Xuan Chen

✉ chen4124@purdue.edu ☎ (765)-413-2541 🌐 xuanchen.me

Education

Ph.D. in Computer Science, Purdue University 08/2021 – 05/2026 (expected)
Advisor: Prof. Xiangyu Zhang
Research Focus: AI security, specializing in LLMs and RL for security

M.S. in Electrical and Computer Engineering, Carnegie Mellon University 08/2019 – 12/2020
Research Focus: Explainable RL

B.E. in Automation, University of Science and Technology Beijing 08/2015 – 06/2019
Outstanding Graduates (top 5%)

Technical Skills

Frameworks & Libraries: PyTorch, Google ADK, Langchain, QLoRA, DeepSpeed, TensorFlow, Transformers

Programming Languages: Python, C/C++, Java, Bash

Machine Learning Skills: LLMs (red-teaming, alignment, jailbreak detection, multi-modal); Reinforcement Learning (Stable Baselines3, RLlib, PyTorch RL); Robustness & Safety (adversarial training, backdoor detection)

Selected Publications

NeurIPS 2025 **Temporal Logic-Based Multi-Vehicle Backdoor Attacks against Offline RL Agents in End-to-end Autonomous Driving**

Xuan Chen, *Shiwei Feng, Zikang Xiong, Shengwei An, Yunshu Mao, Lu Yan, Guanhong Tao, Wenbo Guo, Xiangyu Zhang*

Proposed a novel trajectory-level multi-vehicle backdoor attack against end-to-end autonomous driving. Developed a flexible, temporal logic-based framework to automatically generate and evaluate trigger trajectories. Designed efficient patch trigger generation in the training pipeline to reduces false trigger rates.

NeurIPS 2025 **TAI3: Testing Agent Integrity in Interpreting User Intent**

Shiwei Feng, Xiangzhe Xu, ***Xuan Chen****, Kaiyuan Zhang, Syed Yusuf Ahmed, Zian Su, Mingwei Zheng, Xiangyu Zhang*

Introduced the first API-centric stress-testing framework to systematically uncover intent integrity violations in six LLM agents. Automatically generated realistic tasks from toolkit documentation and applied targeted mutations to effectively expose subtle agent errors while preserving user intent.

ACL 2025 **ASPIRER: Bypassing System Prompts with Permutation-based Backdoors in LLMs**

Lu Yan, Siyuan Cheng, ***Xuan Chen****, Kaiyuan Zhang, Guangyu Shen, Zhuo Zhang, Xiangyu Zhang*

Introduced the first systematic method for bypassing system prompts in LLMs by permutation triggers, which activate only under specific component orderings. Developed a stealthy and adaptive attack strategy resilient to unforeseen user prompts. Demonstrated robust performance achieving up to 100% ASR and CACC.

NeurIPS 2024 **When LLM Meets DRL: Advancing Jailbreaking Efficiency via DRL-guided Search**

Xuan Chen, *Yuzhou Nie, Wenbo Guo, Xiangyu Zhang*

Proposed the first RL-driven black-box jailbreaking attack against LLMs, introducing a novel, low-cost reward space and a customized PPO algorithm. Achieved a 3.5× improvement in red-teaming success rate on `Llama2-70B-instruct` and `GPT-4`, outperforming state-of-the-art baselines.

NeurIPS 2023 **BIRD: Generalizable Backdoor Detection and Removal for DRL**

***Xuan Chen**, Wenbo Guo, Guanhong Tao, Xiangyu Zhang, Dawn Song*

Developed the first method to robustly detect and remove backdoors in pretrained DRL policies without requiring prior knowledge of attack specifications. Formulated trigger restoration as an RL problem and proposed a novel metric to reliably identify compromised policies. Designed a finetuning strategy that removes backdoors while preserving agent performance in clean environments across ten RL environments.

NeurIPS 2023 **ParaFuzz: An Interpretability-Driven Technique for Detecting Poisoned Samples in NLP**

*Lu Yan, Zhuo Zhang, Guanhong Tao, Kaiyuan Zhang, **Xuan Chen** Guangyu Shen, Xiangyu Zhang*

Proposed ParaFuzz, a framework that detects poisoned samples by leveraging ChatGPT paraphrasing to remove triggers. Optimized paraphrasing prompts via fuzzing, introducing sentence coverage and three novel mutation strategies. Achieved a 90.1% F1 score, more than double that of most baselines.

Professional Experience

Research Intern, Google 09/2025 – Present

Team: Agentspace

- Developed a configurable agent-based user simulator using Google ADK, capable of actively probing vulnerabilities in multi-turn agents across different dimensions, enabling more robust and comprehensive evaluation.

Applied Scientist Intern, Amazon 05/2025 – 08/2025

Team: Buyer Risk Prevention

- Designed and implemented the first memory-augmented, multi-modal LLM agent-based framework that dynamically orchestrates five specialized tools and supports episodic memory construction, retrieval, and updating, achieving 96.6% ACC and 91.4% recall (+20% vs. SOTA) with 31% lower latency.

Applied Scientist Intern, Amazon 05/2024 – 08/2024

Team: Buyer Risk Prevention

- Developed a diffusion model-based account takeover library to mitigate catastrophic forgetting, integrating contrastive loss and a tabular transformer encoder to achieve 93.52% AUC and 78.85% Recall@5.
- Partnered with fraud analysis teams to deploy the method into a scalable library, strengthening Amazon's fraud detection framework against dormant re-emerging attack patterns.

Awards and Honors

Amazon Nova AI Challenge, Attack Team Winner, 2025

Women in Science Travel Grant, Purdue University, 2024

Scholar Award, NeurIPS, 2023

Ross Fellowship, Purdue University, 2021

First Class Scholarship, USTB, 2016, 2017, 2018

National Scholarship, Ministry of Education of China (Top 2%), 2015

Academic Services

Reviewer: AISTATS 2026; ICLR 2026; NeurIPS 2025; ICLR 2025; AISTATS 2025; ICML 2025;

Teaching Assistant: CS 529 Security Analysis (Fall 2023), Purdue University

Teaching Assistant: 18661 Introduction to Machine Learning (Spring 2020), Carnegie Mellon University