



Xun Liu

University of Chinese Academy of Sciences
Beijing, China

+86-19568718517

liuxun21@mailsucas.ac.cn

antiquality.github.io

EDUCATION

University of Chinese Academy of Sciences

B.S. in Cybersecurity.

– GPA: 3.83/4.0

– Specialized Courses: *Introduction to Cybersecurity, Stanford CS229: Machine Learning Course* (self-learning).

Aug. 2021 - Present

Beijing, China

RESEARCH EXPERIENCE

Institute of Computing Technology, Chinese Academy of Sciences

Research Assistant, advised by Prof. Fei Sun

– Trustworthy machine learning; LLM Safety.

Sep. 2023 - Present

Beijing, China

Secure Learning Lab, University of Illinois at Urbana-Champaign

Student Intern (Remote), advised by Prof. Bo Li

– Trustworthy machine learning; Agent Safety; CodeAgent benchmark.

Feb. 2024 - Present

Beijing, China

PUBLICATIONS AND PREPRINTS

The Butterfly Effect of Model Editing: Few Edits Can Trigger Large Language Models Collapse

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, Xueqi Cheng. *ACL, 2024.*

SELECTED PROJECTS

• Low-/Mid-Resource Language Jailbreaking LLM (python)

2023

The Trojan Detection Challenge 2023 (LLM Edition)

– **Most Compute-Efficient Special Award** of Large Model Subtrack of Red Teaming Track.

– Placed 5th in development phase of Large Model Subtrack of Red Teaming Track.

– Tools & technologies used: Python, PyTorch, LLaMa2.

– Select 87 low- and 24 mid-resource languages and translate the malicious prompts via Baidu/DeepL/Azure API. Use GPT-3.5 to supervise the tautological transformation of target prompts.

[LEADERBOARD]

TEACHING EXPERIENCE

Peer Mentor: Class 2311

University of Chinese Academy of Sciences

Aug. 2023 - Present

Teaching Assistant: Introduction to Computer Science

University of Chinese Academy of Sciences, Instructor: Prof. Zhiwei Xu

Feb. 2023 - Jul. 2023

HONORS AND AWARDS

First Class Academic Scholarship (Top 5%) *University of Chinese Academy of Sciences* 2022

National Scholarship Nomination (Top 2%) *Ministry of Education of the People's Republic of China* 2022

Outstanding Student Cadre *University of Chinese Academy of Sciences* 2022,2023,2024

Merit Student *University of Chinese Academy of Sciences* 2022,2023,2024

Bronze Medal *The 2021 ICPC Asia Nanjing Regional Contest* 2021

First Prize (Senior Group) *National Olympiad in Informatics in Provinces* 2018,2019

TECHNICAL SKILLS AND INTERESTS

Programming Languages: C/C++, Python, MySQL, Verilog, HTML/CSS, Javascript

Tools and Technologies: Git, Linux

Frameworks: PyTorch

Language: Chinese(native), English(conversational), French(elementary)

Areas of Interest: Photograph, Designing