Chengquan Guo

✉ chengquanguo@zju.edu.cn https://github.com/1mocat

Education

Zhejiang University
B.Eng. in Information Security. GPA: 3.97/4 — Sept. 2021 – Jun. 2025

University of Oxford
Participant in Machine Learning Summer Program — Jul. 2023 – Aug. 2023

University of Chicago
Research Intern, focused on AI safety and security under the guidance of: Prof. Bo Li — Feb. 2024 – Present

Research Interests

My research interests are centered around **AI safety**, **AI security**, and **Trustworthy Machine Learning**. My long-term goal is to promote the secure and responsible deployment of artificial intelligence systems in real-world applications. During my internship, I have focused primarily on the domain of code agents & LLM. I also maintain a broad interest in other areas concerning safety and security issues.

Publications

RedCode: Risky Code Execution and Generation Benchmark for Code Agents
Chengquan Guo*, Xun Liu*, Chulin Xie*, Andy Zhou, Yi Zeng, Zinan Lin, Dawn Song, Bo Li
Neurips 2024

RedCodeAgent: Automatic Red-teaming Agent against Code Agents
Chengquan Guo, Chulin Xie, Yu Yang, Zinan Lin, Bo Li
ICLR 2025, Under Review

MMDT: Decoding the Trustworthiness and Safety of Multimodal Foundation Models
Chejian Xu*, Jiawei Zhang*, Zhaorun Chen*, Chulin Xie*, Mintong Kang*, Zhuowen Yuan*, Zidi Xiong*, Chenhui Zhang, Lingzhi Yuan, Yi Zeng, Peiyang Xu, Chengquan Guo, Andy Zhou, Jeffrey Ziwei Tan, Zhun Wang, Alexander Xiong, Xuandong Zhao, Yu Gai, Francesco Pinto, Yujin Potter, Zhen Xiang, Zinan Lin, Dan Hendrycks, Dawn Song, Bo Li
ICLR 2025, Under Review

Honors & Awards

National Scholarship (top 2%)	2024
Provincial Government Scholarship (top 3%)	2023
First-Class Scholarship in Zhejiang University * 2 (top 3%)	2023 & 2024
Tencent Scholarship	2023
Second-Class Scholarship in Zhejiang University (top 8%)	2022
Outstanding Student * 3	2022 & 2023 & 2024
Pioneer of External Communication * 2	2023 & 2024
Pioneer of Public Service	2022

Research Experience

Red-teaming Agent against Code Agents — Jun. 2024 – Oct. 2024
Advisor: Prof. Bo Li (University of Chicago)

- Propose RedCodeAgent, the first fully automated and adaptive red-teaming agent against code agents, adapting dynamically to developments in red-teaming tools and code agents.
- Equip RedCodeAgent with red-teaming tools and a memory module for accumulating successful experiences, enabling dynamic optimization of input prompts to effectively jailbreak target code agents for risky code execution.
- Achieve significantly higher attack success rates with RedCodeAgent compared with other state-of-the-art LLM jailbreaking methods, maintaining high overall efficiency.

Safety Benchmark for Code Agents — Feb. 2024 – Jun. 2024
Advisor: Prof. Bo Li (University of Chicago)

- Identify an underexplored area of AI safety and propose the risky code execution (RedCode-Exec) and generation (RedCode-Gen) benchmarks for code agents.

- Develop RedCode-Exec with 25 risky scenarios and over 4,000 test cases; RedCode-Gen with 8 categories of malware and 160 prompts. Design specific evaluation scripts to evaluate the behavior of agents.
- Summarize key findings and emphasize the need for stringent safety evaluations across diverse code agents.

Trustworthiness and Safety of Multimodal Foundation models Feb. 2024 – Jun. 2024
Advisor: Prof. Bo Li (University of Chicago)

- Conduct privacy assessments on 14 multimodal foundation models, including GPT-4v, LLaVA, Qwen etc.
- Investigate the potential risks these models pose to geo-privacy through their inference capabilities.
- Find that existing multimodal foundation models barely refuse to predict sensitive locations, suggesting that they are unaware of location privacy risks, potentially leading to misuse

Backdoor Attack and Defense & LLM & Agent Safety Dec. 2023 – Feb. 2024
Advisor: Researcher, Yiming Li (Nanyang Technological University)

- Review and report on over 40 state-of-the-art papers on backdoor attacks, defenses, and copyright protection.
- Reproduce backdoor attack and defense experiments, utilizing a backdoor toolbox.
- Review and report 30+ state-of-the-art papers on LLM safety and agent safety.

Automated Detection of Web Advertisements Mar. 2023 – Apr. 2024
Advisor: Prof. Haitao Xu (Zhejiang University)

- Analyze and report on 10+ state-of-the-art papers concerning advertisement detection.
- Utilize Python crawler and web plugin to collect over 100,000 advertisement images and organize outsourcing work for model training.
- Reproduce AdGraph, Webgraph (state-of-the-art works works in advertisement detection), design and complete comparative experiments.

Projects

Digital Inheritance Deliverer (DID) | Tech Stack: Bootstrap (HTML, CSS, JS), Sass, JavaScript, Solidity

- Create a decentralized solution for managing and transferring digital assets securely.
- Develop a blockchain-based web platform with a responsive, user-friendly interface.
- Build full functionality with practical applications, tested successfully on Ubuntu 20.04 for browser preview.

64-bit RISC-V Software and Hardware Integration | Tech Stack: C, Assembly Language, Verilog, Shell

- Develop a toy operating system running on a 64-bit RISC-V CPU.
- Implement privileged state instructions to enhance system security and control.
- Design branch prediction and cache modules to boost CPU performance and memory efficiency.
- Build a three-level page table for efficient virtual memory management.

Billiards Simulation Game | Programming Language: C

- Develop a realistic billiards simulation game, accurately modeling physics to provide users with an engaging and interactive experience.
- Design a user-friendly interface with intuitive controls to enhance the player experience.
- Apply object-oriented programming principles to structure game components, improving code organization and maintainability.

Skills

Languages: English (Advanced, Toefl: 105, Reading 26, Listening 29, Speaking 23, Writing 27), Mandarin (Native)
Programming: Python, C,C++, Verilog, Shell, HTML
Software Tools: LATEX, MS Office Softwares, IDA, ChatGPT, Zotero
Soft Skills: Willing to communicate, Giving a presentation, Taking responsibility, Organizational skills