Yizhu Wang

yizhu-wa21@mails.tsinghua.edu.cn | https://yizhu-joy.github.io

Education

Tsinghua University *Beijing, China*
B.Eng. in Electronic Engineering GPA: 3.63 Sep. 2021 – Jun. 2026 (expected)

University of British Columbia *Vancouver, Canada*
Undergraduate Exchange in Computer Science Major GPA: 91 Aug. 2023 – Dec. 2023

Research Interests

My research aims to build secure AI systems. I focus on developing robust defenses for large language models and agent systems, and am broadly interested in AI safety, alignment, and interpretability.

Publications

[SaTML 2026] ***Defending Against Prompt Injection With DataFilter***
Yizhu Wang, Sizhe Chen, Raghad Alkhudair, Basel Alomair, David Wagner.
IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), Munich, Bavaria, Germany, Mar. 2026
[Paper] [Code] [Model]

[AISec @ CCS 2025 | Spotlight] ***Defending Against Prompt Injection with a Few Defensive Tokens***
Sizhe Chen, **Yizhu Wang**, Nicholas Carlini, Chawin Sitawarin, David Wagner.
ACM Workshop on Artificial Intelligence and Security (AISec), Taipei, Taiwan, Oct. 2025
[Paper] [Code]

[SOUPS 2025] ***Can You Walk Me Through It? Explainable SMS Phishing Detection using LLM-based Agents***
Yizhu Wang, Haoyu Zhai, Chenkai Wang, Qingying Hao, Nick A. Cohen, Roopa Foulger, Jonathan A. Handler, Gang Wang.
Proceedings of the 21st Symposium on Usable Privacy and Security (SOUPS), Seattle, WA, USA, Aug. 2025
[Paper] [Code] [Video]

[Under Review at CVPR 2026] ***DREAM: Document Recognition with Explicit Adaptive Memory***
Tianqi Zhao, Di WU, Liangrui Peng, Yifan Huang, Kemeng Zhao, Shuo Li, Zhiyu Li, **Yizhu Wang**, Borui Jiang, Yuyang Li
Under Review at Conference on Computer Vision and Pattern Recognition 2026 (CVPR)

[In Submission] ***Human Decision Model in AI-assisted Phishing Detection***
Yizhu Wang*, Haoyu Zhai*, Nick A. Cohen, Roopa Foulger, Jonathan Handler, and Gang Wang.

Research Experience

Defense Against Prompt Injection Attacks in LLMs *University of California, Berkeley*
Project Lead | Advised by Prof. David Wagner and Ph.D. student Sizhe Chen, UCB Mar. 2025 – Present

- Developed and fine-tuned a model-agnostic **data filter** to automatically remove malicious injections from untrusted third-party data.
- Achieved **state-of-the-art robustness** on securing LLM-based agents while maintaining utility.
- Demonstrated effectiveness across both **black-box** and **open-source** (open-weight) LLMs.

Research Assistant | Advised by Prof. David Wagner, UCB

- Proposed a **deployment-friendly defense** inserting lightweight defensive tokens into model prompts to secure open-weight LLMs against prompt injection attacks.
- Achieved robustness comparable to full fine-tuning while incurring minimal utility loss.

Explainable SMS Phishing Detection with LLM-based Agents *University of Illinois Urbana-Champaign*

Project Lead | Advised by Prof. Gang Wang, UIUC Jul. 2024 – Jun. 2025

- Designed an **agentic AI system** to detect and explain SMS phishing messages for lay users.
- Achieved **98.8% detection accuracy** and reduced hallucinations in generated explanations.
- Led **user studies (N=175)** validating explanation quality and usability.
- Pioneered analysis of human responses to AI errors and disagreement cases.

Invited Talks

How Explainable Phishing Detection Works in Human-AI Collaboration *Aug. 2025*

USENIX SOUPS 2025, Seattle, WA **[Video]**

Awards and Honors

Oct. 2025	**Scholarship**: Outstanding Oversea Study Award (Top 1/11)	*Tsinghua University*
Oct. 2025	**Scholarship**: Outstanding Technological Innovation (<4%)	*Tsinghua University*
Aug. 2025	**Grant**: USENIX Student Travel Grant	*USENIX*
Dec. 2023	**Scholarship**: Zheng Gang Alumni Scholarship (<5%)	*Tsinghua University*
Oct. 2023	**Honors**: Excellent Technology Association Backbone (<5%)	*Tsinghua University*
Nov. 2022	**Scholarship**: Social Work Excellence Award (<4%)	*Tsinghua University*
Dec. 2021	**Scholarship**: Freshman Scholarship (<4%)	*Tsinghua University*

Technical Skills

Programming	Python, C/C++, Matlab, JavaScript, HTML/CSS, Verilog, LaTeX
Frameworks/Tools	PyTorch, DeepSpeed, vLLM, Fine-tuning
Hardware	Cadence, Altium Designer, Single-chip Development
Languages	Mandarin Chinese (native), English (TOEFL 109: R30 L27 S26 W26)

Volunteer Services and Social Work

Teaching Assistant, *Synthetical Practice of Electronic System Design* *Jun. 2023 – Oct. 2023*

Tsinghua University

Vice Minister and Department Member, Hardware Department *Jun. 2022 – Jun. 2024*

Student Science Association, Tsinghua University