (Mattern et al., 2023; Shi et al., 2024) are overly optimistic. Our evaluation of the MIA methods for LLMs reveals a crucial confounder: they detect (temporal) distribution shifts rather than the membership of data points (as also concurrently observed by (Duan et al., 2024)). Specifically, we find that these MIAs infer whether an LLM was trained on a *concept* rather than an individual *sentence*. Even when the outputs of such MIAs (weakly) correlate with actual sentence membership, we find that they remain very brittle across sentences from different data distributions, and no single MIA succeeds across all. Based on our experiments, we conclude the discussion of MIAs with guidelines for future researchers to conduct robust experiments, highlighting the importance of using IID splits (between members and non-members), considering various data distributions, and evaluating false positives to mitigate confounding factors.

If membership inference attacks are so brittle, do content writers and private individuals have no recourse to claim that LLM providers unfairly trained on their data? As an alternative to membership inference, we advocate for a shift in focus towards dataset inference (Maini et al., 2021), which is a statistically grounded method to detect if a given *dataset* was in the training set of a model. We propose a new dataset inference method for LLMs that aims at detecting sets of text sequences by specific authors, thereby offering a more viable approach to dataset attribution than membership inference. Our method is presented in Figure 1. The motivation behind dataset inference stems from the observation that in the rapidly evolving discourse on copyright, individual data points have much less agency than sets of data points attributed to a particular creator; and the fact that more often than not, cases of unfair use emerge in scenarios when multiple such sequences or their clusters naturally occur. For instance, consider the Harry Potter series written by J.K. Rowling. Dataset inference tests whether a 'dataset' or a collection of paragraphs from her books was used for training a language model, rather than testing the membership of individual sentences alone. We also outline the specific framework required to operationalize dataset inference, including the necessary assumptions for the same.

We carry out our analysis of dataset inference using LLMs with known training and validation data. Specifically, we leverage the Pythia suite of models (Biderman et al., 2023) trained on the Pile dataset (Gao et al., 2020). This controlled experimental setup allows us to precisely analyze the model behavior on members and non-members when they occur IID (without any temporal shift) as the training and vali-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

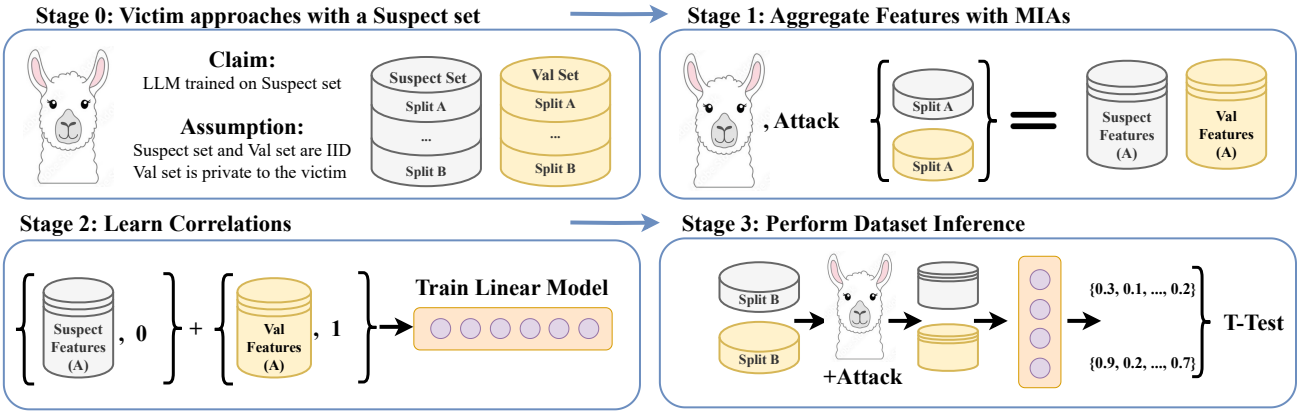

Figure 1. **LLM Dataset Inference.** ***Stage 0:*** *Victim approaches an LLM provider.* The victim's data consists of the suspect and validation (Val) sets. A victim claims that the suspect set of data points was potentially used to train the LLM. The validation set is private to the victim, such as unpublished data (e.g., drafts of articles, blog posts, or books) from the same distribution as the suspect set. Both sets are divided into non-overlapping splits (partitions) A and B. ***Stage 1:*** *Aggregate Features with MIAs.* The A splits from suspect and validation sets are passed through the LLM to obtain their features, which are scores generated from various MIAs for LLMs. ***Stage 2:*** *Learn Correlations (between features and their membership status).* We train a linear model using the extracted features to assign label 0 (denoting potential members of the LLM) to the suspect and label 1 (representing non-members) to the validation features. The goal is to identify useful MIAs. ***Stage 3:*** *Perform Dataset Inference.* We use the B splits of the suspect and validation sets, (i) perform MIAs on them for the suspect LLM to obtain features, (ii) then obtain an aggregated confidence score using the previously trained linear model, and (iii) apply a statistical T-Test on the obtained scores. For the suspect data points that are members, their confidence scores are significantly closer to 0 than for the non-members.

dation splits of PILE are publicly accessible. Across all subsets, dataset inference achieves p-values less than 0.1 in distinguishing between training and validation splits. At the same time, our method shows no false positives, with our statistical test producing p-values larger than 0.5 in all cases when comparing two subsets of validation data. To its practical merit, dataset inference requires only 1000 text sequences to detect whether a given suspect dataset was used to train an LLM.

2. Points of Discourse

Membership Inference for LLMs. In this work, we question the central foundations of research on membership inference in the context of LLMs trained on trillions of tokens of web data. Our findings indicate that current membership inference attacks for LLMs are as good as (and often worse than) random guessing. We demonstrate that past successes in MIAs are often due to specific experimental confounders rather than inherent vulnerabilities. We provide guidelines for future researchers to conduct robust experiments, emphasizing the use of IID splits, considering various data distributions, assessing false positives, and using multiple random seeds to avoid confounders.

Shift to LLM Dataset Inference. Historically, membership inference focused on whether an individual data point was part of a training dataset. Instead, we aggregate multiple data points from individual entities, forming what we now consider a dataset. In our work, we have not only put thought towards the scientific framework of dataset inference but also the ways it will operationalize in real-world settings, for instance, through our running example of a writer who suspects that their books were trained on. Our research demonstrates that LLM dataset inference is effective in minimizing false positives and detecting even minute differences between training and test splits of IID samples.

Limitations A central limitation to dataset inference is the assumptions under which it can be performed. More specifically, we require that the training and validation sets must be IID, and the validation set must be completely private to the victim. While this may appear elusive a priori, we outline concrete scenarios to show how these sets naturally occur. For instance, through multiple drafts of a book, until one gets finalized. The same applies to many artistic and creative uses of LLMs across language and vision today. In terms of data and model access, we assume that the victim or a trusted third party, such as law enforcement, is responsible for running the dataset inference so that there are no privacy-related concerns. This will require the necessary legal framework to be brought in place, or otherwise suspect adversaries may deny querying their model altogether.

References

Germini, https://gemini.google.com/. URL https://gemini.google.com/.

Openai, https://openai.com. URL https://openai.com/.

Getty images vs. stability ai: A landmark case in copyright and ai, 2023. URL https://www.bakerlaw.com/getty-images-v-stability-ai/.

The times sues openai and microsoft over a.i. use of copyrighted work https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html. 2023. URL https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Sarah silverman and authors sue openai and meta over copyright infringement. 2023. URL https://www.nytimes.com/2023/07/10/arts/sarah-silverman-lawsuit-openai-meta.html.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=hvdKKV2yt7T.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.719. URL https://aclanthology.org/2023.findings-acl.719.

Noorjahan Rahman and Eduardo Santacana. Beyond fair use: Legal risk evaluation for training llms on copyrighted text. 2023. URL https://genlaw.org/CameraReady/57.pdf.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, Los Alamitos, CA, USA, may 2017. IEEE Computer Society. doi: 10.1109/SP.2017.41. URL https://doi.ieeecomputersociety.org/10.1109/SP.2017.41.

Xiaodong Wu, Ran Duan, and Jianbing Ni. Unveiling security, privacy, and ethical concerns of chatgpt. *Journal of Information and Intelligence*, 2023.