Matteo Boglioni

Research Statement

Introduction

In recent years, increasing scientific interest has grown in the privacy and security of Machine Learning models, especially for Large Language Models. With an overall advancement in performance, there's a strengthening need for guarantees and measurement of their behavior when deployed in real-world scenarios. I'm particularly interested in red-teaming and model auditing as proactive techniques to uncover existing models' vulnerabilities. I aim to assess and expose these flaws and inconsistencies, both in performance and security, to increase robustness and reliability towards distribution shifts, adversarial inputs, and unintended behaviors. I believe this step is crucial for a responsible and ethical deployment of AI.

Research Background and Contributions

My previous research experience has been focusing mainly on:

Membership Inference Attacks

As part of my Master's curriculum, I conducted an individual research project to investigate Membership Inference vulnerabilities over diffusion-models-based synthetic tabular data. Building on prior work by Carlini et al. [2022, 2023] over generative models, this investigation aimed to extend previous findings, mainly in the white-box scenario where the attacker has access to both the training canaries and the model's weights. The results highlighted a behavior roughly consistent with membership inference attacks over image-generative models, despite the lower dimensionality of tabular data and the respective weaker signal it carries.

Differential-Privacy Auditing

Differential Privacy (DP) has become a popular framework for protecting individual data: algorithms like DP-SGD [Abadi et al., 2016] used to train machine learning models, offer strong formal privacy guarantees against information leakage. However, previous works [Tramer et al., 2022] have highlighted inconsistencies in the theoretical privacy provided in the implementation of these algorithms. DP Auditing aims to provide empirical lower bounds on the privacy guarantees of DP-SGD. My Master's Thesis, as a visiting student at Carnegie Mellon University, has focused on improving efficient and scalable techniques [Bertran et al., 2023, Steinke et al., 2023, Mahloujifar et al., 2024] to reduce the gap between the empirical lower bounds and theoretical upper bounds. This research investigates both low computational requirements Membership Inference Attacks (MIAs) and canary optimization approaches, leading to improvements in both the tightness of the obtained bounds and the resources needed to perform auditing, in a realistic setting such as the black box, where the attacker is not allowed to manually inject gradients during training. While this project is still in progress, an early portion of this research has been accepted as a Workshop Paper at *Theory and Practice of Differential Privacy* (TPDP), 2025.

OOD Generalization

Out-of-distribution (OOD) generalization abilities have gradually become a crucial aspect of LLM deployment, as often the samples observed in real-world scenarios do not match the training data distribution. Previous work assessing LLMs' OOD performance, however, typically focuses on a single out-of-distribution dataset [Mosbach et al., 2023, Bhargava et al., 2021]. This approach may fail to precisely evaluate the capabilities of the model, as the data shifts encountered once a model is deployed are much more diverse. This project investigated whether such OOD generalization results generalize, evaluating correlations between a model's performance across multiple OOD test sets during finetuning. The results highlight the need for several OOD splits to ensure a fair evaluation. Additionally, we observe inconsistencies, in the generalization patterns, depending on both model architecture and fine-tuning setting. This work has been submitted to the *2025 Conference on Empirical Methods in Natural Language Processing* (EMNLP) and is currently under review.

Current Research Interests

My current research interests, aligned with my previous work, include:

OOD LLM Benchmarking

During my previous contribution, the understanding of covariate shifts and their impact on the model performance was primarily empirical. I would be interested in advancing more principled and transferable approaches to benchmarking LLMs on OOD splits. More specifically, investigate which distribution shifts affect the model performance and if these variations might be consistent among different NLP tasks. This involves testing the learned heuristics in a controlled and interpretable way. I believe that this direction is crucial to improve the models' reliability and robustness.

Federated Learning Privacy Auditing

In recent years, Federated Learning (FL) has emerged as a new paradigm for privacy-preserving machine learning. FL enables training models in a decentralized manner without requiring the sharing of private training data. This setup combines a scalable privacy-preserving approach with an overall improvement in model utility. However, other work has evidenced vulnerabilities in these models suggesting interesting research directions [Nasr et al., 2019, Kairouz et al., 2021]. I would be particularly interested in developing privacy auditing tools that can better characterize the privacy risks associated with this new machine learning pattern.

Conclusion

During the years of my Master's program, I have had the opportunity to explore different research areas from LLM Generalization to Differential Privacy, while collaborating with multiple labs to experiment in multiple research environments. Looking ahead, I aim to contribute to the development of machine learning models that are not only powerful but also safe, reliable, and robust.

References

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

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

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Florian Tramer, Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, and Nicholas Carlini. Debugging differential privacy: A case study for privacy auditing. *arXiv e-prints*, pages arXiv–2202, 2022.

Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems*, 36:314–330, 2023.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 36:49268–49280, 2023.

Saeed Mahloujifar, Luca Melis, and Kamalika Chaudhuri. Auditing f-differential privacy in one run. *arXiv preprint arXiv:2410.22235*, 2024.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, 2023.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. Generalization in NLI: Ways (not) to go beyond simple heuristics. In João Sedoc, Anna Rogers, Anna Rumshisky, and Shabnam Tafreshi, editors, *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: $10.18653/v1/2021.insights\text{-}1.18$. URL `https://aclanthology.org/2021.insights-1.18/`.

Milad Nasr, Reza Shokri, and Amir Houmansadr. 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)*, pages 739–753, 2019. doi: $10.1109/SP.2019.00065$.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210, 2021.