

Siyuan Cheng

✉ cheng535@purdue.edu

🎓 Google Scholar

🌐 Homepage

🌐 LinkedIn

🐦 @Megum1



Education

- 8/2021 – Present **Purdue University**, West Lafayette, IN, USA
Ph.D. Candidate, Computer Science
Advised by Samuel Conte Professor *Xiangyu Zhang*
- 9/2016 – 7/2020 **Shanghai Jiao Tong University (SJTU)**, Shanghai, China
Bachelor, Computer Science and Technology (Artificial Intelligence)
Affiliated with the *IEEE Honor Class*

Work Experience

- 8/2021 – Present **Purdue University**, West Lafayette, IN, USA
Research Assistant, Computer Science
Mentored by Samuel Conte Professor *Xiangyu Zhang*
- 5/2024 – 11/2024 **Sony AI**, New York, USA
Research Intern, Privacy-Preserving Machine Learning (PPML) Lab
Mentored by *Lingjuan Lyu* and *Vikash Sehwal*
- 10/2020 – 4/2021 **Hitachi**, Shanghai, China
Applied Research Intern

Publications (* denotes equal contribution)

Conference and Journal

- ☐ **CO-SPY: Combining Semantic and Pixel Features to Detect Synthetic Images by AI**
Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, Vikash Sehwal
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025
- ☐ **ODSCAN: Backdoor Scanning for Object Detection Models**
Siyuan Cheng*, Guangyu Shen*, Guanhong Tao, Kaiyuan Zhang, Zhuo Zhang, Shengwei An, Xiangzhe Xu, Yingqi Liu, Shiqing Ma, Xiangyu Zhang
IEEE Symposium on Security and Privacy (S&P) 2024
- ☐ **LOTUS: Evasive and Resilient Backdoor Attacks through Sub-Partitioning**
Siyuan Cheng, Guanhong Tao, Yingqi Liu, Guangyu Shen, Shengwei An, Shiwei Feng, Xiangzhe Xu, Kaiyuan Zhang, Shiqing Ma, Xiangyu Zhang
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024
- ☐ **UNIT: Backdoor Mitigation via Automated Neural Distribution Tightening**
Siyuan Cheng*, Guangyu Shen*, Kaiyuan Zhang, Guanhong Tao, Shengwei An, Hanxi Guo, Shiqing Ma, Xiangyu Zhang
European Conference on Computer Vision (ECCV) 2024

Publications (* denotes equal contribution) (continued)

- **BEAGLE: Forensics of Deep Learning Backdoor Attack for Better Defense**
Siyuan Cheng, Guanhong Tao, Yingqi Liu, Shengwei An, Xiangzhe Xu, Shiwei Feng, Guangyu Shen, Kaiyuan Zhang, Qiuling Xu, Shiqing Ma, Xiangyu Zhang
Network and Distributed System Security Symposium (NDSS) 2023
- **Deep Feature Space Trojan Attack of Neural Networks by Controlled Detoxification**
Siyuan Cheng, Yingqi Liu, Shiqing Ma, Xiangyu Zhang
AAAI Conference on Artificial Intelligence (AAAI) 2021
- **BAIT: Large Language Model Backdoor Scanning by Inverting Attack Target**
Guangyu Shen*, Siyuan Cheng*, Zhuo Zhang, Guanhong Tao, Kaiyuan Zhang, Hanxi Guo, Lu Yan, Xiaolong Jin, Shengwei An, Shiqing Ma, Xiangyu Zhang
IEEE Symposium on Security and Privacy (S&P) 2025
- **Exploring the Orthogonality and Linearity of Backdoor Attacks**
Kaiyuan Zhang*, Siyuan Cheng*, Guangyu Shen, Guanhong Tao, Shengwei An, Anuran Makur, Shiqing Ma, Xiangyu Zhang
IEEE Symposium on Security and Privacy (S&P) 2024
- **Django: Detecting Trojans in Object Detection Models via Gaussian Focus Calibration**
Guangyu Shen*, Siyuan Cheng*, Guanhong Tao, Kaiyuan Zhang, Yingqi Liu, Shengwei An, Shiqing Ma, Xiangyu Zhang
Advances in Neural Information Processing Systems (NeurIPS) 2023
- **SOFT: Selective Data Obfuscation for Protecting LLM Fine-tuning against Membership Inference Attacks**
Kaiyuan Zhang, Siyuan Cheng, Hanxi Guo, Yuetian Chen, Zian Su, Shengwei An, Yuntao Du, Charles Fleming, Ashish Kundu, Xiangyu Zhang, Ninghui Li
USENIX Security Symposium (USENIX Security) 2025
- **CENSOR: Defense Against Gradient Inversion via Orthogonal Subspace Bayesian Sampling**
Kaiyuan Zhang, Siyuan Cheng, Guangyu Shen, Bruno Ribeiro, Shengwei An, Pin-Yu Chen, Xiangyu Zhang, Ninghui Li
Network and Distributed System Security Symposium (NDSS) 2025
- **Unleashing the Power of Generative Model in Recovering Variable Names from Stripped Binary**
Xiangzhe Xu, Zhuo Zhang, Zian Su, Ziyang Huang, Shiwei Feng, Yapeng Ye, Nan Jiang, Danning Xie, Siyuan Cheng, Lin Tan, Xiangyu Zhang
Network and Distributed System Security Symposium (NDSS) 2025
- **System Prompt Hijacking via Permutation Triggers in LLM Supply Chains**
Lu Yan, Siyuan Cheng, Xuan Chen, Kaiyuan Zhang, Guangyu Shen, Xiangyu Zhang
Findings of the Association for Computational Linguistics (ACL) 2025
- **JailbreakDiffBench: A Comprehensive Benchmark for Jailbreaking Diffusion Models**
Xiaolong Jin, Zixuan Weng, Hanxi Guo, Chenlong Yin, Siyuan Cheng, Guangyu Shen, Xiangyu Zhang
The International Conference on Computer Vision (ICCV) 2025
- **EffiTune: Diagnosing and Mitigating Training Inefficiency for Parameter Tuner in Robot Navigation System**
Shiwei Feng, Xuan Chen, Zhiyuan Cheng, Zikang Xiong, Yifei Gao, Siyuan Cheng, Sayali Kate, Xiangyu Zhang
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2025
- **On Large Language Models' Resilience to Coercive Interrogation**
Zhuo Zhang, Guangyu Shen, Guanhong Tao, Siyuan Cheng, Xiangyu Zhang
IEEE Symposium on Security and Privacy (S&P) 2024

Publications (* denotes equal contribution) (continued)

- **Biscope: Ai-generated text detection by checking memorization of preceding tokens**
Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Kaiyuan Zhang, Guanhong Tao, Guangyu Shen, Xiangyu Zhang
Advances in Neural Information Processing Systems (NeurIPS) 2024
- **Rethinking the Invisible Protection against Unauthorized Image Usage in Stable Diffusion**
Shengwei An*, Lu Yan*, Siyuan Cheng, Guangyu Shen, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, Xiangyu Zhang
USENIX Security Symposium (USENIX Security) 2024
- **Backdoor Attacks without Poisoning**
Guanhong Tao, Siyuan Cheng, Zhenting Wang, Shiqing Ma, Shengwei An, Yingqi Liu, Guangyu Shen, Zhuo Zhang, Yunshu Mao, Xiangyu Zhang
Annual Computer Security Applications Conference (ACSAC) 2024
- **Elijah: Eliminating Backdoors Injected in Diffusion Models via Distribution Shift**
Shengwei An, Sheng-Yen Chou, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, Guangyu Shen, Siyuan Cheng, Shiqing Ma, Pin-Yu Chen, Tsung-Yi Ho, Xiangyu Zhang
AAAI Conference on Artificial Intelligence (AAAI) 2024
- **ROCAS: Root Cause Analysis of Autonomous Driving Accidents via Cyber-Physical Co-mutation**
Shiwei Feng, Yapeng Ye, Qingkai Shi, Zhiyuan Cheng, Xiangzhe Xu, Siyuan Cheng, Hongjun Choi, Xiangyu Zhang
IEEE/ACM International Conference on Automated Software Engineering (ASE) 2024
- **Hard-label Black-box Universal Adversarial Patch Attack**
Guanhong Tao, Shengwei An, Siyuan Cheng, Guangyu Shen, Xiangyu Zhang
USENIX Security Symposium (USENIX Security) 2023
- **ImU: Physical Impersonating Attack for Face Recognition System with Natural Style Changes**
Shengwei An, Yuan Yao, Qiuling Xu, Shiqing Ma, Guanhong Tao, Siyuan Cheng, Kaiyuan Zhang, Yingqi Liu, Guangyu Shen, Ian Kelk, Xiangyu Zhang
IEEE Symposium on Security and Privacy (S&P) 2023
- **FLIP: A Provable Defense Framework for Backdoor Mitigation in Federated Learning**
Kaiyuan Zhang, Guanhong Tao, Qiuling Xu, Siyuan Cheng, Shengwei An, Yingqi Liu, Shiwei Feng, Guangyu Shen, Pin-Yu Chen, Shiqing Ma, Xiangyu Zhang
The International Conference on Learning Representations (ICLR) 2023
- **Detecting Backdoors in Pre-trained Encoders**
Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, Xiangyu Zhang
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2023
- **MEDIC: Remove Model Backdoors via Importance Driven Cloning**
Qiuling Xu, Guanhong Tao, Jean Honorio, Yingqi Liu, Shengwei An, Guangyu Shen, Siyuan Cheng, Xiangyu Zhang
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2023
- **PEM: Representing Binary Program Semantics for Similarity Analysis via A Probabilistic Execution Model**
Xiangzhe Xu*, Zhou Xuan*, Shiwei Feng, Siyuan Cheng, Yapeng Ye, Qingkai Shi, Guanhong Tao, Le Yu, Zhuo Zhang, Xiangyu Zhang
ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE) 2023

Publications (* denotes equal contribution) (continued)

- **Improving Binary Code Similarity Transformer Models by Semantics-driven Instruction Deemphasis**
Xiangzhe Xu, Shiwei Feng, Yapeng Ye, Guangyu Shen, Zian Su, Siyuan Cheng, Guanhong Tao, Qingkai Shi, Zhuo Zhang, Xiangyu Zhang
ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA) 2023
- **Towards Feature Space Adversarial Attack by Style Perturbation**
Qiuling Xu, Guanhong Tao, Siyuan Cheng, Xiangyu Zhang
AAAI Conference on Artificial Intelligence (AAAI) 2021
- **Backdoor scanning for deep neural networks through k-arm optimization**
Guangyu Shen*, Yingqi Liu*, Guanhong Tao, Shengwei An, Qiuling Xu, Siyuan Cheng, Shiqing Ma, Xiangyu Zhang
The International Conference on Machine Learning (ICML) 2021
- **Rational Manager in Bitcoin Mining Pool: Dynamic Strategies to Gain Extra Rewards**
Feifan Yu, Na Ruan, Siyuan Cheng
ACM Asia Conference on Computer and Communications Security (AsiaCCS) 2020

Peer-reviewed Workshop



- **A Systematic Threat Modeling of LLM Applications**
Guanhong Tao*, Siyuan Cheng*, Zhuo Zhang, Junmin Zhu, Guangyu Shen, Wanqing Han, Mu Zhang, Xiangyu Zhang
FSE 2025 Workshop on LLM App Store Analysis (LLMapp)
- **SkewAct: Red Teaming Large Language Models via Activation-Skewed Adversarial Prompt Optimization**
Hanxi Guo, Siyuan Cheng, Guanhong Tao, Guangyu Shen, Zhuo Zhang, Shengwei An, Kaiyuan Zhang, Xiangyu Zhang
NeurIPS 2024 Workshop on Red Teaming GenAI
- **MultiVerse: Exposing Large Language Model Alignment Problems in Diverse Worlds**
Xiaolong Jin, Zhuo Zhang, Guangyu Shen, Hanxi Guo, Kaiyuan Zhang, Siyuan Cheng, Xiangyu Zhang
NeurIPS 2024 Workshop on Safe Generative AI
- **D³: Detoxing Deep Learning Dataset**
Lu Yan, Siyuan Cheng, Guangyu Shen, Guanhong Tao, Kaiyuan Zhang, Yunshu Mao, Xiangyu Zhang
NeurIPS 2023 Workshop on Backdoors in Deep Learning - The Good, the Bad, and the Ugly

Guest Lectures




- Fall 2024 ■ **Safety Alignment of Large Language Models**
CS 59200: AI And Security at *Purdue University*, invited by Prof. *Xiangyu Zhang*.
- Fall 2023 ■ **Advanced Backdoor Attack and Defense in Machine Learning Models**
COMPSCI 360: Introduction to Computer and Network Security at *University of Massachusetts, Amherst*, invited by Prof. *Shiqing Ma*.

Professional Services



Conference Reviewer

- 2025  NeurIPS, ICML, ICLR, CVPR, ICCV, WACV, ACL
- 2024  NeurIPS, CVPR, ACL, EMNLP

Conference External Reviewer

- 2025  OOPSLA, USENIX Security
- 2024  CCS, ISSTA
- 2023  CCS, ICSE, ASE, USENIX Security

Journal Reviewer

- 2025  TDSC, TIFS, TPAMI
- 2024  TDSC, TIFS, TPAMI, TMLR, JETCAS