Evaluating Privacy in Small Language Models: A Membership Inference Attack Benchmark

Roya Arkhmammadova,
Hosein Madadi Tamar

Problem Statement

There is a clear lack of research and experimental data on the vulnerability of smaller LLMs to MIAs. While larger LLMs have been extensively studied, smaller models, which are increasingly being deployed due to their reduced computational requirements and suitability for resource-constrained environments, have received less attention. This lack of research creates a blind spot in security, as these smaller models exhibit different vulnerabilities or sensitivities to MIAs.

We release all the codes for running the benchmark with instructions on how to run it at https://github.com/RoyArkh/COMP430_Project/tree/main

Motivation and Rationale:

Growing Deployment of Smaller LLMs: Smaller LLMs are becoming more prevalent in various applications due to their efficiency and adaptability. This increased deployment necessitates a thorough understanding of their security properties, including their resilience against MIAs.

Lack of Benchmarking: The absence of a standardized benchmarking framework for evaluating MIA vulnerability in smaller LLMs makes it difficult to compare different models and attack methods.

Project Goals and Contributions:

1. *Benchmarking Smaller LLMs*: Conducting extensive experiments on smaller LLMs (Pythia, MobileLLM, GPT-Neo, with varying sizes) using different datasets (Wikitext, AG News, XSum) and a range of MIA methods (Loss Attack, Neighborhood Attack, Reference Attack, Lowercase Attack, MinK, MinK++, Zlib).
2. *Providing benchmarking results*: Developing a reusable benchmarking system that can be used by other researchers to evaluate the MIA vulnerability of their own models. This framework allows for standardized and comparable evaluations.
3. *Identifying Vulnerabilities*: Identifying specific vulnerabilities and patterns in the susceptibility of smaller LLMs to different MIA methods. This information can be used to develop mitigation strategies and improve the security of these models.
4. *Filling the Research Gap*: Contributing to the existing body of knowledge on MIAs by providing empirical evidence and insights into the vulnerability of smaller LLMs.

MIA

In the context of our project, a Membership Inference Attack (MIA) is a type of attack that tries to determine whether a specific data point (a piece of text in our case) was part of the training dataset of a machine learning model, specifically a Large Language Model (LLM).

Here's a breakdown based on our project's focus:

What it is:

Imagine you have trained an LLM (like Pythia, MobileLLM, or GPT-Neo) on a collection of text data (like Wikitext, AG News, or XSum). A MIA attempts to answer the question: "Was this particular text *used* to train the model, or not?"

Why it's a problem:

Privacy Violation: If an attacker can successfully infer membership, it reveals private information about the training data. For instance, if the training data contained sensitive documents or personal communications, a successful MIA could expose these details.

Model Vulnerability: A high success rate of MIA indicates that the model might be overfitting or memorizing parts of the training data, which is undesirable. It can also point to biases in the data.

Security Risk: In real-world applications, such as using LLMs for generating medical reports or financial analyses, revealing training data membership could have severe consequences.

How we made it work:

Our project uses several different MIA methods, which exploit different properties of the model's output. Side note: in the project proposal we only mentioned Neighbourhood, Loss, and Mink. However, we ended up implementing more attacks, namely: neighbourhood, loss, zlib, mink, mink++, lowercase.

- Loss Attack: This attack leverages the fact that a model typically has lower loss (or higher probability) on data points it has seen during training compared to unseen data. The attacker measures the loss of the target text on the model. A lower loss suggests that the text was likely in the training set.
    - *Example:* If we train MobileLLM on Wikitext and then calculate the loss of a sentence *from* Wikitext on the trained model, the loss will likely be lower than the loss of a completely random sentence.
- Neighborhood Attack: This attack explores the "neighborhood" of a given text in the model's embedding space. The idea is that if a text was in the training set, its neighbors (similar texts) might also be closer in the embedding space. We generate "neighbors" of the target text (using the generate_neighbours function in our code) and analyze their log-probabilities.
    - *Example:* If a sentence from AG News was used to train Pythia, then slight variations of that sentence (its neighbors) might also have relatively high probabilities according to Pythia.
- Other Attacks: Lowercase, MinK, MinK++, and Zlib attacks use different heuristics and statistics based on the model's output or compression ratios to infer membership.

Example in our project:

Let's say we train GPT-Neo on a subset of XSum. We then want to test if a specific summary, *S*, was part of the training set. We use the Loss Attack:

1. We calculate the loss of *S* on the trained GPT-Neo model.
2. We also calculate the loss of several summaries *not* in the training set.
3. If the loss of *S* is significantly lower than the average loss of the unseen summaries, we infer that *S* was likely part of the training set.

Datasets

In our project on Membership Inference Attacks (MIAs) against smaller Large Language Models (LLMs), we utilize three distinct text datasets: Wikitext, AG News, and XSum. Each dataset has unique characteristics that influence the training and evaluation of LLMs and, consequently, the effectiveness of MIAs.

Original Datasets:

- Wikitext: This dataset consists of long-term language modeling benchmark data extracted from Wikipedia articles. It's known for its large size and diverse vocabulary, making it suitable for training general-purpose language models. It contains well-formed text with a wide range of topics.

- AG News: This dataset is a collection of news articles from various sources, categorized into four classes: World, Sports, Business, and Sci/Tech. It's a smaller dataset compared to Wikitext and is specifically designed for text classification tasks. The text is more concise and focused on factual reporting.

- XSum: This dataset focuses on extreme summarization. It contains news articles paired with very short, one-sentence summaries. This dataset is challenging for LLMs due to the high level of abstraction required to generate accurate summaries.

Benchmark Datasets (Our Contribution):

To conduct our MIA experiments in a controlled and consistent manner, we created "benchmark" versions of each of the original datasets. These benchmark datasets have the following properties:

- Fixed Size: Each benchmark dataset contains exactly 700 samples.
- Balanced Membership: Each benchmark dataset is carefully constructed to have a balanced membership:
    - 350 samples are *members* (i.e., they are taken from the *training split* of the original dataset).

  - 350 samples are *non-members* (i.e., they are taken from the *test split* of the original dataset).
- Purpose: These benchmark datasets are used *exclusively* for the attack phase of our experiments. This ensures that we are evaluating the effectiveness of the MIAs on a consistent and comparable set of data, eliminating variations due to different dataset sizes or membership imbalances.

Example:

Let's take Wikitext as an example. The original Wikitext dataset is very large. To create our Wikitext benchmark dataset:

1. We select 350 samples from the *training split* of the original Wikitext dataset. These become our *member* samples.
2. We select 350 *different* samples from the *test split* of the original Wikitext dataset. These become our *non-member* samples.
3. We combine these 700 samples (350 members + 350 non-members) to create our Wikitext benchmark dataset.

We repeat this process for AG News and XSum, creating a benchmark dataset of 700 samples for each.

Why Benchmark Datasets are Important for MIA Evaluation:

- Controlled Experiments: Using fixed-size, balanced datasets allows us to directly compare the performance of different MIA methods and the vulnerability of different LLMs without being confounded by dataset size or membership bias.
- Fair Comparison: It provides a level playing field for evaluating different attack strategies.
- Reproducibility: Other researchers can easily reproduce our experiments using the same benchmark datasets.

Vulnerability

By using these carefully constructed benchmark datasets, our project ensures a robust and reliable evaluation of MIAs against smaller LLMs. We can confidently assess the vulnerabilities of these models and provide valuable insights for improving their security.

Observed Vulnerabilities and Trends:

Our experiments, conducted on smaller LLMs (Pythia, MobileLLM, and GPT-Neo) with varying sizes and fine-tuned on diverse datasets (Wikitext, AG News, and XSum), revealed varying degrees of vulnerability to Membership Inference Attacks (MIAs). While the models did not demonstrate complete resistance to MIAs, the effectiveness of the attacks varied significantly depending on the specific attack method employed.

Key Observations:

Varying AUROC Scores: The Area Under the Receiver Operating Characteristic curve (AUROC) scores, used to measure the success of the MIAs, exhibited substantial variation across different attack methods. This indicates that some attacks are more effective than others in inferring membership for these smaller LLMs.

Generally High AUROC: In many cases, the observed AUROC scores were relatively high. This suggests that the attacks could often correctly infer whether a given data point was a member of the training dataset. This finding highlights a potential privacy risk associated with these models, even at smaller scales.

Method-Specific Effectiveness: The effectiveness of each attack method varied. For instance, the Loss Attack might have performed better on certain models or datasets, while the Neighborhood Attack or Reference Attack might have been more effective in other scenarios. This emphasizes the importance of evaluating models against a diverse set of MIA methods to obtain a comprehensive understanding of their vulnerabilities.

Impact of Model Size and Fine-tuning Data: Our experiments also explored the impact of model size and the specific dataset used for fine-tuning. We observed that these factors can influence the vulnerability of the models to MIAs. For example, larger models or models fine-tuned on certain datasets might exhibit different levels of susceptibility.

Visualization with AUROC Plots: We will present our results using AUROC plots. These plots will visually demonstrate the performance of each attack method across different models and datasets, allowing for easy comparison and identification of trends.

Implications:

These findings have important implications for the security and privacy of smaller LLMs. The fact that these models are vulnerable to MIAs, even to varying degrees, underscores the need for developing robust defenses and mitigation strategies. Our benchmarking framework and experimental results provide valuable insights for researchers and practitioners working on improving the privacy and security of these models.

Model size impact on the experiments and results

Our analysis of varying model sizes within each LLM family (Pythia, MobileLLM, and GPT-Neo) revealed a trend of increasing vulnerability to Membership Inference Attacks (MIAs) with larger parameter counts. However, notable differences emerged between model families. MobileLLM demonstrated the highest robustness against MIAs, likely due to its recent development by Meta and its distinct tokenization approach. Conversely, the Pythia family exhibited the greatest susceptibility to these attacks, emerging as the weakest in our experiments. The graphs for AUROC curves are shown in Results 1 section

Impact of Datasets on the results and experiments

Observed Vulnerability Ranking:

- **Wikitext (Most Vulnerable):** Models trained on Wikitext show the highest vulnerability to MIAs.
- **AG News (Moderately Vulnerable):** Models trained on AG News exhibit moderate vulnerability.
- **XSum (Least Vulnerable):** Models trained on XSum are the least susceptible to MIAs among the three datasets.

Explanation and Logic:

Our hypothesis connects this vulnerability ranking to the characteristics of the text in each dataset, specifically focusing on token repetition and variability:

- **Wikitext's High Repetitiveness:** Wikitext, being extracted from Wikipedia, contains a high degree of redundancy and repetition. Articles on similar topics often share similar phrases, sentence structures, and factual information. This repetition can lead to the model "memorizing" specific sequences of tokens or even entire sentences. When a MIA is performed, the model's higher confidence (lower loss) on these memorized sequences makes it easier to infer membership.
    - *Example:* Many Wikipedia articles about historical figures might contain similar phrases like "born in [year]", "died in [year]", or "known for [achievement]". If a model memorizes these common phrases, it will be highly confident when encountering them again, making it vulnerable to loss-based MIAs.
- **AG News' Moderate Variability:** AG News, consisting of news articles, has less repetition than Wikitext. While news articles within the same category might share some vocabulary, the overall sentence structure and content are more varied. This lower repetitiveness makes it slightly harder for the model to memorize specific training examples, reducing its vulnerability to MIAs compared to Wikitext.

- **XSum's High Variability:** XSum, with its focus on extreme summarization, presents the highest variability. Each summary is a concise, one-sentence distillation of a longer article. This process of summarization inherently removes redundancy and focuses on the most salient information. As a result, the summaries in XSum are highly diverse and less likely to be repeated verbatim in the training set. This high variability makes it difficult for the model to memorize training examples, making it the least vulnerable to

MIAs among the datasets.

- *Example:* Even if two news articles are about the same event, their XSum summaries will likely be phrased differently, focusing on different aspects of the story. This high variability makes it harder for the model to overfit and thus reduces its vulnerability to MIAs.

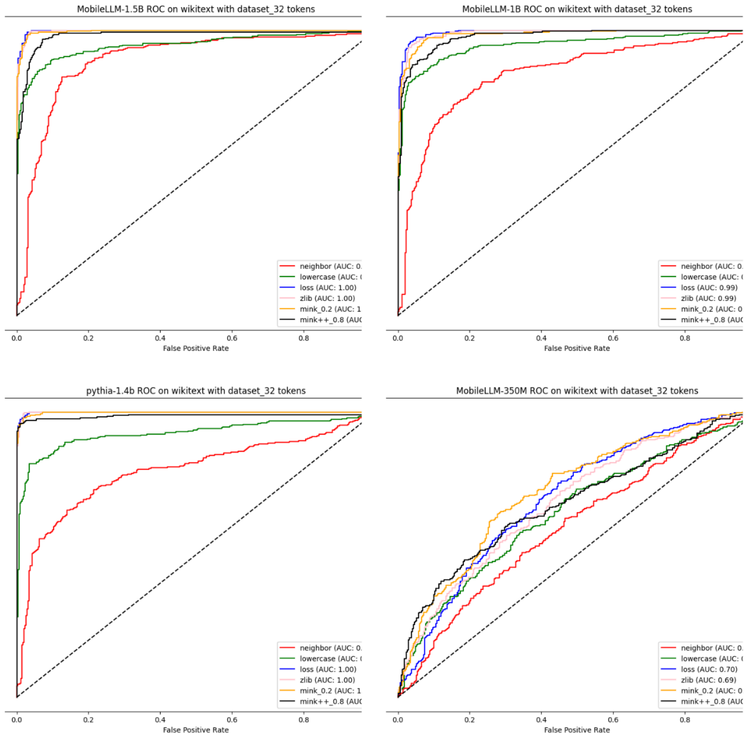
Connecting to MIA Mechanisms:

Loss-based Attacks: These attacks are particularly effective when the model memorizes training examples, as the loss on these examples will be significantly lower. Wikitext's high repetitiveness makes it particularly vulnerable to these attacks.

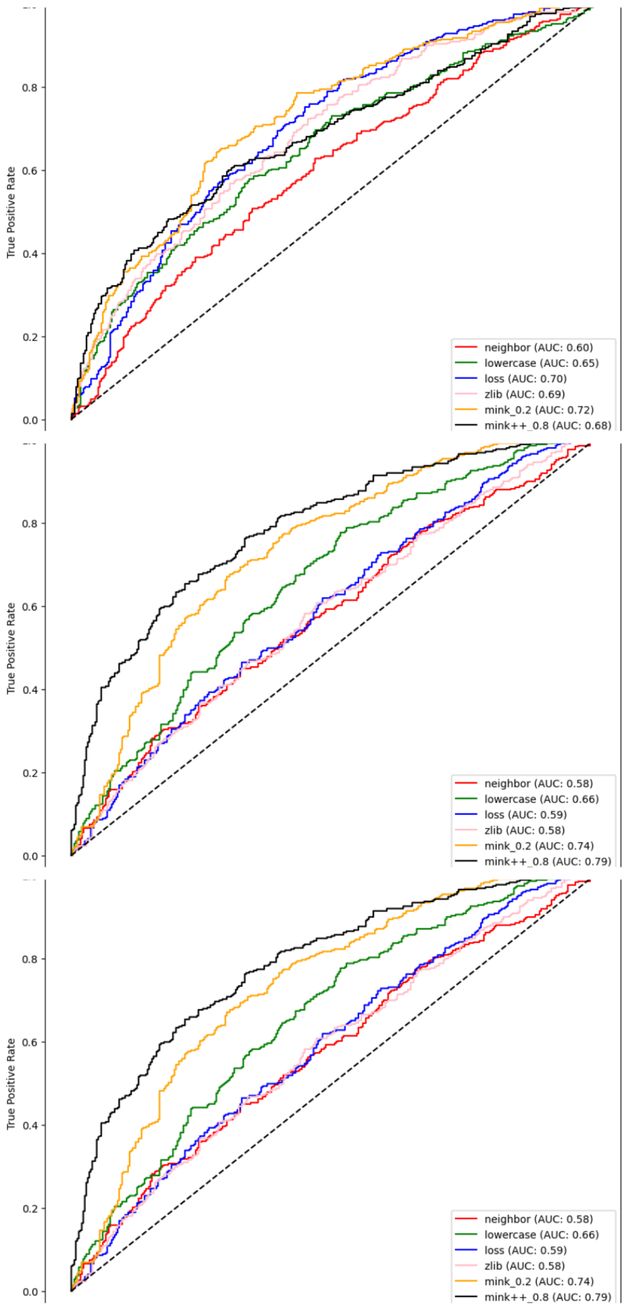
Other Attacks: Attacks that rely on the model's confidence or probability distributions are also affected by memorization. When a model memorizes a training example, its confidence in that example will be higher, making it easier for the attacker to distinguish members from non-members.

In summary: The observed vulnerability ranking (Wikitext > AG News > XSum) can be attributed to the degree of token repetition and variability within each dataset. Higher repetition (like in Wikitext) leads to increased memorization and thus higher vulnerability to MIAs, while higher variability (like in XSum) makes it harder for the model to memorize training examples and thus reduces vulnerability. This analysis highlights the importance of data characteristics in the context of model security and privacy. For more results please check the following link: https://drive.google.com/drive/folders/1IpGWMMVzXVKz3rsvwD2Li7M1jx2SUouM?usp=drive_link

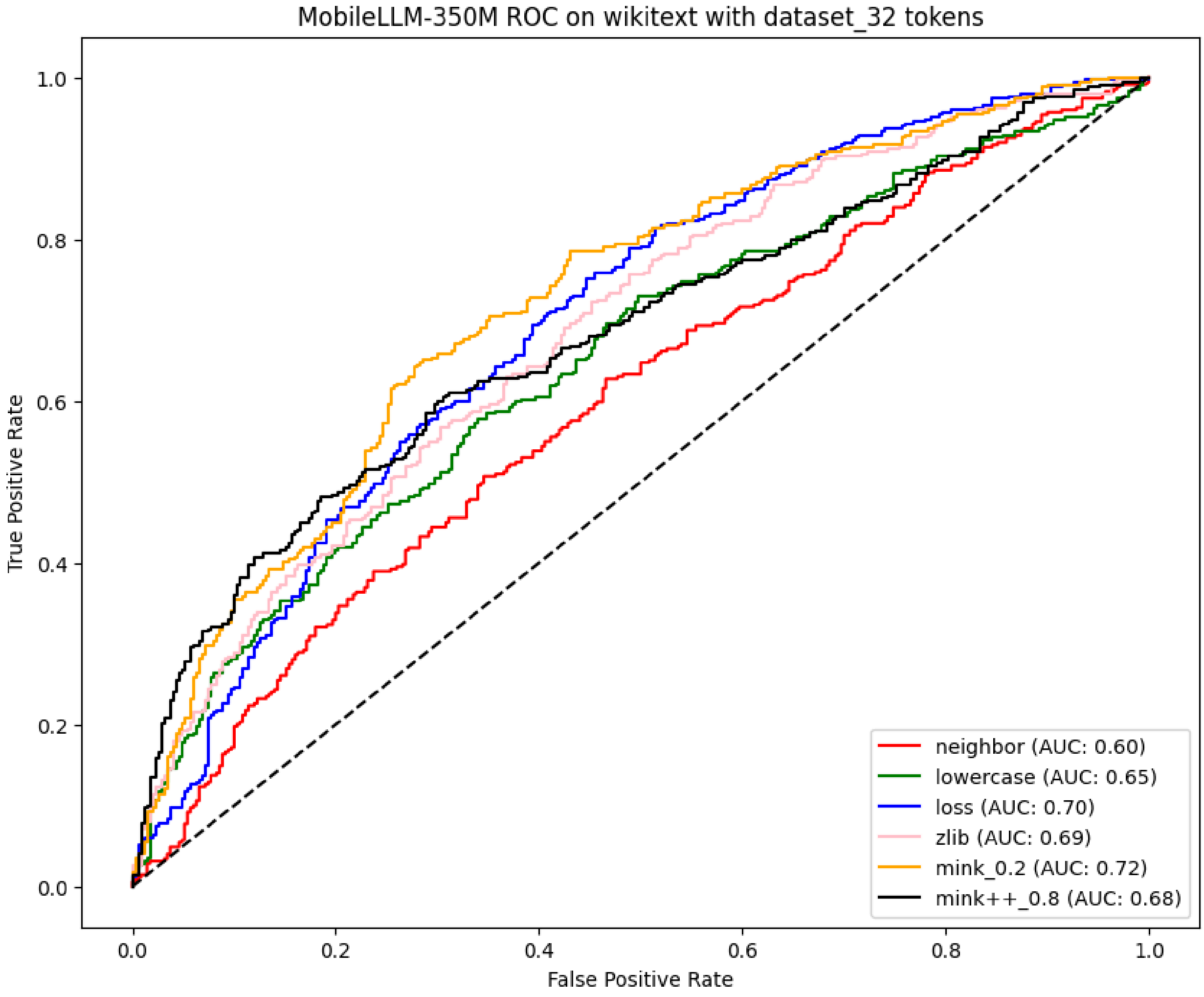
Results 1: for Wikitext dataset and different sizes of models

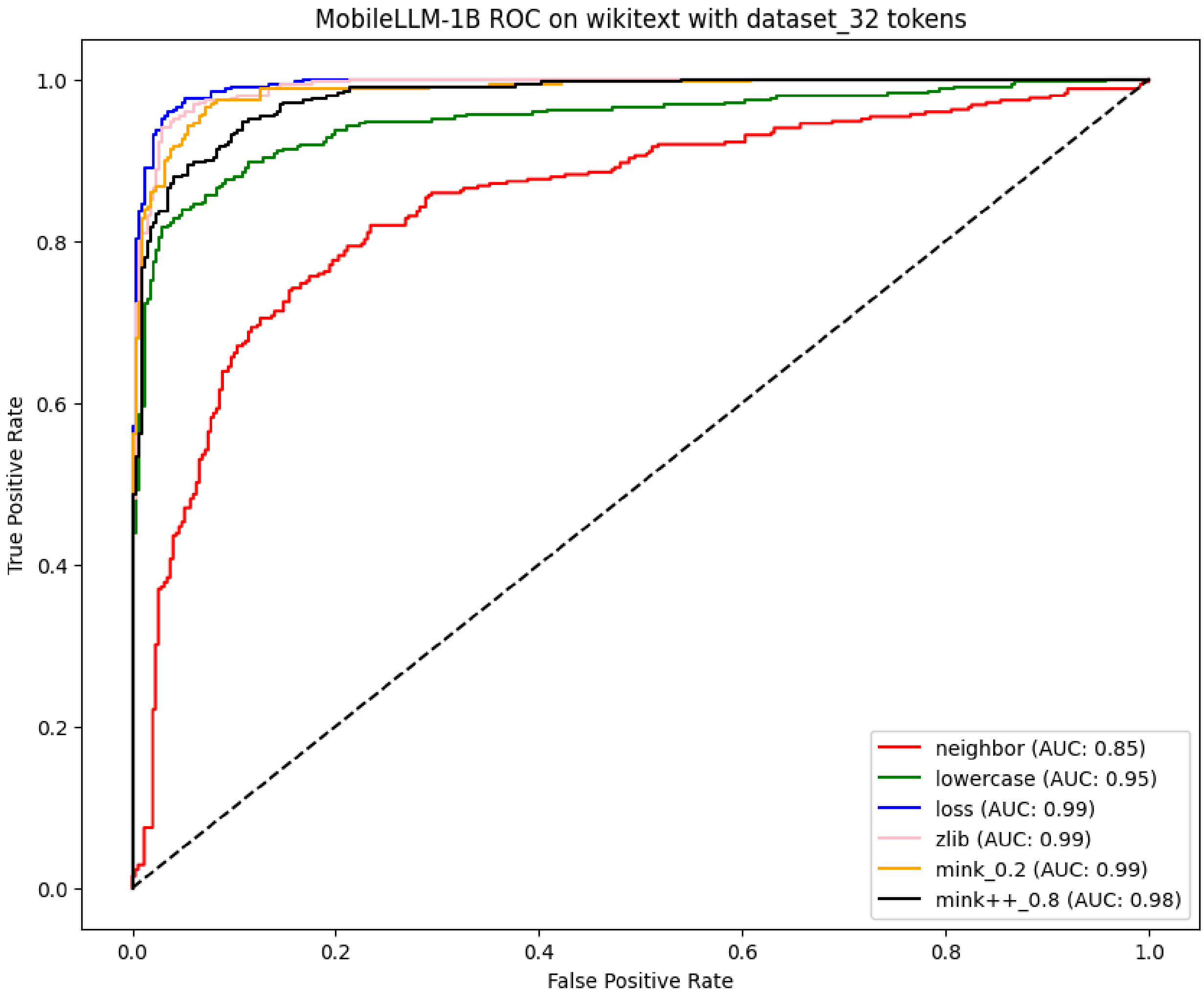


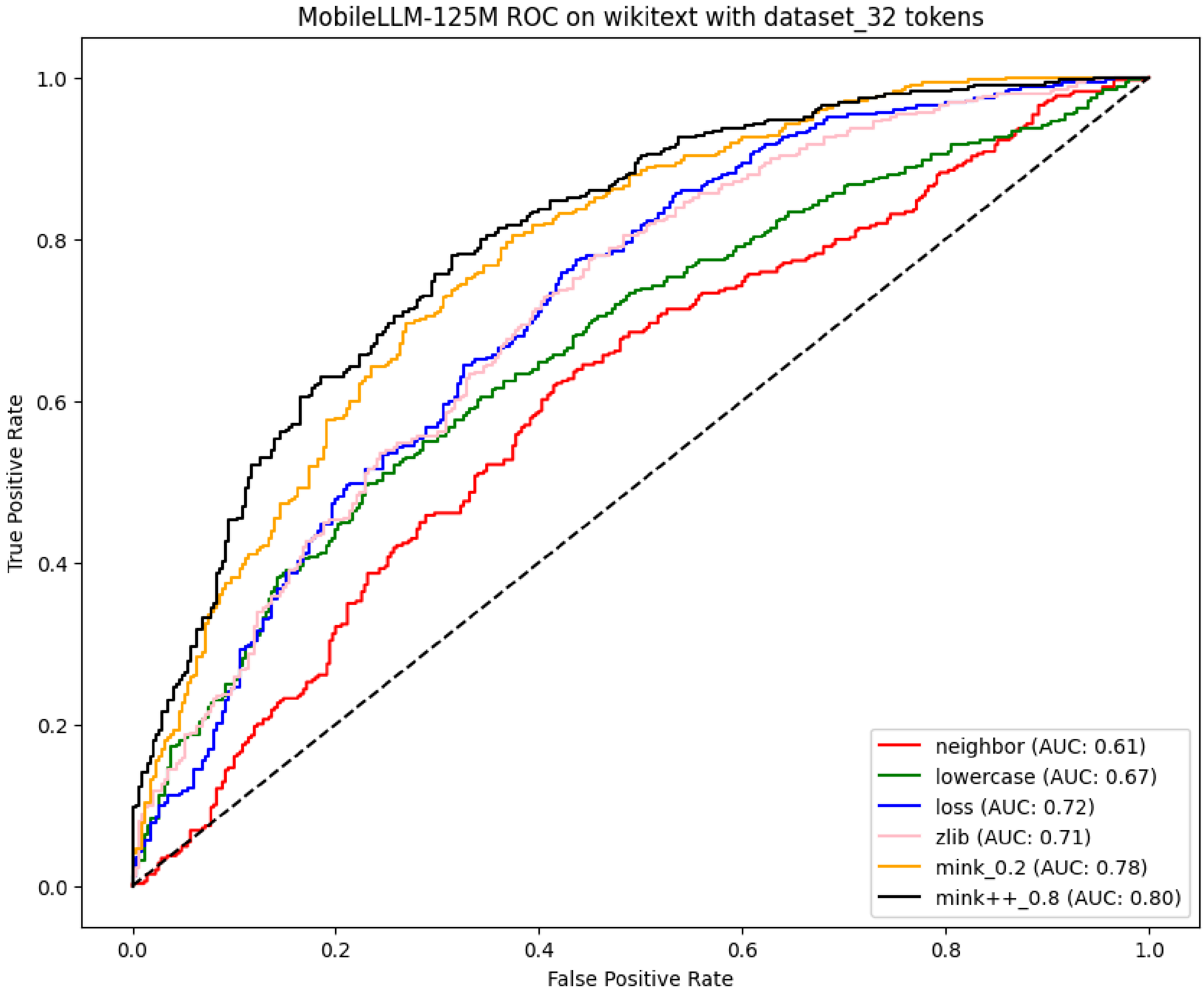
Results 2 : same model different datasets(MobileLLm350m)

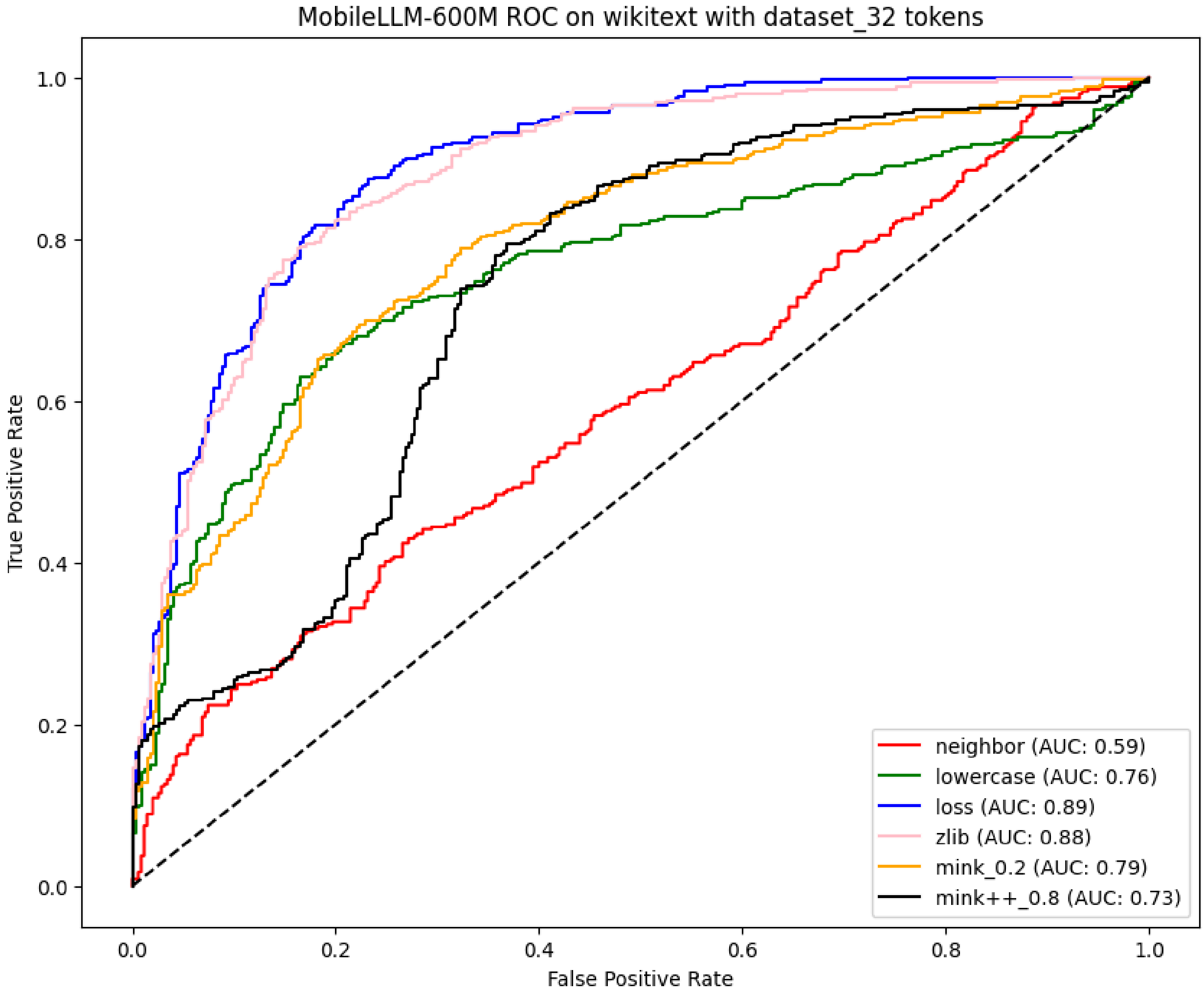


Appendix A: More plots




MobileLLM-1B ROC on wikitext with dataset_32 tokens
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
neighbor (AUC: 0.85)
lowercase (AUC: 0.95)
loss (AUC: 0.99)
zlib (AUC: 0.99)
mink_0.2 (AUC: 0.99)
mink++_0.8 (AUC: 0.98)


MobileLLM-125M ROC on wikitext with dataset_32 tokens
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
neighbor (AUC: 0.61)
lowercase (AUC: 0.67)
loss (AUC: 0.72)
zlib (AUC: 0.71)
mink_0.2 (AUC: 0.78)
mink++_0.8 (AUC: 0.80)


MobileLLM-600M ROC on wikitext with dataset_32 tokens
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
neighbor (AUC: 0.59)
lowercase (AUC: 0.76)
loss (AUC: 0.89)
zlib (AUC: 0.88)
mink_0.2 (AUC: 0.79)
mink++_0.8 (AUC: 0.73)

Sources:

Codes from papers:
[1] https://github.com/zjysteven/mink-plus-plus/blob/main/run.py
[2] https://github.com/justusmattern27/neighbour-mia/tree/main
[3] https://github.com/mireshghallah/neighborhood-curvature-mia/tree/main

Papers:
[4] Zhang, J., Sun, J., Yeats, E., Ouyang, Y., Kuo, M., Zhang, J., ... & Li, H. (2024). Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*.
[5] Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., & Berg-Kirkpatrick, T. (2023). Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*.
[6] Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., ... & Zettlemoyer, L. (2023). Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.
[7] Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018, July). Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)* (pp. 268-282). IEEE.
[8] Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., ... & Hajishirzi, H. (2024). Do membership inference attacks work on large language models?. *arXiv preprint arXiv:2402.07841*.
[9] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 2633-2650).