Noisy Neighbors: Efficient membership inference attacks against LLMs

Filippo Galli[*]
Scuola Normale Superiore
Scuola Superiore Sant'Anna
Pisa, Italy
filippo.galli@sns.it

Luca Melis
Meta Inc.

Tommaso Cucinotta
Scuola Superiore Sant'Anna
Pisa, Italy

Abstract

The potential of transformer-based LLMs risks being hindered by privacy concerns due to their reliance on extensive datasets, possibly including sensitive information. Regulatory measures like GDPR and CCPA call for using robust auditing tools to address potential privacy issues, with Membership Inference Attacks (MIA) being the primary method for assessing LLMs' privacy risks. Differently from traditional MIA approaches, often requiring computationally intensive training of additional models, this paper introduces an efficient methodology that generates *noisy neighbors* for a target sample by adding stochastic noise in the embedding space, requiring operating the target model in inference mode only. Our findings demonstrate that this approach closely matches the effectiveness of employing shadow models, showing its usability in practical privacy auditing scenarios.

1 Introduction

Advancements in natural language processing (Vaswani et al., 2017) have made large language models (LLMs) (Radford et al., 2019) essential for many text tasks. However, LLMs face issues like biases (Narayanan Venkit et al., 2023), privacy breaches (Carlini et al., 2021), and vulnerabilities (Wallace et al., 2021), underscoring the importance of protecting user privacy. The use of large datasets including personal information, has raised privacy concerns, leading to regulations such as GDPR (European Parliament, European Council, 2016) and CCPA (State of California, 2018).

Membership inference attacks (MIA) (Shokri et al., 2017) are effective auditing tools aiming at determining if a specific data point was used in an LLM's training dataset by analyzing its output. Such attacks highlight potential privacy breaches, relying on models' tendency to overfit to familiar data (Carlini et al., 2019). By employing calibration strategies and training shadow models, the accuracy of MIAs can be improved, although challenges such as computational demands and limitations in effectiveness when deviating from training distribution assumptions persist. In this paper, we contribute to this field by: i) exploring membership inference attacks from the standpoint of a privacy auditor, ii) introducing a computationally efficient calibration strategy that sidesteps training shadow models, and iii) empirically assessing its potential in replacing other prevalent strategies.

2 Background

LLMs generate a probability distribution over their vocabulary based on a tokenized input sequence converted into numerical inputs through an embedding layer. This layer maps tokens to a dense representation, which can be learned during training (Radford et al., 2018, 2019) or derived from public *word embeddings* (Devlin et al., 2018). For a model f with input sequence x, we define $\mathbb{P}[w|x] = f_w(x)$ as the conditional probability that the token following x is w. LLMs are typically trained on large datasets of text to minimize a measure of surprise in seeing the next token, called *perplexity*. For a sequence x, it is defined as the average negative log-likelihood of its tokens:

$$ppx(f, x) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \log(f_{x_t}(x_{<t})) \quad (1)$$

with $|x|$ the number of tokens in the sequence.

Membership inference attacks (Shokri et al., 2017; Watson et al., 2021; Carlini et al., 2022) aim to determine whether a particular data record x was used in the training dataset D_{train} of a machine learning model. These methods leverage model outputs like confidence scores or prediction probabilities to compute a score for the targeted sample. For LLMs, the typical assumption is to grant the

[*] Part of this author's work was carried out while at Meta Inc.

Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, pages 1–6
August 15, 2024 ©2024 Association for Computational Linguistics

adversary access to the output probabilities $f(x)$, which may be used to estimate the perplexity on the targeted samples as a score. Given a sample x, the goal of the attacker is to learn a thresholding classifier to output 1 when the perplexity is lower than a certain value γ:

$$A_\gamma(f, x) = \mathbb{1}[ppx(f, x) < \gamma] \tag{2}$$

MIA is a simple and effective tool to measure the privacy risk in a trained machine learning model, and it has interesting connections with other privacy frameworks. In particular, it is known to have a success rate bounded by the privacy parameters of Differential Privacy (DP) (Dwork et al., 2006). A randomized mechanism $\mathcal{M}$ is said to be ε-DP if for any two datasets D, D' that differ in at most one sample, and for any $R \subseteq \text{range}(\mathcal{M})$, we have:

$$\mathbb{P}[\mathcal{M}(D) \in R] \leq e^{\varepsilon}\mathbb{P}[\mathcal{M}(D') \in R] \tag{3}$$

Notably, DP quantifies the worst-case scenario of the privacy risk, so it is a fundamental tool in privacy assessment. From the performance of the thresholding classifier $\tilde{A}_\gamma(f, x)$ one can obtain a lower bound to the *empirical* ε-DP (Kairouz et al., 2015):

$$e^{\varepsilon} \geq \frac{TPR}{FPR} \tag{4}$$

with TPR and FPR being, respectively, the true and false positive rates, given a certain threshold.

3 Related works

Privacy attacks against language models is an active area of research and different refinements have been proposed. Some works have focused on an attacker where data poisoning is allowed, granting the adversary write access to the training dataset, to increase memorization (Tramèr et al., 2022) or in general to induce malicious behaviours (Xu et al., 2023; Wallace et al., 2021; Yan et al., 2023; Shu et al., 2024; Huang et al., 2020) and improve property inference attacks (Mahloujifar et al., 2022). Other works have adopted similar techniques to achieve actual training data extraction from the training set, with only query access to the trained model (Carlini et al., 2021, 2023).

In the context of MIAs with query access to the target model, most research focused on strategies to improve the calibration of the per-sample scores, i.e. techniques to improve the precision and recall in distinguishing members from non-members of the training set. In principle, if we can assert that an out-of-distribution non-member of the training set will induce a high perplexity in a target LLM, there are a number of scenarios where the distinction is not as clear cut, and a thresholding classifier essentially ends up distinguishing between in-distribution from out-of-distribution samples. A refined MIA then employs calibration strategies to tune the scoring function based on the difficulty of classifying the specific sample, as in (Watson et al., 2021). Thus, a relative membership score is obtained by comparing $f(x)$ with one of two results based on whether the adversary is assumed to have access to *neighboring models* $\tilde{f}(x)$ (Carlini et al., 2022; Watson et al., 2021) or *neighboring samples* $f(\tilde{x})$ (Mattern et al., 2023). The new classifier becomes:

$$\tilde{A}_\gamma(f, x) = \mathbb{1}[ppx(f, x) - p\tilde{p}x(f, x) < \gamma] \tag{5}$$

where $p\tilde{p}x(f, x)$ is the calibrated score over a set of neighboring models $ppx(\tilde{f}, x)$ or over a set of neighboring samples $ppx(f, \tilde{x})$. Neighboring models can be obtained by an adversary who is assumed to have some degree of knowledge of the training data distribution and trains a number of shadow models to mimic the behaviour of the target LLM. For instance (Carlini et al., 2022) trains multiple instances of the same architecture on different partitions of the training set, (Carlini et al., 2021) uses smaller architectures trained on roughly the same data, (Watson et al., 2021) leverages catastrophic forgetting of the target model under the assumption of white-box access. Neighboring samples do not require this assumption nor additional training and only need a strategy to craft inputs that are similar to the target sample under a certain distance metric. For instance, (Mattern et al., 2023) crafts neighboring sentences by swapping a number of words with their synonyms, showing good results but applicable primarily when the adversary has limited knowledge of the training data distribution. The authors then base the neighboring relationship in the *semantic* space, which is hard to quantify and fix, resulting in the need to generate a large number of neighbors to reduce the effects of these random fluctuations. Additionally, we emphasize how (Mattern et al., 2023) requires the use of an additional BERT-like model to generate synonyms, thus increasing the computational and memory cost of the attack. In (Tramèr et al., 2022) instead, calibration is done by comparing scores of the true inputs with scores of the lower-cased inputs. These strategies are known to be under-performing when knowledge of the training distribution is available, and are therefore proposed as an effective calibration mechanism when training shadow models is not possible.

4 Method

The intuition behind noisy neighbors is that, fixed a distance from a sample, the target model will show a larger difference in perplexity between a training sample and its neighbors than between a test sample and its neighbors. Thus, if we describe a language model as a composition of layers $f(x) = g(e(x))$ where e is an embedding layer and g is the rest of the network, one can artificially create neighbors in the n-dimensional embedding space by directly injecting random noise at the output of $e(x)$. In particular, if we create noisy neighbors by injecting Gaussian noise such that

$$f(x'_\sigma) = g(e(x)+\rho), \quad \text{with } \rho \sim \mathcal{N}(0, \sigma I_n) \quad (6)$$

then the Euclidean distance between the true and randomized input in the embedding space will be

$$\mathbb{E}[\|e(x) - e(x) - \rho\|] = \mathbb{E}[\|\rho\|] = \sigma\sqrt{n} \quad (7)$$

thus fixing, in expectation, the distance from the true sample at which the perplexity of the models will be evaluated. Generating multiple neighbors for each sample is crucial to mitigate randomness from stochastic noise, requiring repeated LLM inferences. Choosing the standard deviation σ potentially involves a complex parameter search with many model queries. However, the strategy's performance shows a clear peak at the optimal σ value, as shown in Figure 1, which can be efficiently identified using binary search.

We emphasize the challenge of isolating the embedding layer from the remainder of the network in an LLM when considering a scenario where an attacker has only black box access to the model. However, when this limitation does not apply, we think it is still within the capacity of an auditor to utilize a slightly stronger attacker model, where the first embedding layer is exposed, to save computational resources in simulating an adversary without access to the model architecture. Most importantly, in fact, we are inclined to explore this option as a more computationally efficient substitute for training shadow models for calibration, particularly in the context of auditing, rather than viewing it as a novel, realistic attack.

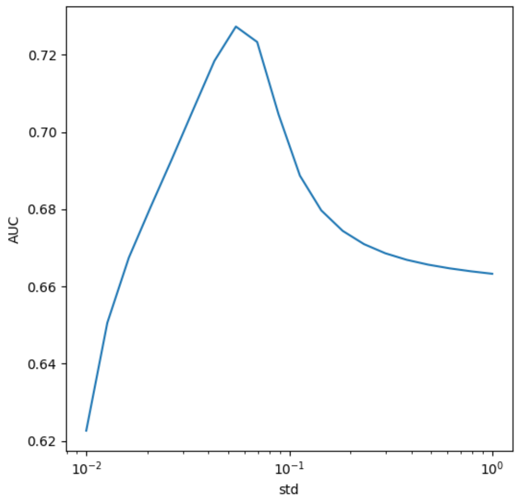


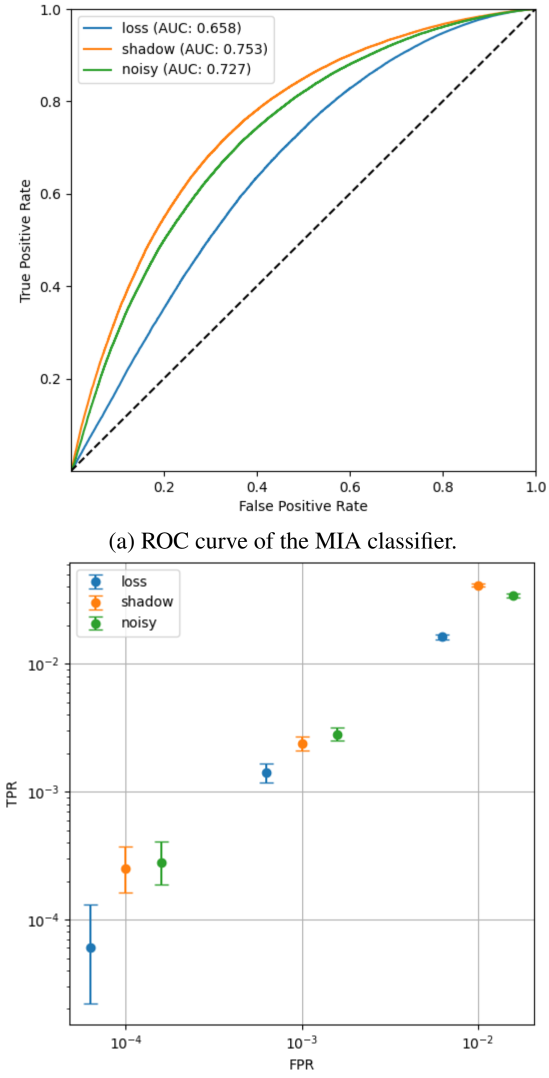
Figure 1: The AUC of the thresholding classifier for MIA shows a single and prominent peak at the optimal σ value in the *noisy neighbors* strategy.

5 Experiments

To validate the noisy neighbor strategy in implementing a calibrated MIA, we run a series of preliminary experiments on an LLM to gauge the risk of memorization of training data. The chosen architecture is GPT-2 *small* (Radford et al., 2019) to compromise learning capacity with memory and computational footprint at about 1.5 billion parameters, especially considering that competing strategies require training multiple LLMs from scratch. The model was pre-trained on OpenWebText (Gokaslan and Cohen, 2019), an open reproduction of the undisclosed WebText in (Radford et al., 2019). The model was then fine-tuned on 60% of the full WikiText corpus (Merity et al., 2016), a large collection of Wikipedia articles. The same data split was then partitioned in 10 subsets used to train 10 shadow models for score calibration, as in (Carlini et al., 2022). Note that Wikipedia articles are filtered out of the OpenWebtext corpus, to avoid data leakage in common benchmarks, such as ours. The remaining portion of 40% of WikiText is thus used as source of non-member, 126-token long samples to analyze the performance of the attack. We generate only 10 synthetic neighbors for each sample. Given a sample and its score, the thresholding classifier yields a binary decision on whether it was part of the training dataset or not. To determine how good the best possible classifier may be, we need to evaluate its accuracy at different thresholds. As it is common for binary classification problems, though, the accuracy does not give a complete picture of the confidence at which the classifier is able to tell apart members and non-members of the dataset. Thus Figure 2a shows the complete range of TPRs versus FPRs for the three main strategies we included in this comparison: score by perplexity (*loss*), by shadow model calibration (*shadow*), and by noisy neighbor (*noisy*) calibration. We have opted not to incorporate the *lowercasing* strategy (Tramèr et al., 2022) and the *semantic neighbor* approach (Mattern et al., 2023) in our study. These methods have, however, shown lower performance levels when information about the training data distribution is accessible, which is contemplated from the auditor point of view. Additionally, we faced challenges replicating some results from (Mattern et al., 2023), possibly due to limitations in the synonym generation technique described in (Zhou et al., 2019). Figure 2a also notes the Area Under the Curve (AUC), which for *noisy* and *shadow* amounts to 0.727 and 0.753 respectively, thus showing a discrepancy of only $\sim 3.4\%$. The AUC is an important metric for binary classifiers as it abstracts from the specific threshold, thus giving an average-case idea of the strength of the attacker. Still, as highlighted in (Carlini et al., 2022), special care should be given to what happens at low FPRs, that is when the attacker can confidently recognize members of the training set. This is what Figure 2b focuses on, again showing a strong overlap of the *shadow* and *noisy* strategies. Following Equation 4, we also provide the perspective of empirical DP, as the privacy community pushes to adopt this framework to comply to regulatory frameworks such as the GDPR (Cummings and Desai, 2018). Empirical DP measures the extent to which individual data points can be inferred or re-identified from the output of the system, and contrary to DP, it is a *post-hoc* measurement, not an *a-priori* guarantee. Figure 3 reports the results, where we see a strong consistency between the *noisy* and *shadow* strategies, especially for FPRs lower than 10^{-2}.

6 Limitations

The effectiveness of the noisy neighbors method depends on assumptions that may not apply universally across models or datasets. Its success also relies on specific noise parameters, potentially limiting its generalizability. Despite being computationally more efficient than shadow model methods, it still requires significant computational resources.



(a) ROC curve of the MIA classifier.

(b) Performance of the attacker at low FPRs.

Figure 2: Efficacy of different strategies for MIA. Confidence intervals are computed with the Clopper-Person method.

7 Conclusion

This work set out to elaborate a strategy for membership inference attacks. Differently from prior research focusing on improving the strength of the attacker, we develop a technique trying to achieve a similar efficacy, while reducing the computational burden for an auditor trying to assess the privacy risk of exposing the query access to a trained LLM. We propose the use of noise injection in the embedding space of the LLM to create synthetic neighbors of the targeted sample, to shift the comparison from the perplexity scored by different models on one sample, to the comparison of different samples by the same model. This approach allows to only use the model in inference mode, thus inherently re-

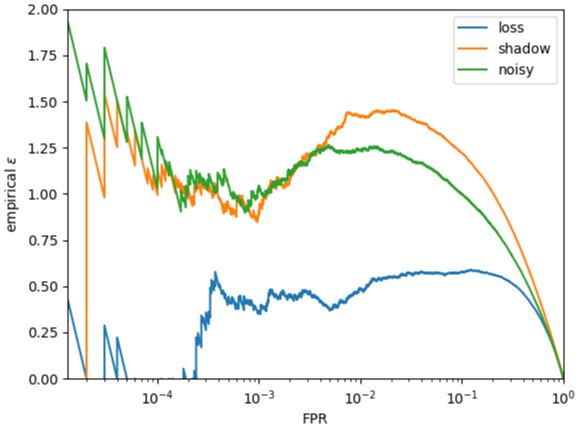


Figure 3: Empirical differential privacy measured downstream of training.

ducing the time and cost of running an MIA. With a number of experiments we assess how our strategy results converge to the results of using shadow models, showing a remarkable alignment.

References

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

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.

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.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270.

Rachel Cummings and Deven Desai. 2018. The role of differential privacy in gdpr compliance. In *FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 20.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

European Parliament, European Council. 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation).

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

W. Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. 2020. Metapoison: Practical general-purpose clean-label data poisoning. In *Advances in Neural Information Processing Systems*, volume 33, pages 12080–12091. Curran Associates, Inc.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2015. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR.

Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. 2022. Property inference from poisoning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1120–1137. IEEE.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2024. On the exploitability of instruction tuning. *Advances in Neural Information Processing Systems*, 36.

State of California. 2018. California Consumer Privacy Act (CCPA).

Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. 2022. Truth serum: Poisoning machine learning models to reveal their secrets. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2779–2792.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed data poisoning attacks on nlp models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. 2021. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2023. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *arXiv preprint arXiv:2305.14710*.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. 2023. Virtual prompt injection for instruction-tuned large language models. *arXiv preprint arXiv:2307.16888*.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.