TONG WU

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

RESEARCH INTEREST

Large Language Model safety, alignment, and reasoning. I am interested in addressing the safety challenges posed by increasingly capable and intelligent LLM systems through effective control mechanisms.

EDUCATION

Princeton University *Aug. 2021 - May 2026*
Ph.D. in Electrical and Computer Engineering
Advisor: Prateek Mittal

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

RESEARCH & PROFESSIONAL EXPERIENCE

Princeton University *Aug. 2021 - Present*
Research Assistant *Princeton, NJ*

· Proposed a novel paradigm to guide the internal reasoning processes of LLMs safely and effectively.
· Developed a certified robust Retrieval-augmented Generation (RAG) system against corrupted retrieval.
· Designed a novel In-context Learning paradigm to mitigate the privacy issues of Large Language Models.

Zoom Video Communications, Inc. *May 2024 - Aug. 2024*
Research Intern *(Remote) Princeton, NJ*

· Developed a method to enhance the LLM system with robustly following the hierarchical instructions.

Microsoft, Inc. (Responsible & OpenAI Research) *Aug. 2023 - Sep. 2023*
Research Intern *(Remote) Princeton, NJ*

· Developed an efficient content moderation model that is 10× faster while retaining 99% accuracy.

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

· Proposed a model personalization (meta-learning) framework for event detection of dialysis patients.

SELECTED PUBLICATIONS

1. **Tong Wu**, Chong Xiang, Jiachen T. Wang, G. Edward Suh, Prateek Mittal. Effectively Controlling Reasoning Models through Thinking Intervention. *ArXiv preprint*, 2025.
2. **Tong Wu**, Shujian Zhang, Kaiqiang Song, Silei Xu, Sanqiang Zhao, Ravi Agrawal, Sathish Reddy Indurthi, Chong Xiang, Prateek Mittal, Wenxuan Zhou. Instructional Segment Embedding: Improving LLM Safety with Instruction Hierarchy. In *International Conference on Learning Representations (ICLR)*, 2025.
3. Chong Xiang*, **Tong Wu***, Zexuan Zhong, David Wagner, Danqi Chen, Prateek Mittal. Certifiably Robust RAG against Retrieval Corruption. *ArXiv preprint*, 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. **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.

6. **Tong Wu**, Liang Tong, Yevgeniy Vorobeychik. Defending Against Physically Realizable Attacks on Image Classification. In *International Conference on Learning Representations (ICLR)*, 2020. **Spotlight Presentation**.

FULL PUBLICATIONS

1. Feiran Jia, **Tong Wu**, Xin Qin, Anna Squicciarini. The Task Shield: Enforcing Task Alignment to Defend Against Indirect Prompt Injection in LLM Agents. *The Association for Computational Linguistics (ACL)*, 2025.

2. Jiachen T. Wang, **Tong Wu**, Dawn Song, Prateek Mittal, Ruoxi Jia. GREATS: Online Selection of High-Quality Data for LLM Training in Every Iteration. In *Neural Information Processing Systems (NeurIPS)*, 2024. **Spotlight Presentation**.

3. Sihui Dai, Chong Xiang, **Tong Wu**, Prateek Mittal. Position Paper: Beyond Robustness Against Single Attack Types. *ArXiv preprint*, 2024.

4. 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.

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. 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**

8. **Tong Wu**, Tianhao Wang, Vikash Sehwag, Saeed Mahloujifar, Prateek Mittal. Just Rotate it: Deploying Backdoor Attacks via Rotation Transformation. In *AISec*, 2022.

9. 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.

REVIEWING

- ICLR'22,24,25; NeurIPS'22,23,24; ICML'23,24,25; CVPR'25; ECCV'24; WCAV'22,24,25; KDD'22; AAAI'21; S&P'21; IJCV

HONORS & AWARDS

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