Xun Liu

University of Chinese Academy of Sciences
Beijing, China

+86-19568718517
liuxun21@mails.ucas.ac.cn
antiquality.github.io

Education

University of Chinese Academy of Sciences *Aug. 2021 - Present*
B.S. in Cybersecurity. Beijing, China
– GPA: 3.88/4.0, Rank: 1/17
– Specialized Courses: *Introduction to Cybersecurity*, *Stanford CS229: Machine Learning Course* (self-learning).

Research Experience

Institute of Computing Technology, Chinese Academy of Sciences *Sep. 2023 - Present*
Research Assistant, advised by Prof. Fei Sun Beijing, China
– Trustworthy machine learning; LLM Safety.

Selected Projects

- **Low-/Mid-Resource Language Jailbreaking LLM (python)** *2023*
  The Trojan Detection Challenge 2023 (LLM Edition)
  – 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.
  – Placed 5th in development phase of Large Model Subtrack of Red Teaming Track. **[Leaderboard]**

Teaching Experience

Teaching Assistant: Introduction to Computer Science *Feb. 2023 - Jul. 2023*
University of Chinese Academy of Sciences, Instructor: Prof. Zhiwei Xu

Honors and Awards

First Class Academic Scholarship (Top 5%) *University of Chinese Academy of Sciences* *2022*
National Scholarship Nomination *Ministry of Education of the People's Republic of China* *2022*
Outstanding Student Cadre *University of Chinese Academy of Sciences* *2022,2023*
Merit Student *University of Chinese Academy of Sciences* *2022,2023*
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

updated on December 4, 2023