Scaling Up Membership Inference: When and How Attacks Succeed on Large Language Models

Anonymous ACL submission

Abstract

Membership inference attacks (MIA) attempt to verify the membership of a given data sample in the training set for a model. MIA has become relevant in recent years, following the rapid development of large language models (LLM). Many are concerned about the usage of copyrighted materials for training them and call for methods for detecting such usage. However, recent research has largely concluded that current MIA methods do not work on LLMs. Even when they seem to work, it is usually because of the ill-designed experimental setup where other shortcut features enable "cheating." In this work, we argue that MIA still works on LLMs, but only when multiple documents are presented for testing. We construct new benchmarks that measure the MIA performances at a continuous scale of data samples, from sentences (n-grams) to a collection of documents (multiple chunks of tokens). To validate the efficacy of current MIA approaches at greater scales, we adapt a recent work on Dataset Inference (DI) for the task of binary membership detection that aggregates paragraph-level MIA features to enable document- and dataset-level MIA. This baseline achieves the first successful MIA on pre-trained and fine-tuned LLMs.

1 Introduction

Large language models (LLMs) are trained on vast datasets, which providers typically keep private. Data owners are concerned that their copyrighted data might be used in LLM training without explicit consent. Membership Inference Attacks (MIAs) attempt to determine if a specific data sample was used to train a model (Shokri et al., 2017). These methods are now being applied to LLMs to address the question of potential data misuse (Shi et al., 2024; Meeus et al., 2024a; Zhang et al., 2024b; Wang et al., 2024a; Xie et al., 2024).

The application of MIAs on LLMs has faced discouraging results so far. Initial claims of success

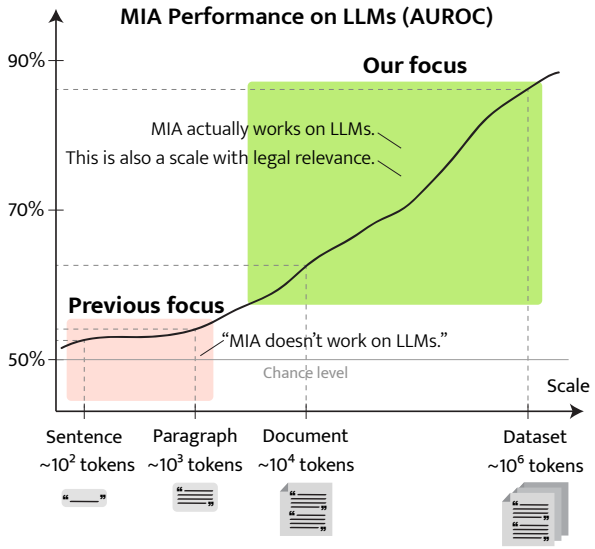

Figure 1: **Focusing on the Right Scale.** MIA has traditionally been considered ineffective for LLMs. However, we argue that MIA remains effective for LLMs when applied at a much larger scale, considering significantly longer token sequences. This *large-scale MIA* is also practically and legally relevant, as copyright is often determined at the document level.

(Shi et al., 2024; Meeus et al., 2024a; Carlini et al., 2021; Mattern et al., 2023) were later disproven by Duan et al. (2024); Das et al. (2024); Maini et al. (2024); Meeus et al. (2024b). Before their work, it was common practice to select true non-members from documents created after the LLM's *cut-off date*. They showed that this approach allowed MIA methods to exploit temporal cues, rather than identifying membership through the model's inherent response characteristics. Duan et al. (2024); Das et al. (2024); Maini et al. (2024) introduced a new evaluation method based on an independent, identically distributed (IID) split between true members and non-members. Since adopting this method, no further MIA studies on LLMs have shown performance significantly better than random. Reported membership detection has remained below 60%

AUROC, close to the random-chance level of 50% AUROC (Duan et al., 2024; Das et al., 2024; Maini et al., 2024; Xie et al., 2024; Zhang et al., 2024b).

We argue that MIA can still be effective on LLMs, provided it is applied to much longer token sequences than previously considered. Earlier MIA approaches often focused on short sequences of tokens, typically ranging from 128 to 256 tokens (Xie et al., 2024; Duan et al., 2024; Shi et al., 2024; Wang et al., 2024b). This use of n-grams faced criticism because of the significant overlap between members and non-members, making the MIA problem poorly defined (Duan et al., 2024). Some later studies suggested analysing entire documents instead of n-grams (Shi et al., 2024; Meeus et al., 2024b), but even then, performance remained near random (Meeus et al., 2024b).

In this work, we demonstrate that MIA approaches begin to show meaningful performance only when applied to much longer token sequences, such as 10K tokens. To show this, we introduce four scales of token sequences: sentences, paragraphs, documents, and datasets, as shown in Figure 1. We propose MIA evaluation protocols and benchmarks for the binary detection of training data samples given at four different scales. As a baseline, we adapt the Dataset Inference (Maini et al., 2024) method for the MIA task at multiple scales. As a result, our aggregation-based MIA demonstrates significant performance improvements for document sets, achieving AUROC scores of 80% or higher.

To explore additional scenarios, we investigate the performance of MIA on fine-tuning data. Our findings show that continual learning MIA is effective at the dataset level (AUROC > 88%), while CoT-based fine-tuning works at both the sentence and dataset levels.

Our contributions are summarized as:

1. We introduce a novel evaluation benchmark and protocol for MIA, covering multiple scales of token sequences.

2. We extend and adapt the first successful MIA Maini et al. (2024) to any data scale, allowing us to conduct a comprehensive analysis of MIA performance in LLMs.

3. We provide additional MIA benchmarks for various LLM fine-tuning scenarios, demonstrating that our method achieves even stronger performance in these contexts.

2 Background & Related Work

Membership inference attacks (MIA) aims to prove that a certain data sample belongs to the training set of a model. Yeom et al. (2018) hypothesize that members have a lower loss (perplexity for LLMs) than non-members and based on this propose to use the loss to infer membership. Carlini et al. (2021) build on top of it and propose to use the ratio of the perplexity and the zlib entropy. Shi et al. (2024) propose to compute the average log-likelihood of the tokens with the lowest probabilities based on the assumption that members would have higher probabilities than non-members.

Since most LLMs do not have training-test set splits, initial benchmarks and recent works use the knowledge cut-off date to define members and non-members (Shi et al., 2024; Meeus et al., 2024a). Works using these evaluation benchmarks show positive results (Shi et al., 2024; Meeus et al., 2024a; Zhang et al., 2024b; Wang et al., 2024a; Xie et al., 2024). However, Duan et al. (2024); Das et al. (2024); Maini et al. (2024) showed that the evaluation method based on cut-offs is flawed as these cut-offs introduce temporal biases that are easily captured by a bag of words. Therefore, they proposed to use LLMs trained on datasets that contain a random train-test split for members and non-members, like The Pile dataset (Gao et al., 2020). In this setup, however, MIA only achieves performance barely above random guessing.

Maini et al. (2024) show that aggregating MIA scores across multiple documents can yield successful dataset-level MIA. However, their work does not provide standard performance metrics like AUROC, leaving the precise effectiveness of their method unclear compared to standard paragraph-level MIA. In this study, we extend their evaluation by applying bootstrapping to compute dataset-level AUROC and adapt their method to any data scale, allowing us to conduct a comprehensive analysis of MIA performance in LLMs.

MIA performance across data scales in LLMs has not yet been explored despite its importance for copyright disputes. Current lawsuits against LLM developers primarily concern the use of entire documents in training sets (Pope, 2024). However, most MIA methods focus on short paragraphs of up to 128 tokens (Xie et al., 2024), 200 words (Duan et al., 2024), and 256 tokens (Shi et al., 2024; Wang et al., 2024b). Document-level MIA did not achieve significantly better results than random guessing

(Meeus et al., 2024a,b), highlighting the challenges in scaling MIA techniques to larger text units. This gap between the granularity of current MIA methods and the document and dataset-level focus of legal disputes underscores the need for more comprehensive research into MIA effectiveness across different scales of textual data in LLMs.

Scale	Definition	#Tokens
Sentence	Natural definition	43 on avg.
Paragraph	LLM context size	512, 1024, 2048, . . .
Document	Natural definition	14K on avg.
Dataset	Multiple docs	14K $\times$ #docs

Table 1: **MIA scales.** We define MIA at four scales.

3 MIA Evaluation

We introduce novel evaluation benchmarks for membership inference attacks (MIA) across various scales and multiple LLM training scenarios.

MIA Evaluation Overview. At a high level, following previous MIA work (Shi et al., 2024; Meeus et al., 2024a; Zhang et al., 2024b; Wang et al., 2024a; Xie et al., 2024), we frame the problem as a binary detection task: determining whether a given token sequence t was used during the training of a target large language model M. Specifically, MIA methods are expected to produce scores $(s^1, \cdots, s^n)$ for each token sequence in a set of instances $(t^1, \cdots, t^n)$. With knowledge of the actual membership status $(m^1, \cdots, m^n)$, indicating whether each sequence was part of the training set, we can evaluate detection performance using the area under the receiver-operating-characteristic curve (AUROC) across different thresholds.

However, creating a binary detection task for long token sequences is non-trivial. Depending on the LLM type, sequences of 10K tokens or more often exceed the context window of current models. This requires assessing membership across a set of sequences, demanding datasets with known members and non-members at the sequence set level.

In this work, we propose four MIA benchmarks on the PILE dataset (Gao et al., 2020), each targeting a different sequence length scale. We also extend previous MIA benchmarks, traditionally focused on detecting pre-training data, to three popular LLM training paradigms, including fine-tuning. In total, we introduce $4 \times 3 = 12$ MIA benchmarks, covering realistic application scenarios.

Data Scales. We define four scales of the MIA tasks: i) sentence, ii) paragraph, iii) document, and iv) dataset, as shown in Table 1.

Sentence-level MIA. We define a sentence as a natural sequence of words ending in a full stop. The average sentence in the Pile contains 43 tokens. This granularity is important because it is used to detect whether specific questions in benchmarks (e.g., question-answering tasks) are contaminated, ensuring fair model evaluation. Due to its short nature, it can be extremely challenging to perform MIA successfully. For instance, Duan et al. (2024) shows large overlaps between member and non-member sentences, which blurs the decision boundary. Additionally, sentence-level MIA can probe privacy leakage inferring membership of personally identifiable information (Kim et al., 2023).

Paragraph-level MIA. We define a paragraph as a sequence of tokens that fits within the context window of a large language model (LLM). Thus, the length of a "paragraph" depends on the specific LLM in use. In our experiments, we use paragraphs of up to 512, 1024, and 2048 tokens, aligning with the context window sizes of the target models. In practice, paragraph-level copyrighted materials are particularly relevant to user-generated content on social media platforms and online forums, where short-form content is most common.

Document-level MIA. We define a document in the conventional sense, such as a single arXiv paper. In The Pile dataset, the average document contains 14,222 tokens. Documents typically consist of multiple paragraphs, often exceeding the context window of many LLMs, such as Pythia, which has a 2048-token limit. MIA approaches must handle these long documents by splitting the token sequence into smaller chunks, or “paragraphs” (as defined earlier), that fit within the model’s context length and then aggregating the model’s responses across these chunks. Performing MIA at the document level is essential for addressing copyright and data ownership concerns, particularly for copyrighted materials like novels, news articles, research papers, and book chapters.

Dataset-level MIA. When MIA targets a collection of documents, we refer to this collection as a dataset. The attacker can choose the dataset’s size. For instance, with 100 documents, a dataset contains approximately 1.4 million tokens. Dataset-level MIA is crucial, as one may question whether

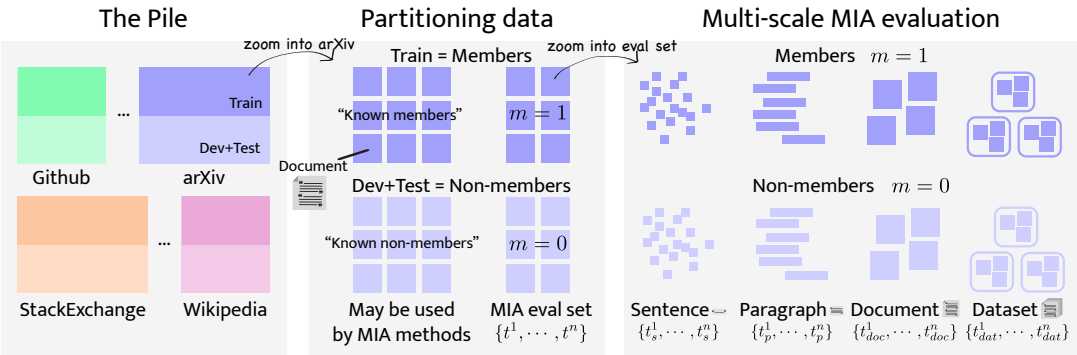

Figure 2: **Preparing MIA Evaluation Datasets.** Each source in The Pile is divided into Train, Dev, and Test splits, where the Train set is used for LLM training. For MIA, we designate the Train set as members and the Dev+Test sets as non-members. The MIA evaluation set $t^1, \cdots, t^n$ consists of datapoints for the binary detection task of predicting membership m. The benchmark makes some "known members" and "known non-members" available to the attackers; MIA methods may choose to use them or not. To support dataset-level MIA, we group documents in n different ways to create member and non-member datasets.

a collection of articles, such as those provided by an internet service provider, has been crawled and used in LLM training. Aggregating individual signals can amplify the detection of the dataset's usage. This is also relevant when examining contaminated benchmarks, as entire datasets may be unintentionally included in the training, leading to an overestimation of model performance.

MIA at scales beyond the document level is legally and practically significant, as copyright disputes often focus on individual articles. Moreover, as we will show in §5, MIA only achieves strong performance at these larger scales.

LLM Training Scenarios. Previous MIA approaches for LLMs have primarily focused on detecting pre-training data. While pre-training poses significant concerns regarding data scraping and exploitation by large tech companies with vast resources, other forms of training, such as fine-tuning, are much more common and affordable for entities worldwide. Although fine-tuning operates on a smaller scale, it may still involve copyrighted materials, and models fine-tuned in this way could be deployed in widely used applications. In such cases, copyright infringement could have serious consequences. Therefore, we consider MIA under three different LLM training scenarios.

Pretraining MIA requires the attacker to identify whether a given data instance was part of the pretraining corpus.

Continual-learning MIA addresses the challenge of keeping models up-to-date, as frequently retraining them from scratch is often infeasible due to high costs. Instead, continual learning offers a solution by incrementally training the model's checkpoint on new data. This approach is crucial for ensuring that LLMs remain current with evolving information (Wu et al., 2024).

Fine-tuning MIA seeks to determine whether specific samples were used during fine-tuning for particular end tasks or domains, such as training data for conversational bots, question-answering systems, and similar applications. Fine-tuning methods are becoming increasingly popular due to their lower costs and flexibility, making them an ideal solution for tasks with large amounts of relevant data (Meta, 2024).

Beyond addressing various practical scenarios, our exploration of non-pretraining MIA highlights the increased effectiveness of existing MIA methods. Since fine-tuning is conducted on smaller datasets, each data instance has a more significant influence on model behaviour, which enhances the likelihood of successful MIA (§5).

Preparing Data for Multiscale MIA Evaluation. For MIA evaluation, it is critical to have access to true members $m = 1$ and true non-members $m = 0$ for a precise evaluation of the membership detection performance. The Pile is one of the few datasets used for training LLMs that provides a clear separation of data used for training (the train split) and data *not* used for training the LLMs (the dev and test splits). We thus re-purpose the Pile dataset for MIA evaluation.

Figure 2 illustrates the procedure of turning the Pile dataset into a MIA evaluation benchmark. For each source in the Pile, we use a subset of the train (members) documents and dev+test (non-members)

documents for the MIA evaluation set $\{t^1, \cdots, t^n\}$. To enable multi-scale MIA evaluation, we adjust the granularity of each document instance t^i by either breaking it down into sentences or paragraphs or by aggregating the documents into sets of documents ("datasets" in our definition).

We encounter a problem at greater scales of MIA evaluation, especially for the dataset-level MIA. Many sources of the Pile contain only a small number of documents as non-members. For instance, arXiv includes only 4.8K non-member documents. When the number of documents in each dataset is above 100, there are at most 48 available non-members to evaluate the MIA performance. To address this problem, we use bootstrapping (Efron, 1992): instead of considering only non-overlapping datasets, allowing overlapping datasets leads to 5×10^{209} possible combinations of 100 documents. This also mirrors real-world scenarios, where overlapping documents are common across collections of documents, such as Wikipedia entries in question-answering datasets (Rogers et al., 2023). In our experiments, we generate 1K datasets for evaluation for members and non-members and conduct experiments varying the number of documents in each dataset between 10 and 500.

The "known members" and "known non-members" splits are the ones that are not used for MIA evaluation. They may be used for improving or adapting MIA to the dataset of interest. Our method in §4, for example, uses the known non-members subset to calibrate the detection.

4 Method

We extend the Dataset Inference (DI) methods introduced by Maini et al. (2024) to multiple MIA scales and adapt them for the binary detection task. While the original DI focused solely on determining whether a single dataset is a member or not, we adapt this methodology for finer granularity, such as document-level MIA. Additionally, we explain how to derive the detection scores necessary for evaluating AUROC performance.

We begin by explaining the Dataset Inference (DI) process. Given a token sequence t, we divide it into smaller paragraphs, $t_{p,1}, \ldots, t_{p,K}$, each small enough to fit within the context window of the large language model (LLM) M under investigation. Following the DI approach, we compute various membership inference attack (MIA) features for each token sequence, using existing sentence-level MIA methods. These include perplexity (Yeom et al., 2018), lowercase perplexity, zlib compression scores (Carlini et al., 2021), and Min-K statistics (with thresholds at 5%, 10%, 20%, ..., 60%; Zhang et al. 2024b). These MIA features, denoted by $A_1, \ldots, A_L$, ($L = 10$) are functions that take the likelihood output of M on a sequence as input and return a scalar representing the membership information for that sequence. Thus, for a given token sequence t, we produce a $K \times L$ array of MIA features, $A_l(M(t_{p,k}))$.

We employ a two-stage aggregation process to reduce the $K \times L$ array into a single statistic representing membership likelihood. In the first stage, we aggregate the L MIA features into a single score using a learned linear map $f : \mathbb{R}^L \to [0, 1]$. This linear map is trained on a dataset consisting of 1,000 known members and 1,000 known non-members (see §3), with the objective of predicting the membership status m of the token sequence t.

After aggregation, we obtain a set of K MIA scores: $\{f(t_{p,1}), \cdots, f(t_{p,K})\}$. These scores are treated as an unordered set, as the criteria for determining membership becomes largely independent of the ordering of documents within a dataset or paragraphs within a document. Following the DI approach, we perform statistical testing by comparing the set of K MIA scores against a baseline set of K scores from known non-members. Specifically, we use the Student's t-test for dataset-level MIA since we aggregate more than 30 samples and the Mann–Whitney U-test for document-level MIA. since we aggregate less than 30 samples from a non-normal distribution.

$$\text{t-score}(t) = \frac{\mu - \mu_{\text{n-m}}}{\sqrt{s^2 + s^2_{\text{n-m}}}} \quad \text{U-score}(t) = -\sum_{k=1}^{K} \text{rank}(f(t_{p,k}))$$

Here, μ and s represent the mean and sample standard deviation of the K MIA scores, with the subscript n-m indicating the scores from the non-members. The term $\text{rank}(f(t_{p,k}))$ refers to the rank of the MIA feature for the query sequence t in the combined set of $2K$ MIA features, which includes both the query sequence and the known non-members.

We conduct our main experiments on Pythia 2.8B and 6.9B (Biderman et al., 2023) with data from The Pile dataset (Gao et al., 2020). See Appendix A for details on the experimental setup.

5 Experiments

With our experiments, we aim to answer the following questions: i) How effective is the aggregation of MIA scores for larger textual units, such

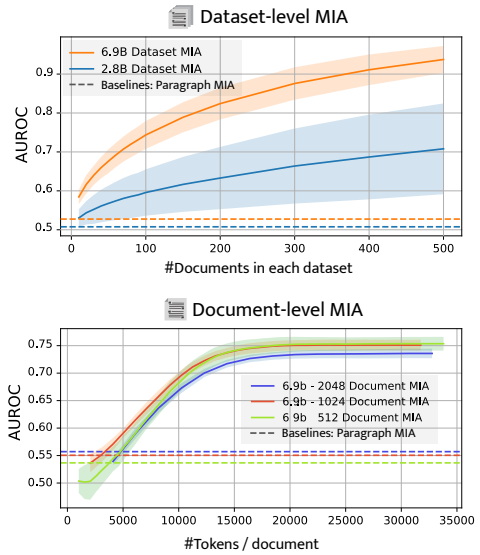

Figure 3: **Effect of aggregation.** We show MIA performances on arXiv at different levels of aggregation. Aggregation becomes more effective as we increase the number of aggregated instances.

Data Scale	ArXiv	HackerNews	Wiki
Sentence	0.501±0.003	0.500±0.003	0.507±0.004
Paragraph	0.528±0.004	0.511±0.015	0.523±0.013
Document	0.697±0.060	0.513±0.040	0.560±0.011
Dataset (500)	0.943±0.025	0.709±0.340	0.844±0.132

Table 2: **Multi-scale MIA results.** We show the AUROC scores for Pythia 6.9B at four scales. For dataset-level MIA, we use sets of 500 documents.

as datasets and documents? (section 5.1), ii) What are the requirements for a successful aggregation? (Section 5.2), iii) What is the nature of the compounding effect observed in MIA score aggregation? (section 5.3), iv) How much does the MIA aggregation benefit from fine-tuning scenarios (Section 5.4)?

5.1 Aggregating Text Subunits is Effective

Table 2 and Figure 3 demonstrate the effectiveness of aggregating multiple text units (e.g., sentences or paragraphs) to perform successful Membership Inference Attacks (MIA) at larger textual scales (e.g., documents or datasets). The figure illustrates a clear trend: MIA performance improves as we increase the number of aggregated text units.

The left plot focuses on dataset MIA for arXiv. It reveals a stark contrast between small and large datasets. While datasets with only dozens of documents yield AUROC scores barely above random chance, those with 500 or more documents achieve significantly higher AUROC. Notably, using the 6.9B model and despite employing only a few MIA methods in our ensemble, we attain a remarkable dataset MIA AUROC of 0.9 for arXiv.

The right plot shows the performance of document MIA on arXiv, where we observe our highest results. This exceptional performance can be attributed to two factors: i) the considerable length of the documents (averaging around 15K tokens), allowing for the aggregation of numerous MIA scores, and ii) a paragraph MIA AUROC exceeding 0.53. The combination of these factors yields an impressive document MIA AUROC of 0.75. To the best of our knowledge, this marks the first successful application of MIA to entire documents.

Known Partition Size and MIA Performance We observe no correlation between the size of the known partition and MIA performance. This finding is significant as it eliminates the need for large held-out collections, enhancing the method's practicality. Data providers often filter out portions of their datasets during their cleaning process; we believe some of this filtered data could be used as the known non-members partition. This approach makes our methodology applicable to existing datasets without additional data collection. We provide a plot that compares the MIA performance across different known partition sizes in Figure 8 in Appendix B.

5.2 Requirements for Successful Aggregations

Figure 4 illustrates common scenarios where the aggregation method fails to achieve MIA performance. The left plot demonstrates that dataset-MIA can be ineffective when the base AUROC is close to random chance. Even with 500 documents per dataset, the dataset-MIA for the 2.8B model barely achieves an AUROC of 0.55. The central plot presents a case where the base AUROC is sufficiently high, but the documents are too short (consisting of only 3 paragraphs of 2048 tokens each), providing insufficient paragraphs for effective aggregation. The right plot shows the opposite situation: while the documents are long enough (comparable to the successful arXiv case), the base AUROC is only 0.5, rendering the aggregation ineffective. These examples highlight the importance of both base performance and sufficient text units for successful MIA score aggregation. Lastly, aggregating sentences to conduct paragraph MIA becomes extremely challenging because sentence-level MIA remains unachievable

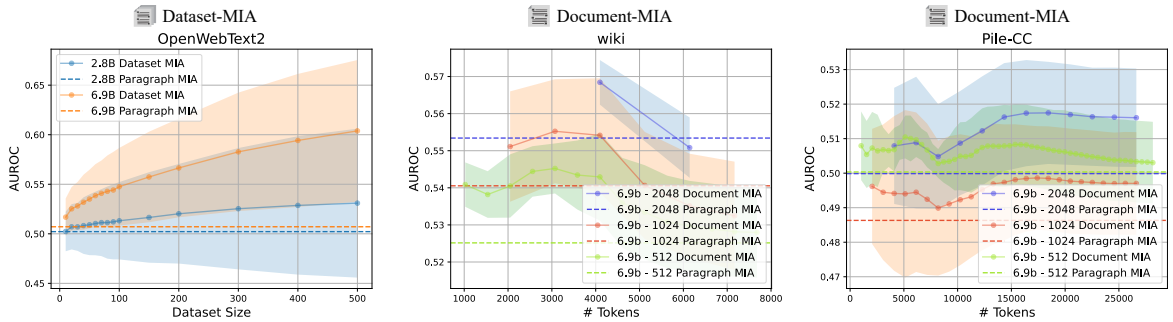

Figure 4: **Failutre modes for aggregation.** Aggregation may not work when the base performance is too low (left and right plots), or when the amount of information to aggregate is too short (centre plot).

and the works of Maini et al. (2021); Duan et al. (2024) suggest it might even be impossible. We provide the plots for all dataset, document, and paragraph aggregations on Appendix B.

5.3 Compounding Effect in MIA Score Aggregation

Our experiments reveal a powerful compounding effect in the aggregation of MIA scores from smaller to larger textual units. This relationship follows an approximately square root function, where small improvements in paragraph-level MIA performance lead to substantial gains at the dataset or document level. Figure 5 illustrates this relationship across all our experiments, encompassing various sources and hyperparameters.

The compounding effect is particularly striking when examining specific thresholds. For instance, a paragraph-level MIA with an AUROC of 0.51 can result in dataset-level MIA AUROCs ranging from 0.5 to 0.65. However, a modest improvement in paragraph-level performance to an AUROC of 0.53 can dramatically boost dataset-level AUROCs to between 0.6 and 0.9. This demonstrates the potential for significant gains in MIA effectiveness through relatively small improvements in paragraph level.

Lastly, our analysis also indicates that our aggregation method remains effective as long as the base AUROC exceeds 0.51 and establishes a threshold for when aggregation becomes a viable strategy.

5.4 MIA Aggregation Benefits from Fine-tuning Scenarios

5.4.1 Continual Learning Fine-tuning

In this section, we investigate the performance of MIA on LLMs that have undergone a continual learning process to adapt to specific domains. Figure 6 and Table 3 show the MIA performance on

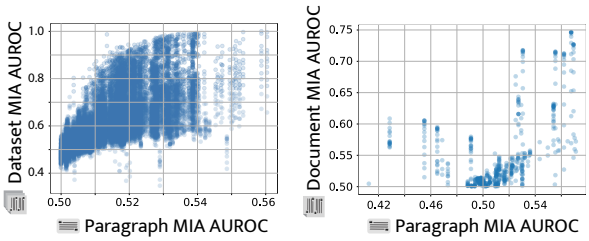

Figure 5: **Impact of paragraph-MIA performance** after aggregation to dataset- and document-level MIA.

Data Scale	Github	Wikipedia
Sentence	0.496± 0.006	0.498± 0.006
Paragraph	0.559± 0.017	0.577± 0.012
Document	0.579± 0.014	0.590± 0.015
Dataset (500)	0.885± 0.064	0.997± 0.007

Table 3: **Multi-scale MIA for continual learning.** We report AUROC for a continually-trained Pythia 2.8B.

Pythia 2.8B after it was further trained on the validation sets of Wikipedia and GitHub (independently) from the Pile dataset.

Our results reveal a significant increase in MIA effectiveness in this continual learning scenario. For Wikipedia, the dataset-level MIA performance achieves an AUROC of over 0.9, with datasets containing only 100 documents. This stands in stark contrast to the pretraining scenario, where the 2.8B model only reached an AUROC of 0.65 for datasets of 500 documents. This substantial improvement can be attributed to an increase in the base AUROC to 0.565, which amplifies the compounding effect (Section 5.3). However, sentence-level MIA remains ineffective, and paragraph-level MIA remains below 0.6.

We observe a similar pattern for GitHub, suggesting a consistent trend across different domains. These findings lead us to conclude that while

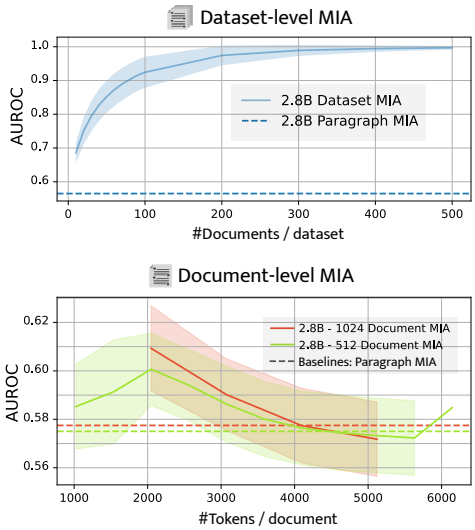

Figure 6: **Continual learning results** for Pythia 2.8B on Wikipedia. Dataset-level MIA shows near-perfect AUROC with sufficient documents, while document-level MIA shows a significant improvement.

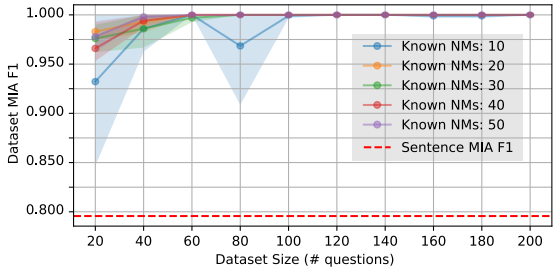

Figure 7: **CoT fine-tuning.** We show dataset-level MIA results on CoT-fine-tuned Phi 2. We evaluate on 100 datasets from 10k unique questions.

LLMs trained with continual learning remain robust against paragraph-level MIA, they become notably vulnerable to MIA when scores are aggregated across larger textual units.

5.4.2 End-Task Fine-tuning

We reuse the fine-tuned Phi-2 for reasoning from Puerto et al. (2024) for our fine-tuning experiments. They fine-tuned the model on multiple question-answering datasets so that the model responded with a chain of thoughts. This task allows us to know if MIA could be used to evaluate the contamination of fine-tuned models at the sentence level, for example, if the model used the evaluation questions for training.

Figure 7 presents the performance of dataset-MIA across varying dataset sizes and known partitions. Notably, even with just 10 members and 10 non-member questions, we achieved perfect classification for datasets containing 60 questions. When using a known partition of 50 members and 50 non-member questions, perfect classification was obtained for datasets with as few as 40 questions. In all cases, the F1 classification score exceeded 92.5, demonstrating the remarkable effectiveness of our approach for fine-tuned LLMs. This suggests that MIA could serve as strong evidence in legal cases to prove the use of data for fine-tuning LLMs, in contrast to the claims made by Zhang et al. (2024a).

We further explored scenarios where datasets contain a mix of both member and non-member questions to evaluate the robustness of our dataset-MIA method in the presence of noise. As shown in Figure 9 in Appendix B, with a 20% contamination rate, the method continued to classify the dataset as members, but at a contamination rate of 50%, the method correctly assigned membership status in 50% of cases. These results confirm the robustness of our approach to handling noisy datasets.

Lastly, we analysed the performance of MIAs at sentence level (i.e., individual questions) using five random seeds. The results revealed an average F1 score of 79.71 ± 1.16, an average AUROC of 81.30 ± 3.50, and a true positive rate at 1% false positive rate (TPR@1%FPR) of 3.71 ± 1.18. This shows that MIA can be very effective in fine-tuned LLMs.

6 Conclusion

In this paper, we have shown the importance of evaluating membership inference attacks (MIA) across different data scales, from sentences to datasets. Each text granularity represents a different and valuable use case that requires investigation. In contrast to prior works that suggest that MIA does not work for LLMs, we have empirically shown that MIA can work for documents and dataset levels with currently available attacks. We have further explored the performance of MIA across different training stages of an LLM and show that while small continual training remains robust towards sentence-level MIA, end-task fine-tuning is vulnerable, making MIA a suitable method to analyse test-set contamination.

In this work, we use simple MIA baselines to test the effectiveness of their aggregation and statistical testing. We believe the addition of more baselines would improve the overall results and leave that for future work.

Limitations

The lack of training-validation-test set splits for training data of LLMs limits the range of models to evaluate. In this work, we only focus on Pythia as in (Duan et al., 2024; Maini et al., 2024). Furthermore, we only employed 2.8B and 6.9B, the smallest models where MIA has been seen to perform a bit better than random chance. Using larger models could boost the reported results. Similarly, we only use three baselines as membership inference attacks. The use of more baselines could further improve our results.

Ethics and Broader Impact Statement

This work adheres to the ACL Code of Ethics. In particular, The Pile and Pythia we used have been shown by prior works to be safe for research purposes. They are not known to contain personal information or harmful content. We also fulfill with their licenses MIT for The Pile and Apache 2.0 for Pythia. Similarly, the model used in Section 5.4 also uses Apache 2.0 and is known to be safe for research purposes. Our method aims to provide data providers with tools to use their rights for the hypothetical cases where an LLM developer uses their data without consent.

Understanding when and how membership inference attacks (MIA) succeed on large language models (LLMs) can provide insights into developing strategies to train LLMs to be robust against these attacks. This poses the risk of making MIA outdated in the future, and therefore, data providers could be back in their current position without tools to use their rights.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Debeshee Das, Jie Zhang, and Florian Tramèr. 2024. Blind Baselines Beat Membership Inference Attacks for Foundation Models. *arXiv preprint*. ArXiv:2406.16201 [cs].

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

Bradley Efron. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 20750–20762. Curran Associates, Inc.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024. LLM Dataset Inference: Did you train on my dataset? *arXiv preprint*. ArXiv:2406.06443 [cs].

Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. 2021. Dataset inference: Ownership resolution in machine learning. In *International Conference on Learning Representations*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership Inference Attacks against Language Models via Neighbourhood Comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada. Association for Computational Linguistics.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. 2024a. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2369–2385.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. 2024b. Inherent Challenges of Post-Hoc Membership Inference for Large Language Models. *arXiv preprint*. ArXiv:2406.17975 [cs].

Meta. 2024. Community stories. Accessed: 2024-10-15.

Audrey Pope. 2024. Nyt v. openai: The times's about-face. *Harvard Law Review Blog*. Accessed: 2024-10-16.

Haritz Puerto, Tilek Chubakov, Xiaodan Zhu, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Fine-tuning with divergent chains of thought boosts reasoning through self-correction in language models. *Preprint*, arXiv:2407.03181.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10).

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Cheng Wang, Yiwei Wang, Bryan Hooi, Yujun Cai, Nanyun Peng, and Kai-Wei Chang. 2024a. Con-ReCall: Detecting Pre-training Data in LLMs via Contrastive Decoding. *arXiv preprint*. ArXiv:2409.03363 [cs].

Cheng Wang, Yiwei Wang, Bryan Hooi, Yujun Cai, Nanyun Peng, and Kai-Wei Chang. 2024b. Con-recall: Detecting pre-training data in llms via contrastive decoding. *ArXiv*, abs/2409.03363.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. 2024. Recall: Membership inference via relative conditional log-likelihoods. *Preprint*, arXiv:2406.15968.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. 2024a. Membership inference attacks cannot prove that a model was trained on your data. *arXiv preprint arXiv:2409.19798*.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. 2024b. Min-K%++: Improved Baseline for Detecting Pre-Training Data from Large Language Models. *arXiv preprint*. ArXiv:2404.02936 [cs].

A Experimental Settings

The Pile Data Splits. We conduct our experiments on 13 out of the 22 subsets available in The Pile dataset. We exclude the nine subsets that contain less than 3000 documents in the validation and test splits (BookCorpus2, Books3, Ubuntu IRC, EuroParl, Enron Emails, OpenSubtitles, PhilPapers, Youtube Subtitles, and Gutenberg) because our evaluation needs enough non-members documents. We did not evaluate document membership on the PubMed Abstracts subset because the documents contain less than 512 tokens, and thus, we cannot aggregate paragraphs. To obtain the sentences for sentence-level MIA, we used the first 2048 tokens of each document and tokenized them using the NLTK sentence tokenizer (Bird et al., 2009).

Implementation. We use the MIA implementations from Shi et al. (2024). We run all our experiments on 5 random seeds. We report the average AUROC with its standard deviation.

Continual Learning Fine-tuning. For the experiments on continual learning, we train Pythia 2.8B twice, once on the Wikipedia documents of the validation set and once on the Github documents of the validation set. We train the model with LoRa (Hu et al., 2021), using the PEFT library (Mangrulkar et al., 2022). Table 4 reports the generic and LoRa-specific hyperparameters. We apply MIA using the validation set of The Pile as the members and the test set as the non-members.

Hardware. All experiments are run using PyTorch 1.13, Tesla V100-PCIE-32GB GPUs, CUDA 12.1 and Ubuntu 20.04.4.

B Additional Figures

C Additional Tables

Hyperparameter	Value
Batch size	8
Epochs	4
Block size	1024
Optimizer	Paged 32-bit AdamW
Learning rate	2×10^{-4}
Gradient clipping norm	0.3
Weight decay	0.001
Quantization	4 bits
LoRa α	16
LoRa rank r	64
LoRa dropout	0.1

Table 4: Hyperparameters used to fine-tune the continual learning models.

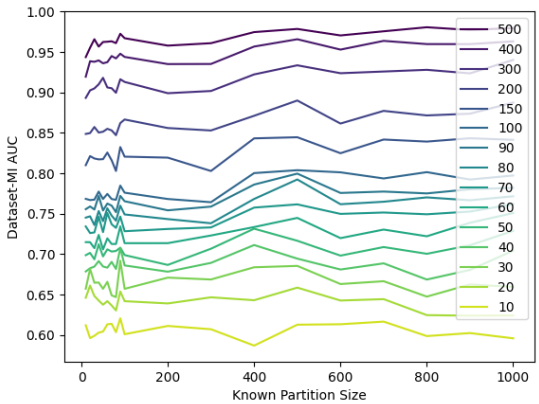

Figure 8: Increasing the size of the known partition does not increase Dataset-MI AUROC for any dataset size.

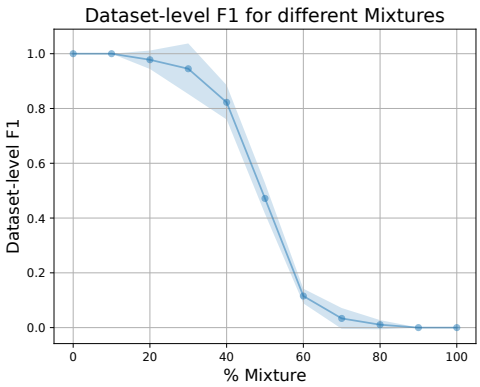

Figure 9: Dataset MI on CoT-based fine-tuned Phi 2 where the datasets contain a % of mix labels.

ThePile Subset	Sentence	Paragraph			Document	Dataset (500 Docs)
		512 Tokens	1024 Tokens	2048 Tokens		
ArXiv	43.3 ± 43.0	497.7 ± 73.3	984.4 ± 177.0	1948.5 ± 403.2	14221.9 ± 13893.0	7102765.9 ± 275483.6
DM Maths	23.0 ± 18.8	512.0 ± 0.0	1024.0 ± 0.0	2048.0 ± 0.0	3693.9 ± 652.9	1846939.7 ± 12665.8
GitHub	127.1 ± 275.9	372.1 ± 178.0	590.0 ± 385.8	857.0 ± 747.0	1765.9 ± 3652.0	880755.8 ± 81174.1
HackerNews	30.8 ± 27.4	333.6 ± 182.6	527.8 ± 399.0	793.9 ± 778.6	1574.6 ± 7769.4	784320.5 ± 142487.2
OpenWebText2	31.9 ± 37.4	392.6 ± 166.0	595.0 ± 356.1	754.7 ± 608.3	922.7 ± 1403.4	461230.0 ± 29294.9
Pile-CC	28.3 ± 24.6	363.5 ± 170.3	536.1 ± 359.3	688.9 ± 622.4	1011.5 ± 2770.4	503924.8 ± 57360.4
PubMed Central	38.9 ± 38.0	466.7 ± 135.1	913.0 ± 294.5	1763.0 ± 642.7	7124.0 ± 6825.5	3560020.9 ± 145890.2
StackExchange	53.0 ± 109.3	393.8 ± 132.0	519.8 ± 289.3	586.5 ± 444.6	627.2 ± 685.5	314027.1 ± 15372.1
Wikipedia	34.2 ± 44.3	323.3 ± 176.8	452.5 ± 348.4	564.5 ± 577.9	726.6 ± 1514.2	363979.1 ± 33328.6

Table 5: **Number of tokens on the data scale**. Average and standard deviation of the number of tokens computed for several text granularities: sentence (sentences longer than 25 characters in the first 2048 tokens), paragraph (the first chunk of 512, 1024 and 2048 tokens), full document, dataset (a set of 100 documents).

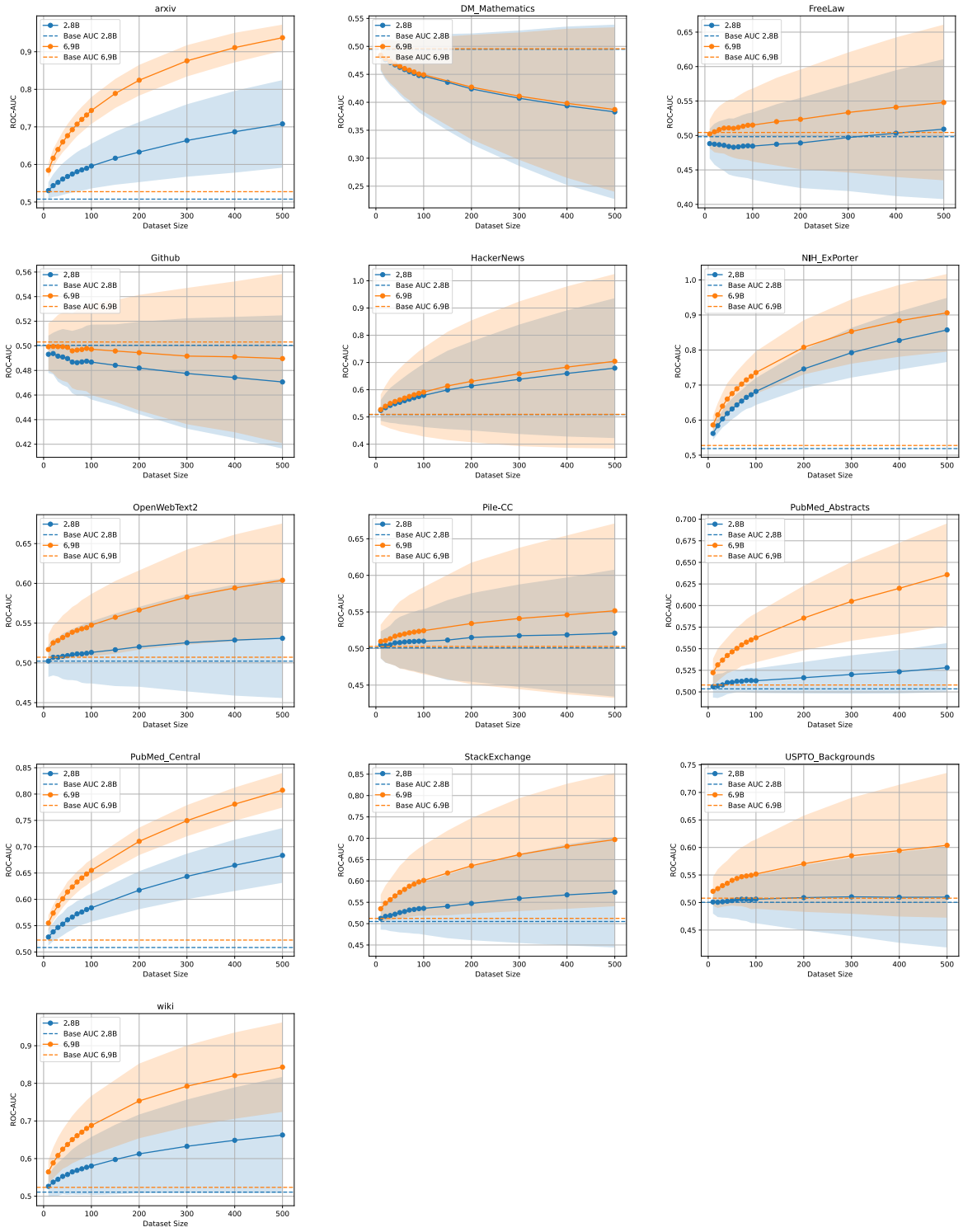

Figure 10: Dataset MIA on pretrained Pythia 2.8B and 6.9B

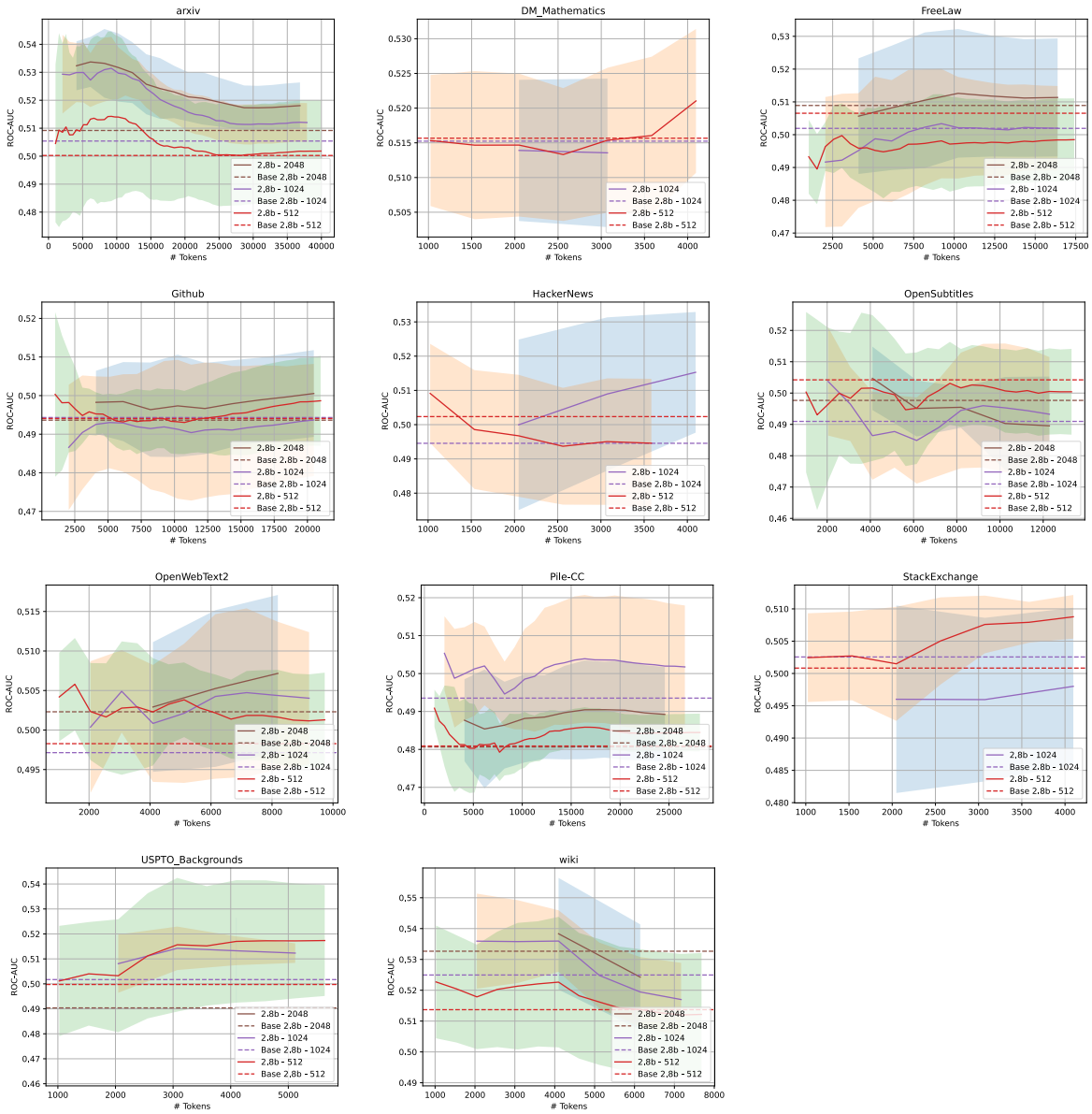

Figure 11: Document MIA on pretrained Pythia 2.8B

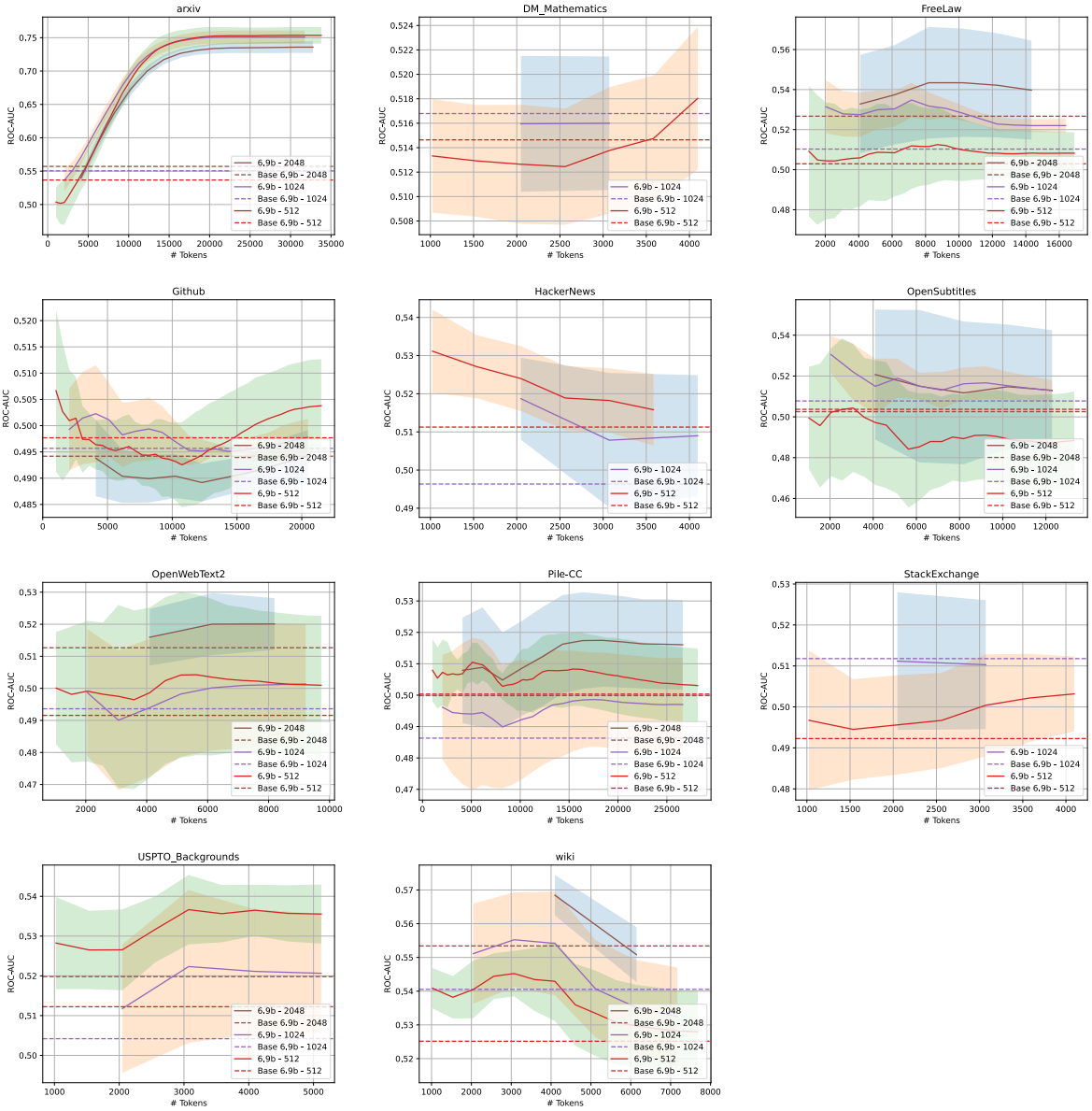

Figure 12: Document MIA on pretrained Pythia 6.9B

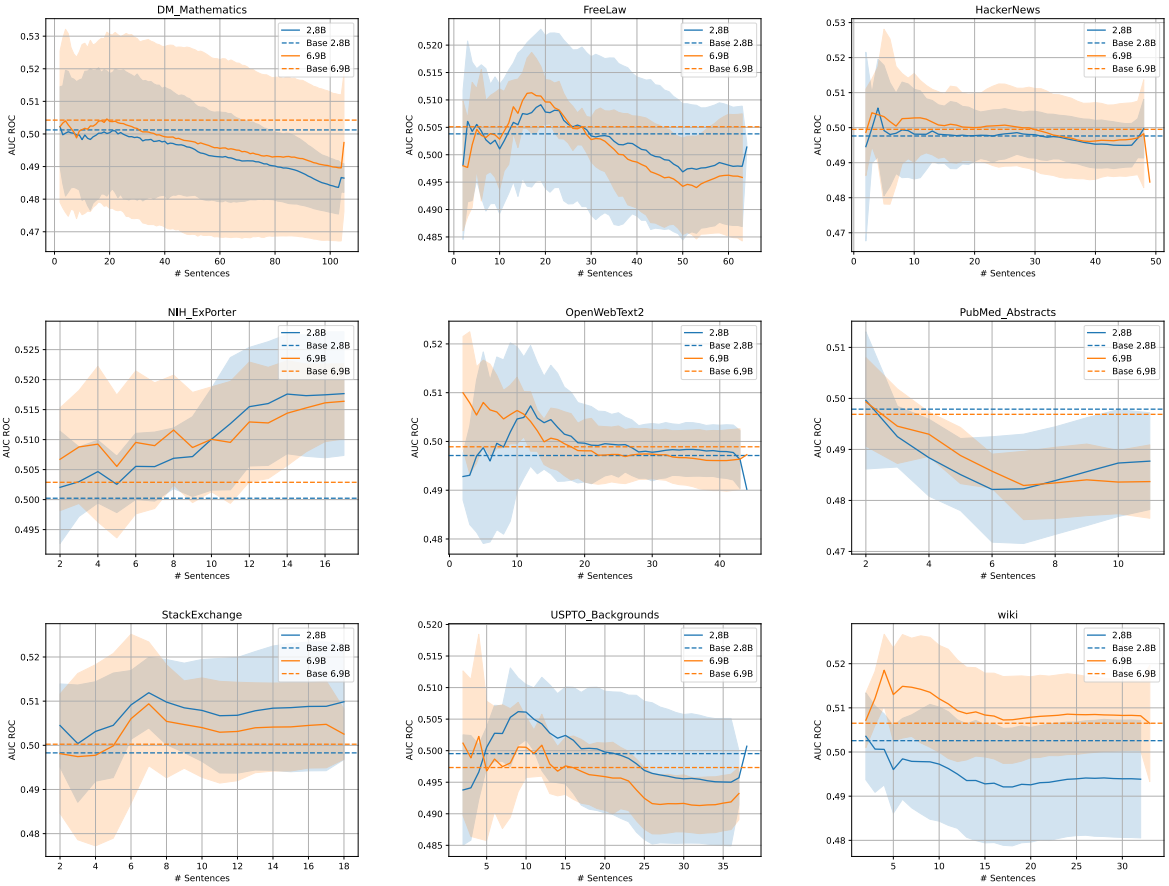

Figure 13: Sent MIA on pretrained Pythia 2.8B and 6.9B

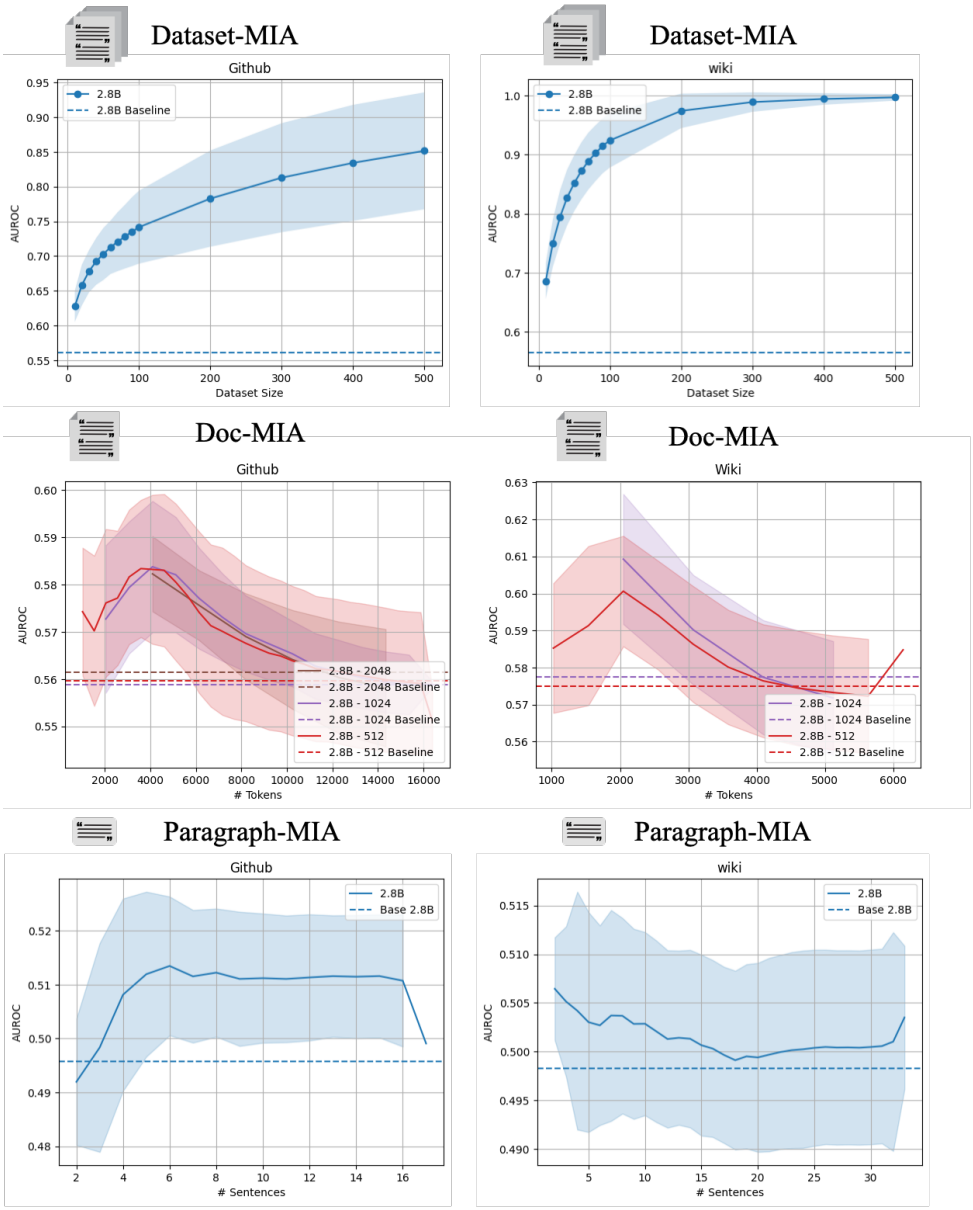

Figure 14: MIA across all data scales Pythia 2.8B trained with continual learning.