Tong Wu

+1-765-720-4989 | tongwu@princeton.edu | Personal Website

LinkedIn | GitHub | Google Scholar | X (Twitter)

Research Interest

My research aims to address the safety challenges of increasingly capable LLMs through simple, scalable methods grounded in rigorous theoretical principles. My work spans:

- **Reasoning Control.** I am intrigued by the frontier reasoning LLMs. My recent work focuses on scalable techniques to improve their reliability and safety through *Thinking Intervention*, a novel paradigm for steering their internal reasoning trajectories. I also investigate the implications of the *Inverse Scaling Law* on robustness in safety tasks.
- **Robust and Private RAG.** I have developed a certifiably robust retrieval-augmented generation system (*RobustRAG*) with responses provably resilient to perturbations in retrieved documents. I also designed a differentially private in-context learning (*DP-ICL*) method that protects the context privacy.
- **Instruction Hierarchy.** I have worked on enhancing instruction hierarchy through architectural innovations (*ISE*) and robust verification mechanisms (*Task Shield*) that correctly prioritize critical instructions in LLM agents.

Education

- **Princeton University** — *Aug. 2021 – Dec. 2025 (Expected)*
  Ph.D. in Electrical and Computer Engineering — Princeton, NJ, USA
  ◦ Advisor: Prof. Prateek Mittal
- **Washington University in St. Louis** — *Aug. 2018 – May 2021*
  B.S./M.S. in Computer Science and Mathematics — St. Louis, MO, USA
  ◦ Advisor: Prof. Yevgeniy Vorobeychik

Experience

- **Princeton University [🌐]** — *Aug. 2021 – Present*
  ◦ Proposed *Thinking Intervention*, a novel paradigm for safe and effective control of LLM reasoning processes.
  ◦ Developed a certifiably robust Retrieval-Augmented Generation (RAG) system resilient to corrupted retrieval.
  ◦ Designed a privacy-preserving in-context learning framework to mitigate LLM privacy risks.
- **Google [🌐]** — *May 2025 – Aug. 2025*
  ◦ Designing rigorous instruction hierarchy mechanisms for the next generation of the Gemini reasoning model.
- **Zoom Video Communications [🌐]** — *May 2024 – Aug. 2024*
  ◦ Proposed a novel architectural framework to strengthen instruction hierarchy in LLM.
- **Microsoft [🌐]** — *Aug. 2023 – Sep. 2023*
  ◦ Developed a high-efficiency content moderation model with a 10× speedup and 99% accuracy retention.
- **NEC Laboratories America [🌐]** — *May 2021 – Aug. 2021*
  ◦ Proposed a meta-learning framework for model personalization in event detection of dialysis patients.

Publications

*= Equal Contribution

Selected Publications

1. **Tong Wu**, Chong Xiang, Jiachen T. Wang, Weichen Yu, Chawin Sitawarin, Vikash Sehwag, Prateek Mittal (2025). **Does More Inference-Time Compute Really Help Robustness?**. *arXiv preprint*.
2. **Tong Wu**, Chong Xiang, Jiachen T. Wang, G. Edward Suh, Prateek Mittal (2025). **Effectively Controlling Reasoning Models through Thinking Intervention**. *arXiv preprint*.
3. **Tong Wu**, Shujian Zhang, Kaiqiang Song, Silei Xu, Sanqiang Zhao, Ravi Agrawal, Sathish Reddy Indurthi, Chong Xiang, Prateek Mittal, Wenxuan Zhou (2025). **Instructional Segment Embedding: Improving LLM Safety with Instruction Hierarchy**. In *International Conference on Learning Representations (ICLR)*.
4. Chong Xiang*, **Tong Wu***, Zexuan Zhong, David Wagner, Danqi Chen, Prateek Mittal (2024). **Certifiably Robust RAG against Retrieval Corruption**. *arXiv preprint*.

5. **Tong Wu***, Ashwinee Panda*, Jiachen T. Wang*, Prateek Mittal (2024). **Privacy-Preserving In-Context Learning for Large Language Models**. In *International Conference on Learning Representations (ICLR)*.
6. Feiran Jia, **Tong Wu**, Xin Qin, Anna Squicciarini (2025). **The Task Shield: Enforcing Task Alignment to Defend Against Indirect Prompt Injection in LLM Agents**. In *The Association for Computational Linguistics (ACL)*.
7. **Tong Wu**, Feiran Jia, Xiangyu Qi, Jiachen T. Wang, Vikash Sehwag, Saeed Mahloujifar, Prateek Mittal (2023). **Uncovering Adversarial Risks of Test-Time Adaptation**. In *International Conference on Machine Learning (ICML)*.
8. **Tong Wu**, Liang Tong, Yevgeniy Vorobeychik (2020). **Defending Against Physically Realizable Attacks on Image Classification**. In *International Conference on Learning Representations (ICLR)*. *Spotlight Presentation*.

Full Publications

9. Sihui Dai, Chong Xiang, **Tong Wu**, Prateek Mittal (2024). **Position Paper: Beyond Robustness Against Single Attack Types**. *arXiv preprint*.
10. Jiachen T. Wang, **Tong Wu**, Dawn Song, Prateek Mittal, Ruoxi Jia (2024). **GREATS: Online Selection of High-Quality Data for LLM Training in Every Iteration**. In *Neural Information Processing Systems (NeurIPS)*. *Spotlight Presentation*.
11. Chong Xiang, **Tong Wu**, Sihui Dai, Jonathan Petit, Suman Jana, Prateek Mittal (2024). **PatchCURE: Improving Certifiable Robustness, Model Utility, and Computation Efficiency of Adversarial Patch Defenses**. In *USENIX Security Symposium (USENIX)*.
12. Jiachen T. Wang, Saeed Mahloujifar, **Tong Wu**, Ruoxi Jia, Prateek Mittal (2023). **A Randomized Approach for Tight Privacy Accounting**. In *Neural Information Processing Systems (NeurIPS)*.
13. Xiangyu Qi, Tinghao Xie, Jiachen T. Wang, **Tong Wu**, Saeed Mahloujifar, Prateek Mittal (2023). **Towards a Proactive ML Approach for Detecting Backdoor Poison Samples**. In *USENIX Security Symposium (USENIX)*.
14. Chong Xiang, Chawin Sitawarin, **Tong Wu**, Prateek Mittal (2023). **Short: Certifiably Robust Perception Against Adversarial Patch Attacks: A Survey**. In *VehicleSec*. *Best Short/WIP Paper Award Runner-Up*.
15. **Tong Wu**, Tianhao Wang, Vikash Sehwag, Saeed Mahloujifar, Prateek Mittal (2022). **Just Rotate It: Deploying Backdoor Attacks via Rotation Transformation**. In *Artificial Intelligence and Security (AISec)*.
16. Shaojie Wang, **Tong Wu**, Ayan Chakrabarti, Yevgeniy Vorobeychik (2022). **Adversarial Robustness of Deep Sensor Fusion Models**. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
17. Yevgeniy Vorobeychik, **Tong Wu**, Liang Tong (2021). **Systems and methods for defending against physical attacks on image classification**. *US Patent*

Peer-review Service

- International Conference on Learning Representations (ICLR): 2022, 2024, 2025
- Conference on Neural Information Processing Systems (NeurIPS): 2022, 2023, 2024, 2025
- International Conference on Machine Learning (ICML): 2023, 2024, 2025
- Conference on Language Modeling (COLM): 2025
- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 2025
- European Conference on Computer Vision (ECCV): 2024
- IEEE/CVF Winter Conference on Applications of Computer Vision (WACV): 2022, 2024, 2025
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD): 2022
- AAAI Conference on Artificial Intelligence (AAAI): 2021
- IEEE Symposium on Security and Privacy (S&P): 2021
- International Journal of Computer Vision (IJCV): 2021

Honors and Awards

- Research Excellence Award, Washington University in St. Louis, 2021
- AAMAS Student Scholarship, AAMAS 2021
- Washington University Undergraduate Research Conference Travel Award, 2020
- Tau Beta Pi Honor Society, 2020

Teaching Experience

- Information Security, Princeton University, 2024
- Introduction to Machine Learning, Washington University in St. Louis 2019, 2020, 2021