

Jiahao Yu

✉ jiahaoyu04@gmail.com • [sherdencooper.github.io/](https://github.com/sherdencooper)

Education

Northwestern University
Ph.D., Computer Science

Evanston, United States
Jan. 2022 – Now

- Advisor: Prof. Xinyu Xing
- GPA: 3.85/4.0

Shanghai Jiao Tong University
B.S, School of Electronic Information and Electrical Engineering

Shanghai, China
Sep. 2017 – July 2021

- Advisor: Prof. Liyao Xiang
- Zhiyuan Honor Program
- GPA: 3.6/4.0

Professional Appointments

New York University Abu Dhabi
Tenure-Track Assistant Professor (TTAP), Computer Engineering

Abu Dhabi, United Arab Emirates
Starting Aug. 1, 2026

Research Interests

My research is situated at the intersection of AI and security, with a broad focus on leveraging artificial intelligence to advance cybersecurity capabilities. Specifically, I concentrate on using Large Language Models (LLMs) and Reinforcement Learning (RL) for critical security tasks, including automated program patch generation, vulnerability identification, and securing blockchain systems. From a machine learning perspective, my work aims to improve LLMs' explainability, adversarial robustness, noise resistance, and overall effectiveness of these models and algorithms.

Published Work

*Denotes equal contribution.

- 2026.....
- Decoupled Alignment for Robust Plug-and-Play Adaptation
 - Haozheng Luo*, Jiahao Yu*, Wenxin Zhang, Jialong Li, Chenghao Qiu, Yimin Wang, Eric Hanchen Jiang, Jerry Yao-Chieh Hu, Yan Chen, Binghui Wang, Xinyu Xing, Han Liu
 - In Proceedings of the Conference on Language Modeling (COLM) 2026
 - Contrastive Reasoning Alignment: Reinforcement Learning from Hidden Representations
 - Haozheng Luo, Yimin Wang, Jiahao Yu, Binhui Wang, Yan Chen
 - In Proceedings of the International Conference on Machine Learning (ICML) 2026
 - Locus: Agentic Predicate Reasoning for Directed Fuzzing
 - Jie Zhu, Chihao Shen, Ziyang Li, Jiahao Yu, Yizheng Chen, Kexin Pei
 - In Proceedings of the International Conference on Software Engineering (ICSE) 2026
 - PROMPTFUZZ: Harnessing Fuzzing Techniques for Robust Testing of Prompt Injection in LLMs
 - Jiahao Yu*, Yangguang Shao*, Hanwen Miao, Junzheng Shi
 - Transactions on Information Forensics & Security
- 2025.....
- Mind the Inconspicuous: Revealing the Hidden Weakness in Aligned LLMs' Ethical Boundaries
 - Jiahao Yu, Haozheng Luo, Jerry Yao-Chieh Hu, Wenbo Guo, Yan Chen, Han Liu, Xinyu Xing
 - In Proceedings of the USENIX Security Symposium (USENIX Security) 2025 (Long Talk)
 - PatchAgent: A Practical Program Repair Agent Mimicking Human Expertise
 - Zheng Yu, Ziyi Guo, Yuhang Wu, Jiahao Yu, Meng Xu, Dongliang Mu, Yan Chen, Xinyu Xing
 - In Proceedings of the USENIX Security Symposium (USENIX Security) 2025 (Long Talk)
 - This work was the foundation of the auto patch generation for our winning solution (Team 42-b3yond-bug)

in the DARPA AI Cyber Challenge (AIXCC), securing a **\$1,000,000 grant for proposal** and **\$2,000,000 prize in semi-final**.

- This work was selected as Runner-up in Technical Impact in **CSAW 2025**.

The Illusion of Role Separation: Hidden Shortcuts in LLM Role Learning (and How to Fix Them)

- Zihao Wang, Yibo Jiang, Jiahao Yu, Heqing Huang
- In Proceedings of the International Conference on Machine Learning (ICML) 2025

GPO: Learning from Critical Steps to Improve LLM Reasoning

- Jiahao Yu, Zelei Cheng, Xian Wu, Xinyu Xing
- In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) 2025
- This work was featured in **MIT Technology Review China a** and **MIT Technology Review China b** (October 2025).

BlockScan: Detecting Anomalies in Blockchain Transactions

- Jiahao Yu, Xian Wu, Hao Liu, Wenbo Guo, Xinyu Xing
- In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) 2025

Knowledge-Distilled Memory Editing for Plug-and-Play LLM Alignment

- Haozheng Luo*, Jiahao Yu*, Wenxin Zhang*, Jialong Li, Jerry Yao-Chieh Hu, Yan Chen, Binhui Wang, Xinyu Xing, Han Liu
- ICML 2025 Workshop on the Impact of Memorization on Trustworthy Foundation Models (ICLM@MemFM)

UTF: Undertrained Tokens as Fingerprints A Novel Approach to LLM Identification

- Jiacheng Cai*, Jiahao Yu*, Yangguang Shao, Yuhang Wu, Xinyu Xing
- ACL 2025 Workshop on LLMsec (ACL@LLMsec)

2024.....

Soft-Label Integration for Robust Toxicity Classification

- Zelei Cheng*, Xian Wu*, Jiahao Yu*, Shuo Han, Xin-Qiang Cai, Xinyu Xing
- In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) 2024
- This work was featured in **MIT Technology Review China** (December 2023).

LLM-Fuzzer: Scaling Assessment of Large Language Model Jailbreaks

- Jiahao Yu, Xingwei Lin, Zheng Yu, Xinyu Xing
- In Proceedings of the USENIX Security Symposium (USENIX Security) 2024
- Originally released as GPTFuzzer, this work received the **Geekcon 2023 Annual Themed Debate Break-through Award**.
- The tool has been widely adopted by major LLM providers including OpenAI, ByteDance, Ant Group, ScaleAI and Anthropic, and is integrated into Microsoft Azure's PyRIT.
- The associated open-source tool has been downloaded over 146,000 times.

RICE: Breaking Through the Training Bottlenecks of Reinforcement Learning with Explanation

- Zelei Cheng*, Xian Wu*, Jiahao Yu*, Sabrina Yang, Gang Wang, Xinyu Xing
- In Proceedings of the International Conference on Machine Learning (**ICML Spotlight**) 2024

BandFuzz: A Practical Framework for Collaborative Fuzzing with Reinforcement Learning

- Wenxuan Shi, Hongwei Li, Jiahao Yu, and Wenbo Guo, Xinyu Xing
- ICSE Workshop on Search-Based and Fuzz Testing (ICSE@SBFT) 2024
- This work won the First Prize at the SBFT 2024 Fuzzing Competition, ranking 1st across all evaluation metrics.
- It was featured in **McCormick School of Engineering News**.
- It served as the foundational fuzzing component for our winning solution in the DARPA AI Cyber Challenge (AIXCC) semi-final.
- In the AIXCC final, this framework discovered **the highest number of zero-day vulnerabilities among all competitors**.

Assessing Prompt Injection Risks in 200+ Custom GPTs

- Jiahao Yu, Yuhang Wu, Dong Shu, Mingyu Jin, Xinyu Xing
- ICLR 2024 Workshop on Secure and Trustworthy Large Language Models (ICLR@SeTLLM)

- This research was featured in **WIRED** (November 2023).

2023.....

StateMask: Explaining Deep Reinforcement Learning through State Mask

- Zelei Cheng*, Xian Wu*, Jiahao Yu*, Wenbo Guo, Wenhai Sun, Xinyu Xing
- In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) 2023

AIRS: Explanation for Deep Reinforcement Learning based Security Applications

- Jiahao Yu, Wenbo Guo, Qi Qin, Gang Wang, Ting Wang, Xinyu Xing
- In Proceedings of the USENIX Security Symposium (USENIX Security) 2023

Publications Prior to Ph.D. Program

2021.....

Matrix Gaussian Mechanism for Differentially-Private Learning

- Jungang Yang, Liyao Xiang, Jiahao Yu, Xinbing Wang, Bin Guo, Bin Guo, Zhetao Li, Baochun Li
- In IEEE Transactions on Mobile Computing (TMC) 2021

Speedup robust graph structure learning with low-rank information

- Hui Xu, Liyao Xiang, Jiahao Yu, Anqi Cao, Xinbing Wang
- In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM) 2021

2020.....

Voiceprint Mimicry Attack Towards Speaker Verification System in Smart Home

- Lei Zhang, Yan Meng, Jiahao Yu, Chong Xiang, Brandon Folk, Haojin Zhu
- In Proceedings of the IEEE International Conference on Computer Communications (INFOCOM) 2020

Research on Application of Artificial Intelligence Technology in Electrical Automation Control

- Chao Jiang, Xiaorui Xiong, Tanqing Zhu, Jiajia Cao, Jiahao Yu
- Journal of Physics: Conference Series, Vol. 1578, 2020

Unpublished Manuscripts & Preprints

2026.....

Building Coding Agents via Entropy-Enhanced Multi-Turn Preference Optimization

- Jiahao Yu, Zelei Cheng, Xian Wu, Xinyu Xing
- arXiv preprint arXiv:2509.12434
- It ranks **1st in SWEBench-Lite** and **5th in SWEBench-Verified**(only suppressed by models 10x larger than us) official open-weight leaderboard.

2025.....

A Survey on Explainable Deep Reinforcement Learning

- Zelei Cheng*, Jiahao Yu*, Xinyu Xing
- arXiv preprint arXiv:2502.06869

BandFuzz: An ML-powered Collaborative Fuzzing Framework

- Wenxuan Shi, Hongwei Li, Jiahao Yu, Xinqian Sun, Wenbo Guo, Xinyu Xing
- arXiv preprint arXiv:2507.10845

GenoArmory: A Unified Evaluation Framework for Adversarial Attacks on Genomic Foundation Models

- Haozheng Luo, Chenghao Qiu, Yimin Wang, Shang Wu, Jiahao Yu, Han Liu, Binghui Wang, Yan Chen
- <https://github.com/MAGICS-LAB/GenoArmory>

POISONCRAFT: Practical Poisoning of Retrieval-Augmented Generation for Large Language Models

- Yangguang Shao, Xinjie Lin, Haozheng Luo, Chengshang Hou, Gang Xiong, Jiahao Yu, Junzheng Shi
- arXiv preprint arXiv:2505.06579

2023.....
 GPTFUZZER : Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts
 • Jiahao Yu, Xingwei Lin, Xinyu Xing
 • arXiv preprint arXiv:2309.10253

Professional Service

- Workshop Organizer
 - **The 2nd Workshop on Efficient Reasoning at COLM** 2026
- Program Committee Member
 - USENIX Security Symposium ('27 Cycle 1)
 - USENIX Security Symposium ('26 Cycle 1)
 - USENIX Security Symposium ('26 Cycle 2)
 - Conference on Computer and Communications Security (CCS) 2026 (Track A)
 - Conference on Computer and Communications Security (CCS) 2026 (Track B)
 - LLM4Code@ICSE 2026
- Area Chair
 - ES-Reasoning@ICLR 2026
- Reviewer
 - Conference on Language Modeling (COLM) 2026
 - AAAI Conference on Artificial Intelligence (AAAI) 2026
 - IEEE Transactions on Software Engineering (TSE) 2026
 - International Conference on Learning Representations (ICLR) 2025 (Notable Reviewer), 2026
 - IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2026
 - Conference on Neural Information Processing Systems (NeurIPS) 2024, 2025
 - International Conference on Machine Learning (ICML) 2025
 - International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, 2026
 - The Web Conference (WWW) 2025
 - ICLR Workshop on Secure and Trustworthy Large Language Models 2024
 - IEEE Transactions on Information Forensics and Security (TIFS) 2025
 - IEEE Transactions on Knowledge and Data Engineering (TKDE) 2024
 - Journal of Orthopaedic Surgery 2025
- Artifact Evaluation Committee Member
 - USENIX Security Symposium 2025 (Cycle 1, Cycle 2), 2026 (Cycle 1, Cycle 2)
 - Network and Distributed System Security Symposium (NDSS) 2025 (Summer, Fall)
- External Reviewer
 - ACM Conference on Computer and Communications Security (CCS) 2023 (Track B), 2022 (Track B)
 - IEEE Symposium on Security and Privacy (S&P) 2025 (Cycle 2), 2023 (Cycle 3), 2022 (Poster)
 - International Conference on Automated Software Engineering (ASE) 2024
 - International Conference on Learning Representations (ICLR) 2024
 - International Symposium on Research in Attacks, Intrusions and Defenses (RAID) 2022
 - Network and Distributed System Security Symposium (NDSS) 2025
 - USENIX Workshop on Offensive Technologies (WOOT) 2024

Awards

- ICLR Notable Reviewer 2025
- CSAW Runner-up in Technical Impact 2025
- DARPA AIxCC Semi-Finalist (Team 42-b3yond-bug) 2024
- OpenAI Researcher Access Program 2024
- First Prize, SBFT Fuzzing Competition 2024

o Geekcon Breakthrough Award 2023

Working Experiences

LLM Red-teaming Benchmarking
ByteDance

San Francisco, USA
Sep. 2024 – Dec. 2024

- Research Intern
- Advisor: Dr. Zhenqing Luo

LLM for Vulnerability Detection
University of Chicago

Chicago, USA
June. 2024 – Sep. 2024

- Research Assistant
- Advisor: Dr. Kexin Pei

Automatically Red-teaming Large Language Models
Ant Group

Hangzhou, China
June. 2023 – Aug. 2023

- Research Intern
- Advisor: Xingwei Lin

Adversarial Attack against PowerShell Malware Detector
MSRA

Beijing, China
Aug. 2020 – Mar. 2021

- Research Intern
- Advisor: Dr. Bin Zhu