Wenxiao Wang

wwx@relai.ai | **Head of AI** at **RELAI, Inc**

Education

University of Maryland, College Park — *College Park, US*

Ph.D. in Computer Science (Dissertation: Towards Reliable Agentic LLMs) — *Sept. 2021 - 2025*

Advisor: Prof. Soheil Feizi

Institute for Interdisciplinary Information Sciences (a.k.a. Yao class), Tsinghua University — *Peking, China*

B.Eng. in Computer Science and Technology. — *Sept. 2016 - June. 2020*

Experience

RELAI, Inc — *Maryland, US*

Head of AI — *July 2025 - Present*

Sony AI — *Remote, US*

research intern mentored by **Weiming Zhuang** and **Lingjuan Lyu** — *May. 2023 - Aug. 2023*

Bytedance — *Remote, US*

research intern mentored by **Linjie Yang**, **Heng Wang** and **Yu Tian** — *June 2022 - Nov. 2022*

Institute for Interdisciplinary Information Sciences, Tsinghua University — *Peking, China*

research assistant mentored by Prof. **Hang Zhao** — *Sep. 2020 - Aug. 2021*

University of California, Berkeley — *Berkeley, US*

visiting student researcher advised by **Xinyun Chen**, **Ruoxi Jia** and Prof. **Dawn Song** — *Apr. 2019 - Aug. 2019*

Bytedance AI Lab — *Peking, China*

intern in visual search group mentored by **Yi He** and **Lei Li** — *May. 2018 - Nov. 2018*

Preprints

Chain-of-Defensive-Thought: Structured Reasoning Elicits Robustness in Large Language Models against Reference Corruption [url]

Wenxiao Wang, Parsa Hosseini, Soheil Feizi — *2025*

Publications

Gaming Tool Preferences in Agentic LLMs [url]

Kazem Faghih*, **Wenxiao Wang***, Yize Cheng*, Siddhant Bharti, Gaurang Sriramanan, Sriram Balasubramanian, Parsa Hosseini, Soheil Feizi (*equal contribution) — *2025*

Conference on Empirical Methods in Natural Language Processing (EMNLP)

DyePack: Provably Flagging Test Set Contamination in LLMs Using Backdoors [url]

Yize Cheng*, **Wenxiao Wang***, Mazda Moayeri, Soheil Feizi (*equal contribution) *2025*

Conference on Empirical Methods in Natural Language Processing (EMNLP)

Can AI-Generated Text be Reliably Detected? [url]

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, **Wenxiao Wang**, Soheil Feizi *2025*

Transactions on Machine Learning Research (TMLR)

Media Coverage: **[Washington Post] [Wired] [New Scientist] [The Register] [TechSpot] [UMD Science]**

Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks [url]

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, **Wenxiao Wang**, Soheil Feizi *2024*

International Conference on Learning Representations (ICLR)

Media Coverage: **[Wired] [MIT Tech Review] [Bloomberg News] [The Register]**

DRSM: De-Randomized Smoothing on Malware Classifier Providing Certified Robustness [url]

Shoumik Saha, **Wenxiao Wang**, Yigitcan Kaya, Soheil Feizi, Tudor Dumitras *2024*

International Conference on Learning Representations (ICLR)

Temporal Robustness against Data Poisoning [url]

Wenxiao Wang, Soheil Feizi *2023*

Conference on Neural Information Processing Systems (NeurIPS)

Spuriosity Rankings: Sorting Data for Spurious Correlation Robustness [url]

Mazda Moayeri, **Wenxiao Wang**, Sahil Singla, Soheil Feizi *2023*

Conference on Neural Information Processing Systems (NeurIPS)[**spotlight**]

Lethal Dose Conjecture on Data Poisoning [url]

Wenxiao Wang, Alexander Levine, Soheil Feizi *2022*

Conference on Neural Information Processing Systems (NeurIPS)

Improved Certified Defenses against Data Poisoning with (Deterministic) Finite Aggregation [url]

Wenxiao Wang, Alexander Levine, Soheil Feizi *2022*

International Conference on Machine Learning (ICML)

On Feature Decorrelation in Self-Supervised Learning [url]

Tianyu Hua*, **Wenxiao Wang***, Zihui Xue, Sucheng Ren, Yue Wang, Hang Zhao (*equal contribution) *2021*

International Conference on Computer Vision (ICCV)[**oral**]

DPlis: Boosting Utility of Differentially Private Deep Learning via Randomized Smoothing [url]

Wenxiao Wang, Tianhao Wang, Lun Wang, Nanqing Luo, Pan Zhou, Dawn Song, Ruoxi Jia *2021*

Privacy Enhancing Technologies Symposium (PETS)

REFIT: A Unified Watermark Removal Framework For Deep Learning Systems With Limited Data [url]

Xinyun Chen*, **Wenxiao Wang***, Yiming Ding, Chris Bender, Ruoxi Jia, Bo Li, Dawn Song (*equal contribution) *2021*

ACM Asia Conference on Computer and Communications Security (AsiaCCS)

The Secret Revealer: Generative Model Inversion Attacks Against Deep Neural Networks [url]

Yuheng Zhang*, Ruoxi Jia*, Hengzhi Pei, **Wenxiao Wang**, Bo Li, Dawn Song (*equal contribution) *2020*

Conference on Computer Vision and Pattern Recognition (CVPR)[**oral**]

Leveraging Unlabeled Data for Watermark Removal of Deep Neural Networks [url]

Xinyun Chen*, **Wenxiao Wang***, Yiming Ding, Chris Bender, Ruoxi Jia, Bo Li, Dawn Song (*equal contribution) *2019*

ICML2019 Workshop on Security and Privacy of Machine Learning

Talks

- **Temporal Robustness against Data Poisoning**, AI TIME Youth PhD Talk, November 2023.
- **Lethal Dose Conjecture: From Few-shot Learning to Potentially Nearly Optimal Defenses against Data Poisoning**, TMLR Group, Hong Kong Baptist University, December 2022.
- **Lethal Dose Conjecture on Data Poisoning**, AI TIME Youth PhD Talk, November 2022.
- **Improved Certified Defenses against Data Poisoning with (Deterministic) Finite Aggregation**, AI TIME Youth PhD Talk, August 2022.

Services

Program Committee / Reviewer of: NeurIPS, ICML, ICLR, ICCV, CVPR, TPAMI, TMLR, ...

Awards

Gold Medal(**4th place**)	National Olympiad in Informatics	2015
Gold Medal(**1st place**)	Asia and Pacific Informatics Olympiad in China District	2015
Gold Medal(**10th place**)	China Team Selection Competition	2015
Gold Medal	National Olympiad in Informatics	2014
Bronze Medal	Asia and Pacific Informatics Olympiad in China District	2014
Silver Medal	China Team Selection Competition	2014

Teaching

- Teaching Assistant of *CMSC828W: Foundations of Deep Learning*, Fall 2022, University of Maryland, College Park.
- Teaching Assistant of *CMSC422: Introduction to Machine Learning*, Spring 2022, University of Maryland, College Park.
- Teaching Assistant of *CMSC351: Algorithms*, Fall 2021, University of Maryland, College Park.