Tinghao Xie

✉ thx@princeton.edu · 🔗 https://tinghaoxie.com · vtu81

Education

Princeton University, Princeton, United States 08/2022 – Present

Ph.D. candidate, Electrical Computer Engineering (ECE)

- **Advisor**: Prof. *Prateek Mittal*

Princeton University, Princeton, United States 08/2022 – 09/2024

M.A., Electrical Computer Engineering (ECE)

Zhejiang University, Zhejiang, China 09/2018 – 06/2022

B.E., Computer Science and Technology (CS)

- **GPA**: 3.99/4.00 (92.07/100)
- **Rank**: 1st/186

Research Interests

Safe AI; Secure and Reliable AI Systems; Robust and Adversarial ML.

Publications & Manuscripts

Red-teaming NSFW Image Classifiers with Generative AI Tools
Tinghao Xie, Yueqi Xie, Alireza Zareian, Shuming Hu, Felix Juefei-Xu, Xiaowen Lin, Ankit Jain, Prateek Mittal, Li Chen
Will be online soon (Under Review)

On Evaluating the Durability of Safeguards for Open-Weight LLMs
Xiangyu Qi*, Boyi Wei*, Nicholas Carlini, Yangsibo Huang, **Tinghao Xie**, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, Peter Henderson
ICLR 2025

SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal Behaviors 🔗, 📄, </>
Tinghao Xie*, Xiangyu Qi*, Yi Zeng*, Yangsibo Huang*, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, Prateek Mittal
ICLR 2025

Fantastic Copyrighted Beasts and How (Not) to Generate Them 🔗, 📄, </>
Luxi He*, Yangsibo Huang*, Weijia Shi*, **Tinghao Xie**, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan Zhang, Danqi Chen, Peter Henderson
ICLR 2025

AI Risk Management Should Incorporate Both Safety and Security 📄
Xiangyu Qi, Yangsibo Huang, Yi Zeng, Edoardo Debenedetti, Jonas Geiping, Luxi He, Kaixuan Huang, Udari Madhushani, Vikash Sehwag, Weijia Shi, Boyi Wei, **Tinghao Xie**, Danqi Chen, Pin-Yu Chen, Jeffrey Ding, Ruoxi Jia, Jiaqi Ma, Arvind Narayanan, Weijie J Su, Mengdi Wang, Chaowei Xiao, Bo Li, Dawn Song, Peter Henderson, Prateek Mittal
Preprint (Under Review)

Assessing the brittleness of safety alignment via pruning and low-rank modifications 🔗, 📄, </>
Boyi Wei*, Kaixuan Huang*, Yangsibo Huang*, **Tinghao Xie**, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, Peter Henderson
ICML 2024

Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! 🔗, 📄, </>, 📰

Xiangyu Qi*, Yi Zeng*, **Tinghao Xie***, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal†, Peter Henderson†
ICLR 2024 (Oral)
*This work was exclusively reported by 📰 **New York Times**, and covered by many other social medias!*

BaDExpert: Extracting Backdoor Functionality for Accurate Backdoor Input Detection 📄, </>

Tinghao Xie, Xiangyu Qi, Ping He, Yiming Li, Jiachen T. Wang, Prateek Mittal
ICLR 2024

Towards A Proactive ML Approach for Detecting Backdoor Poison Samples 🔗, 📄, </>

Xiangyu Qi, **Tinghao Xie**, Jiachen T. Wang, Tong Wu, Saeed Mahloujifar, Prateek Mittal
USENIX Security Symposium 2023

Revisiting the Assumption of Latent Separability for Backdoor Defenses 🔗, 📄, </>

Xiangyu Qi*, **Tinghao Xie***, Yiming Li, Saeed Mahloujifar, Prateek Mittal
ICLR 2023

Towards Practical Deployment-Stage Backdoor Attack on Deep Neural Networks 🔗, 📄, </>

Xiangyu Qi*, **Tinghao Xie***, Ruizhe Pan, Jifeng Zhu, Yong Yang and Kai Bu
CVPR 2022 (Oral)

Work Experience

Meta GenAI (Media Safety Team), Menlo Park, United States 05/2024 – 08/2024

Research Intern

- **Mentor**: *Li Chen*
- **Project**: Red-teaming NSFW Image Classifiers and Text-to-Image System Safety

Honors and Awards

Francis Robbins Upton Fellowship	08/2022
Outstanding Graduate Thesis	06/2022
Champion of Zhejiang University Bodybuilding Competition (70kg Level)	05/2022
Elite Liu Yongling Scholarship (1/802)	2020 – 2021
Tencent Scholarship (5/802)	2020 – 2021
The 2nd Class Prize in ASC 20-21 Student Supercomputer Challenge	01/2021
Narada Scholarship (1/372)	2019 – 2020
Champion of Zhejiang University DFM Hip-hop Crew Battle	2019

Services

Reviewing for CVPR 2025, ICLR 2025, Neurips 2025, ICML 2024, ICLR 2024 Set LLM Workshop, Neurips 2024, Neurips 2023 (*Top Reviewer*), Neurips 2023 BUGS Workshop.

Skills

- **Programming**: C/C++, Python, JavaScript, CUDA, Verilog, Shell, MATLAB, ActionScript, HTML.
- **Software**: LaTeX, Vivado, Adobe {Photoshop, Premiere Pro, After Effects, Audition}.
- **Languages known**: English(fluent), Chinese(native), Cantonese(native).
- **Hobbies**: Climbing, Skiing, Dance (Hip-hop, House, etc.), Power Lifting, Swimming, Basketball, Billiards.

Latest update date: 01/25/2025