Wenxin Ding

Research Interest: Machine learning security and privacy

Education

University of Chicago Chicago, IL
Ph.D. in Computer Science (4.0 / 4.0) 2021.9 – 2026.6

Carnegie Mellon University Pittsburgh, PA
M.S. in Computer Science – Research Thesis (4.11 / 4.33) 2020.8 – 2021.8

B.S. in Computer Science and B.S. in Mathematical Sciences (3.92 / 4.0) 2016.8 – 2020.5

Work Experience

Research Assistant, University of Chicago (Chicago, IL) 2021.9 – Present
Advisors: Prof. Heather (Haitao) Zheng and Prof. Ben Y. Zhao

My research focuses on the security and privacy of machine learning systems, with an emphasis on bridging the gap between theoretical understanding and empirical practice. Specifically, I study the safety and robustness of machine learning models when exposed to strategically optimized or adversarial training data. I have developed methodologies for model recovery using augmented datasets and derived theoretical guarantees on the optimal loss achievable by robust classifiers. More recently, I have explored vulnerabilities in text-to-image diffusion models — I designed and evaluated data poisoning attacks against generative models and derived a theoretical analysis on model performance under a range of adversarial scenarios. My PhD research received University of Chicago's Harper Dissertation Fellowship award.

Machine Learning Engineer, Qualcomm (San Diego, CA) 2025.6 – 2025.9
Mentors: Dr. Jonathan Petit and Dr. Cong Chen

At Qualcomm, I worked in the AI safety team to build benchmarks for evaluating safety of vision language models (VLMs) commissioned by MLCommons. I evaluated existing jailbreak attacks against a variety of VLMs and designed effective attacks on safety benchmark datasets, demonstrating the lack of robustness against jailbreak attacks in existing VLMs. My research at Qualcomm contributed to a white paper released by MLCommons and a technical paper under submission.

Research Assistant, Carnegie Mellon University (Pittsburgh, PA) 2019.2 – 2021.8
Advisors: Prof. Nihar B. Shah and Prof. Weina Wang

Motivated by the need to improve the peer-review system in use, I conducted research on releasing essential peer-review data and practicing calibration in peer-review process. I applied differential privacy to protect the anonymity of privacy-sensitive peer-review data and designed efficient algorithm to enhance utility of the protected data. I proved the Pareto-optimal strategy to calibrate peer-review ratings, balancing between privacy of reviewers and utility of the peer-review outcome.

Publications

- **Wenxin Ding**, Cong Chen, Jean-Philippe Monteuuis, and Jonathan Petit. "Safe but not Robust: Security Evaluation of VLM by Jailbreaking MSTS." *under submission.*
- James Goel et al. "AILuminate Security Introducing v0.5 of the Jailbreak Benchmark from MLCommons." *under submission.*
- **Wenxin Ding**, Cathy Li, Shawn Shan, Ben Y. Zhao, Haitao Zheng. "Understanding Implosion in Text-to-Image Generative Models." *in Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS), 2024.*

- Shawn Shan, **Wenxin Ding**, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models." *in Proceedings of IEEE Symposium on Security and Privacy (S&P), 2024.*
- **Wenxin Ding**, Arjun Nitin Bhagoji, Ben Y. Zhao, and Haitao Zheng. "Towards Scalable and Robust Model Versioning." *in Proceedings of IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 2024.*
- Sihui Dai*, **Wenxin Ding***, Arjun Nitin Bhagoji, Daniel Cullina, Ben Y. Zhao, Haitao Zheng, and Prateek Mittal. "Characterizing the Optimal 0-1 Loss for Multi-class Classification with a Test-time Attacker." *in Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2023*. Spotlight paper.
  (* for equal contribution)
- Shawn Shan, **Wenxin Ding**, Emily Wenger, Haitao Zheng, and Ben Y. Zhao. "Post-breach recovery: Protection against white-box adversarial examples for leaked DNN models." *in Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS), 2022*.
- **Wenxin Ding**, Gautam Kamath, Weina Wang, and Nihar B. Shah. "Calibration with privacy in peer review." *in Proceedings of IEEE International Symposium on Information Theory (ISIT), 2022*.
- **Wenxin Ding**, Nihar B. Shah, and Weina Wang. "On the privacy-utility tradeoff in peer-review data analysis." *AAAI Privacy-Preserving Artificial Intelligence (PPAI) workshop, 2021*. Spotlight paper.

Services

Technical Program Committee

- ACM Conference on Computer and Communications Security (CCS), *2025 & 2026*
- IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), *2025 & 2026*
- ACM Workshop on Artificial Intelligence and Security (AISec), *2024*

Reviewer

- Nature
- SIAM Journal on Mathematics of Data Science (SIMODS)
- ACM SIGGRAPH Asia, *2025*
- The Conference on Uncertainty in Artificial Intelligence (UAI), *2024 & 2025*

Awards

- 2025 University of Chicago William Rainey Harper Dissertation Fellowship
- 2024 University of Chicago UU Fellowship
- 2021—2026 University of Chicago Eckhardt Scholar
- 2020 Carnegie Mellon University Senior Leadership Recognition
- 2019 Mark Stehlik SCS Alumni Undergraduate Impact Scholarship
- 2017 William Lowell Putnam Mathematical Competition (Rank: 255 / 4638)

Teaching Experience

Teaching Assistant

- Adversarial Machine Learning (University of Chicago)
- Mathematical Foundations of Machine Learning (University of Chicago)
- Principles of Computing (Carnegie Mellon University)
- Introduction to Computer Systems (Carnegie Mellon University)
- Distributed Systems (Carnegie Mellon University)