Pandora's White-Box: Precise Training Data Detection and Extraction in Large Language Models

Anonymous Authors[1]

Introduction. As large language models (LLMs) are increasingly trained and fine-tuned on sensitive data in domains ranging from financial services to healthcare, the issue of *privacy* has emerged as an important consideration in the responsible deployment of these models. Recent work has focused on two types of privacy attacks especially relevant to LLMs: (i) membership inference attacks (MIA) (Shokri et al., 2017), and (ii) training data extraction (Carlini et al., 2020). In an MIA, an adversary with some level of access to the model seeks to reliably distinguish training samples from test samples. Highly accurate MIAs against LLMs would be useful to demonstrate privacy leakage as well as other applications like detecting train/test contamination, verifying if copyrighted data was in training data, quantifying memorization during training (Zhou et al., 2023), and evaluating unlearning success (Kurmanji et al., 2023; Pawelczyk et al., 2023; Hayes et al., 2024). Despite recent papers studying MIAs against LLMs (Duan et al., 2024; Li et al., 2023; Mattern et al., 2023), existing MIAs barely outperform random guessing when evaluated on aggregate, and some actually perform *worse than random guessing* at low False Positive (FPRs). The latter is particularly important because any attack which can extract even a subset of training data with high confidence poses a serious privacy risk (Carlini et al., 2022). This failure even prompted (Duan et al., 2024) to speculate that resistance of pretrained LLMs against MIAs might be an inherent property of the training process and distribution of train/test examples, rather than a shortcoming of existing attacks.

Pre-trained MIAs. Here we develop the first strong MIAs against pretrained LLMs, with True Positive Rates (TPRs) at low FPRs that are hundreds of times better than baseline attacks (Figure 1). We train a supervised attack that predicts training dataset membership given access to a small subset of training and test data. As input to the classifier we use the various p-norms of the gradient of loss with respect to weights at different layers in the network as well as the p-norms of the gradient with respect to the input embedding of a prompt ($p \in \{1, 2, \infty\}$). This approach is similar to that of (Nasr et al., 2019) who train supervised white-box attacks against classifiers, but with the complication of the high dimensionality of model gradients in LLMs. This necessitates dimensionality reductions on the scale of millions. For all Pythia model sizes up to 6.9B, we train a logistic regression $\texttt{LogReg}_\theta$ and a MLP $\texttt{NN}_\theta$ with 40,000 points and evaluate on 4000, and find that they significantly outperform the baselines across model sizes with high TPRs at low FPRs. We present the results for Pythia-1B in Table 1.

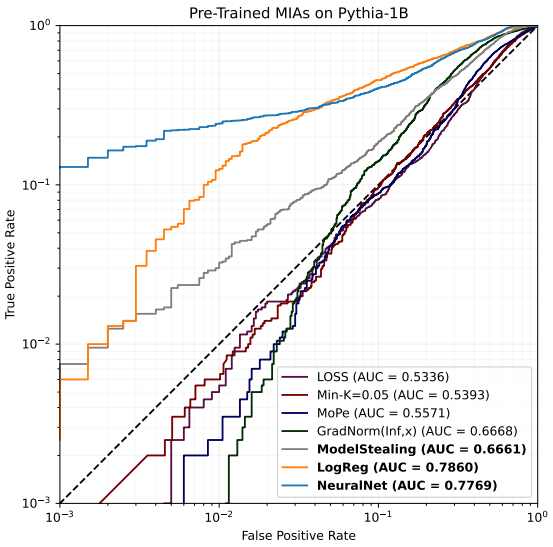

Figure 1: Our new attacks, **NeuralNet** and **LogReg**, perform *hundreds of times* better than previous state-of-the-art in low FPR regimes. Even without white-box access, our gray-box **ModelStealing** attack has competitive low FPR performance with white-box attacks. Full results in Table 1.

Gray-box MIA. Many of the most performant models like GPT-4o or Claude 3 remain closed source. Thus, we develop a less restrictive "gray-box" variant of our supervised MIA leveraging the recent model stealing attack of (Carlini et al., 2024). (Carlini et al., 2024) show, that given logit access to the model (e.g. via an API), one can recover the projec-

[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.

MIA	AUC	$\text{TPR}_{1\%}$
$\texttt{LOSS}_\theta$(Yeom et al., 2018)	0.534	0.00550
$\texttt{Min-K}_\theta(k = 0.05)$ (Shi et al., 2023)	0.539	0.00600
$\texttt{MoPe}_\theta$(Li et al., 2023)	0.557	0.00250
$\texttt{GradNorm}_\infty$ (Nasr et al., 2019)	0.667	0.000500
$\texttt{ModelStealing}_\theta$	0.666	0.032
$\texttt{LogReg}_\theta$	**0.786**	0.126
$\texttt{NN}_\theta$	0.777	**0.243**

Table 1: We record AUC and TPR value at FPR$= 1\%$ on the Pythia-1B model for a variety of pretrained MIAs, evaluated on 2,000 samples from The Pile's train and validation sets.

tion layer up to a rotation. This variant of our supervised MIA uses the loss gradient with respect to the projection layer as input features, which we can derive (up to some symmetries) using the model stealing attack. Table 1 shows this attack ($\texttt{ModelStealing}_\theta$) achieves state-of-the-art AUCs compared to prior black-box MIAs on pretrained LLMs.

Fine-tuned MIAs. Next, we consider fine-tuned models. These are increasingly relevant from a privacy perspective since companies and other institutions will often fine-tune open source pretrained language models on sensitive internal data, rather than training their own LLM from scratch. Prior work on memorization finds that data seen more recently during training is more vulnerable to membership inference (Zhou et al., 2023), and other work on MIAs against classifiers shows that access to multiple checkpoints can be combined to boost MIA attack accuracy (Jagielski et al., 2022). We are first to show that a simple MIA ($\texttt{FLoRa}_{\theta,\theta'}$) of a ratio of the losses of the fine-tuned and pretrained models is *nearly perfect* against fine-tuned LLMs; on Pythia-1B, $\texttt{FLoRa}_{\theta,\theta'}$ achieves an **AUC** of 0.987 and a $\textbf{TPR}_{1\%}$ of 92.8%.

Fine-tuned data extraction. We then turn to *data extraction*, where the adversary wishes to extract verbatim training examples. Using our MIAs in the generate-then-rank data extraction pipeline for LLMs (Carlini et al., 2020), we find that 50%+ of fine-tuning datasets can be extracted with $\texttt{FLoRa}_{\theta,\theta'}$ after 3 epochs of fine-tuning (standard in practice), with the simple strategy of (i) prompting the model 20 times and (ii) selecting the generation with the highest $\texttt{FLoRa}_{\theta,\theta'}$ score as a candidate training sample (Table 2). We study the discoverable extraction setting, where the adversary can prompt the model with the first $k = 50$ tokens of the example, and seeks to generate the remaining 50. Most suffixes are likely to be generated after 2 epochs of fine-tuning, and $\texttt{FLoRa}_{\theta,\theta'}$ is able to accurately identify $> 90\%$ of them.

Pretrained data-extraction. In the pretrained setting, unlike with fine-tuned models, the probability of generating training data across pretrained models remains low, so the probability of extracting a *specific* example is also low. As such, due to compute limitations, we cannot run a generate-then-rank pipeline for extracting pretrained LLM data. However, recent work has found that the upper limits of discoverable extraction are high. (Nasr et al., 2023) show that $> 1\%$ of generations from LLMs including Pythia and Llama-7B/Llama-7B-chat models are memorized training samples. This suggests that even though memorization levels are low, a mechanism that identifies training examples with high accuracy at low FPRs is an important threat vector. We explicitly train supervised white-box attacks (denoted $\texttt{Gen-LogReg}$, $\texttt{Gen-NN}$) for the task of classifying train points versus samples that share the same 50-token prefix but have a model-generated suffix. On Pythia-1.4B, this results in attacks that can identify the correct suffix 90.1% and 88.3% of the time out of 20 generations.

Model	Metric	Epoch 1	Epoch 2	Epoch 3
L7B	Precision	0.006	0.098	0.602
	01-Hamming	0.082	0.396	0.792
	$p > 0.1$	0.006	0.134	0.686
	$p > 0.05$	0.008	0.156	0.696
L7BC	Precision	0.004	0.144	0.622
	01-Hamming	0.103	0.478	0.812
	$p > 0.1$	0.008	0.186	0.658
	$p > 0.05$	0.008	0.216	0.668

Table 2: **L7B** refers to Llama-7B and **L7BC** refers to Llama-7B-Chat; we present metrics of extractability varying the number of epochs of fine-tuning. We report the attack precision and 01-Hamming (average proportion of tokens that match between the top-ranked generation and true suffix) with 20 generations and $\texttt{FLoRa}_{\theta,\theta'}$ as a ranking metric (rows 1-2), and the proportion of suffixes with generation probabilities $p > .01, .05, .01$ (rows 3-5). Across all suffixes, 66.8% had probability > 0.05 of being generated. After ranking 20 generations with $\texttt{FLoRa}_{\theta,\theta'}$, we accurately recover 62.2% of suffixes (precision) and have an average token-by-token match rate between top-ranked generation and true suffix of 81.2% (01-Hamming).

Conclusion. All together, our results have the following implications: 1) Very strong MIAs are possible in white-box settings against pre-trained LLMs, and in the black-box setting against fine-tuned LLMs. These MIAs are of independent interest outside of privacy, e.g. to detect if copyrighted data was used in training, prevent test/train contamination in evaluations, and evaluate machine unlearning. 2) The success of supervised white-box attacks can be partially extended to the black-box setting with recent model stealing attacks. 3) Finally, when LLMs are fine-tuned for over an epoch, they are highly vulnerable to data extraction attacks.

References

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D. X., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *USENIX Security Symposium*, 2020. URL https://api.semanticscholar.org/CorpusID:229156229.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles, 2022.

Carlini, N., Paleka, D., Dvijotham, K. D., Steinke, T., Hayase, J., Cooper, A. F., Lee, K., Jagielski, M., Nasr, M., Conmy, A., Wallace, E., Rolnick, D., and Tramèr, F. Stealing part of a production language model, 2024.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. Do membership inference attacks work on large language models?, 2024.

Hayes, J., Shumailov, I., Triantafillou, E., Khalifa, A., and Papernot, N. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy, 2024.

Jagielski, M., Wu, S., Oprea, A., Ullman, J., and Geambasu, R. How to combine membership-inference attacks on multiple updated models, 2022.

Kurmanji, M., Triantafillou, P., Hayes, J., and Triantafillou, E. Towards unbounded machine unlearning, 2023.

Li, M., Wang, J., Wang, J., and Neel, S. MoPe: Model perturbation based privacy attacks on language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13647–13660, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.842. URL https://aclanthology.org/2023.emnlp-main.842.

Mattern, J., Mireshghallah, F., Jin, Z., Schoelkopf, B., Sachan, M., and Berg-Kirkpatrick, T. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 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.

Nasr, M., Shokri, R., and Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, May 2019. doi: 10.1109/sp.2019.00065. URL http://dx.doi.org/10.1109/SP.2019.00065.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models, 2023.

Pawelczyk, M., Neel, S., and Lakkaraju, H. In-context unlearning: Language models as few shot unlearners, 2023.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models, 2023.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models, 2017.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting, 2018.

Zhou, Z., Xiang, J., Chen, C., and Su, S. Quantifying and analyzing entity-level memorization in large language models, 2023.