Tianhao Li

tianhao.li tianhao.li@duke.edu LinkedIn Google Scholar Github

Research Interests and Overview

I have been working on AI safety and security research for over two years, one year in industry and one year in academia, recently focusing on LLM agents.

Education Background

Duke University **August 2024 - May 2026 (Expected)**
Master of Science Candidate in DKU Medical Physics *GPA: 3.5/4*

Research: Security and Privacy in Foundation Models and Agentic Systems for Healthcare

With: Prof. Neil Gong, Prof. Chaowei Xiao

North China University of Technology **September 2020 - June 2024**
Bachelor of Engineering in Information Security *GPA: 3.9/4*

Project: A Multi-modal LLM Red Teaming Platform Based on NVIDIA/Garak

Honors: Outstanding Thesis Award (1/101), Outstanding Student Scholarship (1/329)

Work Experiences

ByteDance **November 2025 – February 2026 (Expected)**
ByteIntern - LLM Agent DataOps *Beijing, China*

- Served as lead author for research projects about LLM Agent, resulting in industry-validated and peer-reviewed publications.

SIGMIR **January 2025 – Present**
Co-founder & CEO - Interdisciplinary Research *CA, United States*

- Founded SIGMIR, a 501(c)(3) non-profit; built the core team, led early operations, and oversee all projects.
- Facilitating connections between academia and industry and among individuals from diverse disciplines to cultivate collaborative research and innovation.

TOPSEC **March 2024 – September 2024**
Security Researcher (Intern) - Adversarial ML - with FTE return offer (declined) *Beijing, China*

- Developed a prototype software system for ML model adversarial robustness evaluation;
- Contributed to a phishing email detection system using SFT-LLMs, with responsibility for dataset collection and refinement.

NSFOCUS **September 2023 – March 2024**
Security Researcher (Intern) - LLM Red Teaming *Beijing, China*

- Contribute to algorithms design and prototype development of NSFOCUS LLM Security Assessment System (LSAS);
- Performed offensive security test on multiple public and private large language models to evaluate vulnerabilities;
- Reviewed academic papers and Gartner reports on trustworthy AI to enrich the threat intelligence database.

Publications

- **Tianhao Li**, Chuangxin Chu, Yujia Zheng, Bohan Zhang, Neil Zhenqiang Gong, and Chaowei Xiao. (2025). A2ASecBench: A Protocol-Aware Security Benchmark for Agent-to-Agent Multi-Agent Systems. ICLR 2026 Under Review - openreview.net/forum?id=LfdFnakqGJ.

- **Tianhao Li**, Jingyu Lu, Chuangxin Chu, Tianyu Zeng, Yujia Zheng, Mei Li, Haotian Huang, Bin Wu, Zuoxian Liu, Kai Ma, Xuejing Yuan, Xingkai Wang, Keyan Ding, Huajun Chen, and Qiang Zhang. (2024). SciSafeEval: A comprehensive benchmark for safety alignment of large language models in scientific tasks. *AAAI 2025 AI for Cybersecurity - ACM TIST Under Review - arXiv:2410.03769.*

- Yujia Zheng[1st], **Tianhao Li**[1st], Haotian Huang, Tianyu Zeng, Jingyu Lu, Chuangxin Chu, Yuekai Huang, Ziyou Jiang, Qian Xiong, Yuyao Ge and Mingyang Li (2025). Are all prompt components value-neutral? Understanding the heterogeneous adversarial robustness of dissected prompt in large language models. *ACL ARR Oct 2025 - Under Review - arXiv:2508.01554.*

- Tianyu Zeng[1st], **Tianhao Li**[1st], Yujia Zheng[1st], Kaizhu Huang, and Zhenyu Yang. (2025). Do Synthetic Medical Images Pose Unseen Risks to Clinical Segmentation Tasks?. *ACM TIST Under Review*.
- Weizhi Ma, Ying Li, **Tianhao Li**, Haowei Yang, Zhengping Li, Lijun Wang, and Junyu Xuan. (2025). SFSWTS: A spatial-frequency shifted windows and time self-attention network for EEG emotion recognition. *Neurocomputing* - *doi.org/10.1016/j.neucom.2025.130309*.
- Qian Xiong, Yuekai Huang, Ziyou Jiang, Zhiyuan Chang, Yujia Zheng, **Tianhao Li**, and Mingyang Li. (2025). Butterfly effects in toolchains: A comprehensive analysis of failed parameter filling in LLM tool-agent systems. *EMNLP 2025 Findings* - *arXiv:2507.15296*.
- Qiang Zhang, Xiang Zhuang, Chenyi Zhou, Yihang Zhu, Tong Xu, **Tianhao Li**, Dianbo Liu, Shengchao Liu, Keyan Ding, Emine Yilmaz, Haofen Wang, and Huajun Chen. (2025). Beyond Accuracy: Comprehensive Alignment for AI-Driven Molecular Design. *Nature Computational Science Major Revision*.
- Yuanyuan Wei, Xianxian Liu, Yao Mu, Changran Xu, Guoxun Zhang, **Tianhao Li**, Zida Li, Wu Yuan, Ho-Pui Ho, and Mingkun Xu. (2025). From droplets to diagnosis: AI-driven imaging and system integration in digital nucleic acid amplification testing. *Biosensors and Bioelectronics* - *doi.org/10.1016/j.bios.2025.117741*.
- Guofan Zhang, Xuanzhi Wang, Qirui Sun, Hanchi Zhao, Jiaao Han, Haoyuan Yang, Hanyu Wang, Chengshu Tian, Aizhen Kuang, Lanxiang Lai, Chulong Zhang, **Tianhao Li** and Rui Liu (2025). A versatile method for accurately predicting electronic absorption spectra of tetrapyrrole macrocycles. *ChemRxiv* - *doi.org/10.26434/chemrxiv-2025-9gsp2*.
- Yepeng Feng, Yujia Zheng., **Tianhao Li** and Xuan Tian. (2025). Personalized Query Recommendation for Large Language Models Guided by Dynamic User Profiles and Bidirectional Intention. *Neurocomputing Submit*.
- Yujia Zheng, Tianyu Zeng, Haotian Huang, Chulong Zhang, **Tianhao Li**, Jingyu Lu, Chuangxin Chu, Manju Liu, Chunhao Wang, Kaizhu Huang, Fang-Fang Yin and Zhenyu Yang. (2025). Assessing the Risks of Synthetic Data in Deep Learning-Based PET Tumor Segmentation. *International Journal of Radiation Oncology, Biology, Physics* - *doi.org/10.1016/j.ijrobp.2025.06.3861*
- Weizhi Ma, Yujia Zheng, **Tianhao Li**, Zhengping Li, Ying Li, and Lijun Wang. (2024). A comprehensive review of deep learning in EEG-based emotion recognition: classifications, trends, and practical implications. *PeerJ Computer Science* - *doi.org/10.7717/peerj-cs.2065*.
- Yuyao Ge, Lingrui Mei, Zenghao Duan, **Tianhao Li**, Yujia Zheng, Yiwei Wang, Lexin Wang, Jiayu Yao, Tianyu Liu, Yujun Cai, Baolong Bi, Fangda Guo, Jiafeng Guo, Shenghua Liu, and Xueqi Cheng. (2025). A Survey of Vibe Coding with Large Language Models. *arXiv:2510.12399*.

Technical Reports

- Ghosh et al. (2025). AILuminate: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons. *MLCommons* - *arXiv:2503.05731* - mlcommons.org/en/ailuminate.
- Goel et al. (2025). AILuminate Security: Introducing v0.5 of the Jailbreak Benchmark from MLCommons. *MLCommons* - *wp-content* - mlcommons.org/ailuminate/jailbreak/.
- Laszewski et al. (2025). AI Benchmarks Carpentry and Democratization. *MLCommons*.

Academic Talks

- Invited talk - Toward Trustworthy Generative Foundation Models and Autonomous Agent System (25 mins). *Institute of Software, Chinese Academy of Sciences*, Aug 2025, Beijing, China.
- Contributed talk - Guardians of Trust: Safety & Privacy for Healthcare Foundation Models (15 mins). *Duke Kunshan Graduate Student Colloquium*, Apr 2025, Jiangsu, China.
- Contributed talk - SciSafeEval: A Comprehensive Benchmark for Safety Alignment of Large Language Models in Scientific Tasks (20 mins). *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI-2025), Artificial Intelligence for Cyber Security (AICS)*, Mar 2025, Philadelphia, PA, USA.

Professional Services

- Journal Referee: ACM TIST, IEEE TAI, IEEE TBME, IEEE JBHI, IJHCI, ESWA, EAAI, RESS, COMPUT NETW, PATTERN RECOGN LETT.
- Program Committee Member: ICLR, AAAI, ACL ARR, IJCAI, COLING, IJCNN, ISBI, CHIL, LLMSEC

Open Source Activities

NVIDIA/garak

github.com/nvidia/garak - Generative AI red-teaming & assessment kit *Contributor*

- Extended from single modality (text-to-text) only to multimodal (any-to-any) capability with #587: implemented a FigStep-based [(image+text)-to-text] jailbreaking pipeline for LLaVA - the first multi-modal red-teaming tool.
- Improved system reliability through bug fixes (#296, #401) and enhanced exception handling (#566) across core modules.

Google/adk-samples

github.com/google/adk-samples - A collection provides ready-to-use agents built on top of the google-adk *Contributor*

- Enhancing Data Science Agent's configurability by introducing environment-based dataset management (#220, #221).

A2AProject/a2a-python

github.com/a2aproject/a2a-python - Official python SDK for the Google Agent2Agent (A2A) protocol *Contributor*

- Enhanced reliability by redesigning the streaming error-handling architecture to accurately surface server status and messages. (#502, #505)

SIGMIR-ORG/TARMAS

github.com/sigmir-org/tarmas - A framework for threat analysis and Risk metrics in agent systems *Owner*

- Lead project development and research direction, overseeing design, implementation, and maintenance

Media Coverage

- Concordia AI. AI Action Summit: Public Interest AI. https://concordia-ai.com/wp-content/uploads/2024/11/Concordia-AI-French-AI-Action-Summit-Public-Interest-AI-contribution.pdf
- Concordia AI. AI Action Summit: AI of Trust - Security & Safety. https://concordia-ai.com/wp-content/uploads/2024/11/Concordia-AI-French-AI-Action-Summit-AI-of-Trust-Security-Safety-contribution.pdf
- Shanghai AI Lab. Frontier AI Risk Management Framework. https://research.ai45.shlab.org.cn/safework-f1-framework.EN.pdf
- IEEE Spectrum. One Chatbot Safety Benchmark To Test Them All: AILuminate aims to create an industry standard for large language model safety. https://spectrum.ieee.org/ai-safety-standard
- BusinessWire. MLCommons Launches AILuminate, First-of-Its-Kind Benchmark to Measure the Safety of Large Language Models. https://businesswire.com/news/home/20241204285618/en/MLCommons-Launches-AILuminate-First-of-Its-Kind-Benchmark-to-Measure-the-Safety-of-Large-Language-Models
- Binary Verse AI. Vibe Coding Tools: The Definitive 2025 Guide, Based on a Landmark AI Survey. https://binaryverseai.com/best-vibe-coding-tools-top-agentic-cursor-ai/
- PYMNTS Logo. Vibe Coding Won't Replace Humans Anytime Soon, Data Shows https://www.pymnts.com/artificial-intelligence-2/2025/vibe-coding-wont-replace-humans-anytime-soon-data-shows/

Competition Awards

- Aug.2024, Second Prize in 17th CISCN Security Project Contest [certificate] *National*
- Jul.2024, Top 5 out of 89 in Galaxy Generative AI Safety Contest Attack Track *National*
- Feb.2024, S Prize in 2024 Mathematical Contest In Modeling (MCM) [paper] *National*
- Nov.2023, Second Prize in 11th Digital Media Technology and Creativity Contest [certificate] *National*
- Nov.2023, Third Prize in 8th National Cryptography Contest [certificate][list][news] *National*
- Aug.2023, Third Prize in 16th CISCN Security Project Contest [certificate][report][slide][video] *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 [certificate][paper] *National*
- Apr.2023, First Prize in 14th Lanqiao Software Development and Algorithm Contest [certificate] *Provincial*
- Jun.2022, Winning Prize in 15th CISCN Security Project Contest [certificate][report][slide] *National*
- 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 [certificate] *National*
- Apr.2021, First Prize in 12th Lanqiao Software Development and Algorithm Contest [certificate] *Provincial*