TONG WU

tongwu@princeton.edu
https://tongwu2020.github.io/tongwu/

RESEARCH INTEREST

Trustworthy machine learning and large language model. Recently, I have built **robust** and **private** retrieval-augmented generation (RAG) systems with theoretical guarantees.

EDUCATION

Princeton University *August 2021 - Present*
Ph.D. in Electrical and Computer Engineering
Advisor: Prateek Mittal

Washington University in St. Louis *August 2018 - May 2021*
B.S./M.S. in Computer Science and Mathematics
Advisor: Yevgeniy Vorobeychik

RESEARCH & PROFESSIONAL EXPERIENCE

Princeton University *August 2021 - Present*
Research Assistant *Princeton, NJ*

- Developed the first certifiably robust retrieval-augmented generation (RAG) system.
- Designed a novel In-context Learning paradigm to mitigate the privacy issues of Large Language Models.
- Uncovered hidden adversarial risks of deploying test-time adaptation (transductive learning paradigm).
- ***Key words***: Large Language Models, Retrieval-augmented Generation, Security & Privacy.

Microsoft, Inc. (Responsible & OpenAI Research) *August 2023 - September 2023*
Research Intern *Princeton, NJ*

- Conducted model distillation to significantly reduce the size of a content moderation model, achieving a ∼10× reduction in size with a mere 2% degradation in performance.

NEC Laboratories America, Inc. *May 2021 - August 2021*
Research Intern *Princeton, NJ*

- Proposed a model personalization (meta-learning) framework for event detection of dialysis patients.
- Adapted covariance transfer and adversarial attacks to conduct OOD detection in few-shot learning.

Washington University in St. Louis *Dec 2018 - May 2021*
Research Assistant *St. Louis, MO*

- Studied the problem of defending deep neural network approaches from physically realizable attacks and demonstrated that the state-of-the-art robust models exhibit limited effectiveness.
- Proposed a new abstract model, ROA, where an adversary places a small crafted rectangle that fools the image classifier, and adversarial training using ROA achieved much better robustness than all SOTA.

PUBLICATIONS

1. Chong Xiang*, **Tong Wu***, Zexuan Zhong, David Wagner, Danqi Chen, Prateek Mittal. Certifiably Robust RAG against Retrieval Corruption. *ArXiv preprint*, 2024.
2. Sihui Dai, Chong Xiang, **Tong Wu**, Prateek Mittal. Position Paper: Beyond Robustness Against Single Attack Types. *ArXiv preprint*, 2024.

3. Chong Xiang, **Tong Wu**, Sihui Dai, Jonathan Petit, Suman Jana, Prateek Mittal. PatchCURE: Improving Certifiable Robustness, Model Utility, and Computation Efficiency of Adversarial Patch Defenses. In *USENIX Security Symposium (Security)*, 2024.
4. **Tong Wu***, Ashwinee Panda*, Jiachen T. Wang*, Prateek Mittal. Privacy-Preserving In-Context Learning for Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024.
5. Jiachen T. Wang, Saeed Mahloujifar, **Tong Wu**, Ruoxi Jia, Prateek Mittal. A Randomized Approach for Tight Privacy Accounting. In *Neural Information Processing Systems (NeurIPS)*, 2023.
6. Xiangyu Qi, Tinghao Xie, Jiachen T. Wang, **Tong Wu**, Saeed Mahloujifar, Prateek Mittal. Towards A Proactive ML Approach for Detecting Backdoor Poison Samples. In *USENIX Security Symposium (Security)*, 2023.
7. **Tong Wu**, Feiran Jia, Xiangyu Qi, Jiachen T. Wang, Vikash Sehwag, Saeed Mahloujifar, Prateek Mittal. Uncovering Adversarial Risks of Test-Time Adaptation. In *International Conference on Machine Learning (ICML)*, 2023.
8. Chong Xiang, Chawin Sitawarin, **Tong Wu**, Prateek Mittal. Short: Certifiably Robust Perception Against Adversarial Patch Attacks: A Survey. In *VehicleSec*, 2023. **Best Short/WIP Paper Award Runner-Up**
9. **Tong Wu**, Tianhao Wang, Vikash Sehwag, Saeed Mahloujifar, Prateek Mittal. Just Rotate it: Deploying Backdoor Attacks via Rotation Transformation. In *AISec*, 2022.
10. Shaojie Wang, **Tong Wu**, Ayan Chakrabarti, Yevgeniy Vorobeychik. Adversarial Robustness of Deep Sensor Fusion Models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* , 2022.
11. **Tong Wu**, Liang Tong, Yevgeniy Vorobeychik. Defending Against Physically Realizable Attacks on Image Classification. In *International Conference on Learning Representations (ICLR)*, 2020. **Spotlight Presentation**.

PATENTS

1. Yevgeniy Vorobeychik, **Tong Wu**, Liang Tong. Systems and Methods for Defending against Physical Attacks on Image Classification. US Patent App. 17/214,071, 2021.

REVIEWING

- *Journals:* IJCV
- *Conferences:* ICLR'22,24, NeurIPS'22,23, 24 ICML'23,24, IEEE S&P'21, ECCV'24, KDD'22, WCAV'22, 24, AAAI'21, AML-CV'21

TEACHING EXPERIENCE

- Teaching Assistant of Information Security (Spring 2024), Princeton University.
- Teaching Assistant of Introduction to Machine Learning (Spring 2019, Fall 2019, Spring 2020, Spring 2021), Washington University in St. Louis.

HONORS & AWARDS

- Research Excellence Award at Washington University, 2021
- AAMAS 2021 Student Scholarship, 2021
- Member of Tau Beta Pi Association