When MIA meets (L)LMs

MIA on LMs. One MIA approach on LMs solely based on the target model's loss for each sample, which is shown to be inconclusive [1]. On the other hand, similar to the shadow model setup proposed by Shokri et al. [2], the reference-based likelihood ratio attack [1] involves training a reference model on samples from the underlying population distribution that generates the training data for the target model. A record can be decided as a membership or not based on a threshold [1]. However, these MIA approaches only consider masked LM pre-training [3] or supervised fine-tuning [1, 5], where models are usually trained for more than 10 epochs [4]. Besides, it is observed that in general, member data seen more recently by the given checkpoint contributes to better MIA performance, resulting in high MIA performance on fine-tuning datasets [4].

Characteristics of LLMs. MIAs on LLMs are still a largely unexplored area [4]. Due to the massive scale of data and the tendency to overfit quickly, LLMs are typically trained for approximately one epoch [8,9]. Duan et al. (2024) hypothesize that *the large pretraining corpora characteristic of LMs decreases MIA performance as larger pretraining datasets lead to better generalization* [4]. These characteristics of LLMs limit the memorization of individual training data points, leading to lower MIA success rates. Furthermore, experiments show that even minor changes to a sample can alter the classification outcome [4].

Barriers to Effective MIAs on LLMs. Inherent ambiguity exists in natural language documents with a very high overlap between members and non-members. Furthermore, Selecting an appropriate reference model is difficult for LLMs, as it should be trained on the data disjoint from the dataset used to train the target model (representing the worst-case scenario for the attacker) [2, 4]. However, enforcing this strict assumption is difficult due to the vast scale of LLM pre-training corpora [4].

Future directions.

- **Revisiting membership.** Redefine membership in the context of information leakage in generative models by extending to sufficiently similar samples, measured by lexical or semantic distance [4].
- **User-level inference.** Inference attacks can target the user level to determine if a user's data was included in a dataset, extending membership inference from individual data samples to the broader privacy concerns of users who may contribute multiple samples [6].
- **Contextual privacy**. Explore the reasoning capabilities of LLMs in identifying and handling sensitive data within interactive, context-driven scenarios, where LLMs may generate and share outputs derived from sensitive input data [7].
- **Extraction attack.** Extraction attacks can be performed on LLMs by using sufficient length prefixes and with additional measures to reveal leakage risks [4, 6]. Recently this 'extractability' is also used to investigate the memorization across models [4].

Reference

[1] Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., & Shokri, R. (2022). Quantifying privacy risks of masked language models using membership inference attacks. arXiv preprint arXiv:2203.03929.

[2] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.

[3] Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., & Wallace, B. C. (2021). Does BERT pretrained on clinical notes reveal sensitive data?. arXiv preprint arXiv:2104.07762.

[4] 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.

[5] Fu, W., Wang, H., Gao, C., Liu, G., Li, Y., & Jiang, T. (2023). Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. arXiv preprint arXiv:2311.06062.

[6] Kandpal, N., Pillutla, K., Oprea, A., Kairouz, P., Choquette-Choo, C. A., & Xu, Z. (2023). User inference attacks on large language models. arXiv preprint arXiv:2310.09266.

[7] Mireshghallah, N., Kim, H., Zhou, X., Tsvetkov, Y., Sap, M., Shokri, R., & Choi, Y. (2023). Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. arXiv preprint arXiv:2310.17884.

[8] Komatsuzaki, A. (2019). One epoch is all you need. arXiv preprint arXiv:1906.06669.

[9] Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., ... & Raffel, C. A. (2023). Scaling data-constrained language models. Advances in Neural Information Processing Systems, 36, 50358-50376.