Tianhao Li

(+86) 13121515269 tianhao.li tianhao.li2@duke.edu LinkedIn Google Scholar Orcid Github

Recent Experiences

SplxAI **August 2024 – Present**
Research Scientist (Part-time) *New York, United States*

- **Duties**: **(i)** Research on industry-leading generative AI red team safety assessment and alignment framework, outputting algorithmic prototype systems and academic papers.
- **Supervisor**: CEO Kristian Kamber, CTO Ante Gojsalić, Lead Red Team Data Scientist Dorian Granoša

Beijing TopSec Network Security Technology **March 2024 – August 2024 (6 mos)**
Security Researcher @ Department of Innovative Technology Research *Beijing, China*

- **Duties**: **(i)** Contributed to the development of a pioneering prototype system for machine learning model security assessment, focusing on the design and implementation of a robust backend evaluation module. Successfully debugged and mitigated adversarial and poisoning attacks, enhancing the system's resilience and accuracy.; **(ii)** Contribute to a phishing email detection project, overseeing dataset collection and enhancement initiatives. Successfully applied Large Language Models (LLMs) and Supervised Fine Tuning (SFT) to develop a robust detection system.
- **Supervisor**: GM of Innovation Research Dr. Wei Wang

NSFOCUS Technologies Group Co.,Ltd **September 2023 – March 2024 (7 mos)**
Security Researcher @ Dubhe Lab, Innovation Institute *Beijing, China*

- **Duties**: **(i)** Design core algorithms and develop a prototype for NSFOCUS LLM Security Assessment System (LSAS); **(ii)** Enhance the security & privacy performance of NSFGPT and other private LLMs through offensive testing, i.e. red teaming; **(iii)** Conduct literature reviews encompassing research papers and Gartner's consulting reports on trustworthy AI, enriching the threat intelligence database; **(iv)** Write white paper on secure LLM.
- **Impacts**: The release of LSAS, SecLLM NSFGPT, and white paper Enhancing Network Security with SecLLM.
- **Supervisor**: Principal Security Researcher Dr. Xingkai Wang, CIO & GM of Innovation Research Dr. Wenmao Liu

High-Tech Innovation Institute **July 2022 – July 2023 (1 yr 1 mo)**
Research Assistant @ VR & AR Technology Innovation Center *Beijing, China*

- **Duties**: **(i)** Participate in the national key research and development program; **(ii)** Research on gait-based intelligence, surveillance and reconnaissance (ISR) system; **(iii)** Research on classification of human emotion based on electroencephalogram (EEG) and Swin-Transformer; **(iv)** Research on early warning mechanism for aviation safety based on QAR stream data; **(v)** Research on threat model and mitigation methods in AI-XR-based Metaverse.
- **Impacts**: Co-authored 7 papers indexed in SCI/EI, secured victory in 4 national professional contests, delivered 3 oral presentations and presented 1 poster at international conferences.
- **Supervisor**: Prof. Dr. Lijun Wang, Prof. Dr. Zhengping Li, Assoc. Prof. Dr. Ying Li

Selected Projects

LLM Security Vulnerability Assessment System **September 2023 – June 2024 (Expected)**
Algorithm Researcher, Full-stack Developer Opening Defense | Mid Inspection | Thesis Defense

- **As algorithm researcher**: Designed probes and detectors for diverse vulnerabilities (e.g., adversarial attack, jailbreak, DAN). Implemented multi-modal LLM adoption for testing across text, image, and audio domains. (LLM red teaming tool)
- **As full-stack developer**: Developed Python and Django-based B/S platform supporting ModelOps, model monitoring, and privacy protection features. (Practice of Gartner's AI TRiSM conceptual framework)

Gait-based Intelligence, Surveillance, Reconnaissance (ISR) System **May 2023 – August 2023**
Team Leader, Algorithm Engineer, Backend Developer Demo Video 1 | Demo Video 2 | About Page

- **As team leader**: Assigned team roles based on strengths, led system security capacity building, led technical architecture design, and prepared technical report.
- **As algorithm engineer**: Implemented gait recognition pipeline, trained and evaluated models using PyTorch, and developed system security features including authentication, tracking, and early warning mechanisms for physical attacks.
- **As backend developer**: Developed encryption and network modules utilizing SM2/SM3/SM4, including secure key distribution and data encryption/decryption on public channels.

Open Source Activities

- leondz/garak(Generative AI Red-teaming Tool Kit, Metasploit in the field of LLM Security. My main contributions include multi-modal capabilities and the development of intelligent red teaming capabilities.):
  - **Merged**: Python interpreter version bug fix#296; Replicate LLM generator exception error fix#401; Add exception handling logic to the Detector base class#566; Multi-modal Jailbreaking Attack on LLaVA #587 (Make it the **world's first** tool with multimodal red teaming capabilities.). (v0.9.0.9, v0.9.0.11, v0.9.0.13)
  - **Ongoing**: Prompt Architecture Enhancement for Better Multi-modal Red Teaming (#658); Network Proxy Feature For Generator (DavidLee528/garak:proxy_dev); AutoFD: LLM as smart failure detector, a generic multi-level failure detection algorithm (Research Project).
- jdyjjj/All-in-One-Gait(Prototype system of gait recognition, includes three processes: object detection, silhouette segmentation, and gait feature extraction. It involves three deep learning models: ByteTrack, PaddleSeg, and GaitBase. My main contributions were algorithm performance optimization.)
  - **Merged**: **68% performance enhance** of silhouette segmentation module#12.
- EasyJailbreak/EasyJailbreak(Hybrid Jailbreak Attack Prompt Generation Framework)
  - **Merged**: Python interpreter version bug fix#16#17#18

Publications

- Tianhao Li, Weizhi Ma, Yujia Zheng, Xinchao Fan, Guangcan Yang, Ying Li, Lijun Wang, Zhengping Li*. A Survey on Gait Recognition Against Occlusion: Taxonomy, Dataset and Methodology. *PeerJ Computer Science, 2024 (SCI, JCR-Q1, Major Revision)*
- Weizhi Ma, Ying Li*, Tianhao Li, Haowei Yang, Zhengping Li, Lijun Wang, Junyu Xuan. SFSWTS: EEG Emotion Recognition Based on Spatial-frequency Shifted Windows Time Self-attention Neural Network. *Pattern Recognition, 2024 (SCI, JCR-Q1, Peer Review)*
- Weizhi Ma, Yujia Zheng, Tianhao Li, Zhengping Li, Ying Li, Lijun Wang*. A comprehensive review of deep learning in EEG-based emotion recognition: classifications, trends, and practical implications. *PeerJ Computer Science, 2024 (SCI, JCR-Q1)*
- Tianhao Li, Yujia Zheng, Weizhi Ma, Guangshuo Wang, Zhengping Li, and Lijun Wang*. Trustworthy Metaverse: A Comprehensive Investigation into Security Risks and Privacy Issues in Artificial Intelligence-Extended Reality Systems. *SID Symposium Digest of Technical Papers, 2024 (ICDT)*
- Yujia Zheng, Tianhao Li, Weizhi Ma, Jiaxiang Zheng, Zhengping Li*, and Lijun Wang. Unveiling Privacy Challenges: Big Data-Driven Digital Twins in Smart City Applications. *SID Symposium Digest of Technical Papers, 2024 (ICDT)*
- Yujia Zheng, Tianhao Li, Weizhi Ma, Zhengping Li, Lijun Wang, and Ying Li*. Automated Pricing and Replenishment Decision for Vegetable Products Based on Hybrid Machine Learning Models. *Electronics, Communications and Networks, 2024*
- Weizhi Ma, Zhengping Li, Tianhao Li, Yujia Zheng, and Lijun Wang. Application of Virtual Reality Technology in the Diagnosis and Treatment of Psychological Disorders: An Electroencephalography (EEG)-Based Approach. *SID Symposium Digest of Technical Papers, 2024 (ICDT)*
- Tianhao Li, Yujia Zheng, Haoan Zhang, Weizhi Ma, and Ying Li*. Research on Real-time Early Warning Mechanism of Aviation Safety Based on Finite State Machine Underlying in QAR Stream Data. *Proceedings of the 2023 7th International Conference on Big Data and Internet of Things, 2023 (BDIoT, Oral)*

Talks

- A Survey on Gait Recognition Against Occlusion: Taxonomy, Dataset and Methodology *(English Poster Presentation)* @ 2023 5th International Conference on Machine Learning and Intelligent System (MLIS'23), Macau SAR, China. [certificate]
- Automated Pricing and Replenishment Decision For Vegetable Products Based on Hybrid Machine Learning Models *(English Oral Presentation)* @ 2023 5th International Conference on Machine Learning and Intelligent System (MLIS'23), Macau SAR, China. [certificate]
- A Comprehensive Review of Deep Learning in EEG-based Emotion Recognition: Classifications, Trends, and Practical Implications *(English Oral Presentation)* @ 2023 5th International Conference on Machine Learning and Intelligent System (MLIS'23), Macau SAR, China. [certificate]
- Dive Into QAR Stream Data - A Real-time Early Warning Mechanism *(English Oral Presentation)* @ 2023 7th International Conference on Big Data and Internet of Things (BDIoT'23), Beijing, China. [certificate][slide]

Honors and Awards

- Jul.2024, Top 8 in Galaxy Generative AI Safety Contest Attack Track *Top9.0%, National*
- Feb.2024, S Prize in 2023 Mathematical Contest In Modeling (MCM) [paper] *Top?%, National*
- Nov.2023, Second Prize in 11th Digital Media Technology and Creativity Contest [certificate] *Top9.8%, National*
- Nov.2023, Third Prize in 8th National Cryptography Contest [certificate][list][news] *Top12.4%, National*
- Sep.2023, Second Prize in Contemporary Undergraduate Mathematical Contest in Modeling [cert.] *Provincial*
- Sep.2023, Merit-based Scholarships For Outstanding Students *Top0.15%, School*
- Aug.2023, Third Prize in 16th CISCN Security Project Contest [cert.][report][slide][video] *Top13.1%, National*
- Jun.2023, Third Prize in 16th CISCN AWDP (Attack with Defense Plus) Contest [certificate] *Regional*
- Apr.2023, First Prize in 13th MathorCup Mathematical Modeling Challenge [cert.][paper] *Top5.7%, National*
- Apr.2023, First Prize in 14th Lanqiao Software Development and Algorithm Contest [certificate] *Provincial*
- Sep.2022, Merit-based Scholarships For Outstanding Students *Top2.91%, School*
- Jun.2022, Third Prize in 15th CISCN CTF (Capture the Flag) Contest [certificate] *Regional*
- Jun.2021, Third prize in 12th Lanqiao Software Development and Algorithm Contest [cert.], *Top15.3%, National*
- Sep.2021, Merit-based Scholarships For Outstanding Students *Top17.61%, School*
- Apr.2021, First Prize in 12th Lanqiao Software Development and Algorithm Contest [certificate] *Provincial*

Education

Duke University **August 2024 – May 2026 (Expected)**

Master of Science Student in Medical Physics

Research Focus: Trustworthy ML(Security, Privacy, Robustness, Fairness); Large Language Model; AI for Medicine

Core Courses: Academic Research, Modern Diagnostic Imaging Systems, Machine Learning, Deep Learning

North China University of Technology **September 2020 – June 2024**

Bachelor of Engineering in Information Security *GPA: 3.89/4 Grade: 90.88/100 transcript*

Core Courses: Applied Cryptography, Data Security, Image Processing, Principles of Computer Composition, Operating System, Reverse Engineering, Digital Forensic, Software Security, Network Attack and Defense

Honors: Outstanding Dissertation Award (Rank 1/101, Top 0.99%), Elite Student Scholarship (Rank 1/329, Top 0.30%)

Skills

Programming Language & OS: C/C++11, Python3, GNU/Linux(Ubuntu)

Teamwork & Documentation: Mandarin (Native), English (Bilingual proficiency), Git, SVN, LaTeX, Markdown

Software Framework: Machine Learning(PyTorch, Tensorflow, LangChain, TensorRT, OpenCV), ML Vulnerability Assessment(Garak, PyRIT, Counterfit), Cryptography(OpenSSL, GmSSL), Development(Django, BoostC++, MFC)

AI Red Teaming: M-LLM, Adversarial Attack, Prompt Injection, Insecure Plugins, Fine-tuning, RLHF, CoT, RAG

Miscellaneous: Driving (Since Sep 2020), Photography, Adobe PS/PR/AE/AI/AU

Extracurricular Activities

- Conference Attendance: ISC 2016, GeekPwn 2020, GeekCon 2023, CyberSecurity 2023, CCF NCCA 2024.
- Training Experience: Google Digital Talent Development Program (certificate), China Computer Federation (CCF) Artificial Intelligence Course (certificate), Ministry of Industry and Information Technology of the People's Republic of China Industry and Information Technology Talents Project (certificate), Rescue Skills (CPR and Care For Injures) Training of Red Cross Society of China Beijing Branch (certificate).
- Mountain Biking: See POV videos on bilibili channel. Top 9 racer in 2023 GDL Thaiwoo downhill race.
- Road/Trail Running and Hiking: Finisher in 2023 Chongli 168 Ultra Trail(Biggest in Asia) DTC-100KM race.
- Photography and Off-road Driving: Keen on wildlife photography and exploring the unknown.
- Voluntary service: Nearly 400 hours were recorded by Beijing Volunteer Association (As of July 2023)