Keyan Guo

917-374-6606 | keyanguo@buffalo.edu | linkedin.com/in/keyan96 | keyanUB.github.io

Research Overview

My research centers on **Generative AI Security**—designing trustworthy generative AI against threats such as jailbreak attacks and insecure code generation—and **Generative AI for Security**—leveraging generative AI to detect and mitigate online harms such as hateful memes, unsafe user-generated content, and cyberbullying. I have published at top-tier venues including IEEE S&P, USENIX Security, NDSS, ACM CHI, and EMNLP.

Education

University at Buffalo, SUNY — Buffalo, NY
Ph.D. in Computer Science and Engineering — *January 2022 – December 2026 (Expected)*
- Advisor: Prof. Hongxin Hu

University at Buffalo, SUNY — Buffalo, NY
M.S. in Engineering Science — *July 2019 – June 2021*

Qingdao University — Qingdao, China
Bachelor of Engineering in Information Engineering — *July 2014 – June 2018*

Publications

† indicates equal contributions.

Selected Publications

- Ruchi Panchanadikar, Yang Hu†, **Keyan Guo**†, Amelia L. Hall, Hongxin Hu, Nishant Vishwamitra, Guo Freeman. "Beyond Age-Based Restrictions: Rethinking Children's Online Safety Through Comparing Parent–Child Perspectives of Risks in User-Generated Content Games." In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2026
- Yong Zhuang†, **Keyan Guo**†, Juan Wang, Yiheng Jing, Xiaoyang Xu, Wenzhe Yi, Mengda Yang, Bo Zhao, Hongxin Hu. "I know what you MEME! Understanding and Detecting Harmful Memes with Multimodal Large Language Models." In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2025
- Shenyi Zhang, Yuchen Zhai, **Keyan Guo**, Hongxin Hu, Shengnan Guo, Zheng Fang, Lingchen Zhao, Chao Shen, Cong Wang, Qian Wang. "JBShield: Defending Large Language Models from Jailbreak Attacks through Activated Concept Analysis and Manipulation." In *Proceedings of 34th USENIX Security Symposium (USENIX Security)*, 2025
- Yiheng Jing, Mingming Zhang, Yong Zhuang, Jiacheng Guo, Juan Wang, Xiaoyang Xu, Wenzhe Yi, **Keyan Guo**, Hongxin Hu. "HVGuard: Utilizing Multimodal Large Language Models for Hateful Video Detection." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025
- **Keyan Guo**, Ayush Utkarsh, Wenbo Ding, Isabelle Ondracek, Ziming Zhao, Guo Freeman, Nishant Vishwamitra, Hongxin Hu. "Moderating Illicit Online Image Promotion for Unsafe User Generated Content Games Using Large Vision-Language Models." In *Proceedings of 33rd USENIX Security Symposium (USENIX Security)*, 2024
- Nishant Vishwamitra†, **Keyan Guo**†, Farhan Tajwar Romit, Isabelle Ondracek, Long Cheng, Ziming Zhao, Hongxin Hu. "Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models." In *Proceedings of 45th IEEE Symposium on Security and Privacy (S&P)*, 2024

Under Review

- Rupam Patir, **Keyan Guo**, Haipeng Cai, Hongxin Hu. "GRASP: Fortifying LLM-Based Code Generation with Graph-Based Reasoning on Secure Coding Practices." Under review at *ACM Conference on Computer and Communications Security (CCS)*, 2026
- Tian Zhang, Yiwei Xu, Juan Wang, **Keyan Guo**, Xiaoyang Xu, Bowen Xiao, Quanlong Guan, Jinlin Fan, Jiawei Liu, Zhiquan Liu, Hongxin Hu. "AgentSentry: Mitigating Indirect Prompt Injection in LLM Agents via Temporal Causal Diagnostics and Context Purification." *arXiv preprint, 2026*

Other Publications

- Jaden Mu, David Cong, Helen Qin, Ishan Ajay, **Keyan Guo**, Nishant Vishwamitra, Hongxin Hu. "Detecting Cyberbullying in Visual Content: A Large Vision-Language Model Approach." In *Proceedings of 23rd IEEE International Conference on Machine Learning and Applications*, 2024
- Ebuka Okpala, Nishant Vishwamitra, **Keyan Guo**, Song Liao, Long Cheng, Hongxin Hu, Xiaohong Yuan, Jeannette Wade, Sajad Khorsandroo. "AI-Cybersecurity Education Through Designing AI-based Cyberharassment Detection Lab." In *Proceedings of 28th Colloquium for Information Systems Security Education (CISSE)*, 2024
- **Keyan Guo**, Guo Freeman, Hongxin Hu. "Moderating Embodied Cyber Threats Using Generative AI." In *CHI 2024 Workshop on Novel Approaches for Understanding and Mitigating Emerging New Harms in Immersive and Embodied Virtual Spaces*, 2024
- Nishant Vishwamitra, Ebuka Okpala, Song Liao, **Keyan Guo**, Sandeep Shah, Hongxin Hu, Xiaohong Yuan and Long Cheng. "Enhancing AI-Centered Social Cybersecurity Education through Learning Platform Design." In *Proceedings of 28th Colloquium for Information Systems Security Education (CISSE)*, 2024
- **Keyan Guo**, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, Hongxin Hu. "An Investigation of Large Language Models for Real-World Hate Speech Detection." In *Proceedings of 22nd IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2023
- Nishant Vishwamitra, **Keyan Guo**, Liao Song, Jaden Mu, Zheyuan Ma, Long Cheng, Ziming Zhao, Hongxin Hu. "Understanding and Analyzing COVID-19-related Online Hate Propagation Through Hateful Memes Shared on Twitter." In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2023
- Ebuka Okpala, Nishant Vishwamitra, **Keyan Guo**, Liao Song, Long Cheng, Hongxin Hu, Yongkai Wu, Xiaohong Yuan, Jeannette Wade, Sajad Khorsandroo. "AI-Cybersecurity Education Through Designing AI-based Cyberharassment Detection Lab." In *Proceedings of IEEE Frontiers in Education Conference (FIE)*, 2023
- Nishant Vishwamitra, **Keyan Guo**, Hongxin Hu, Ziming Zhao, Long Cheng, Feng Luo. "Understanding and Measuring Robustness of Vision and Language Multimodal Models." In *Proceedings of International Conference on Secure Knowledge Management (SKM)*, 2023
- Wenbo Ding, Liao Song, **Keyan Guo**, Ziming Zhao, Hongxin Hu. "Exploring Vulnerabilities in Voice Command Skills for Connected Vehicles." In *Proceedings of EAI International Conference on Security and Privacy in Cyber-Physical Systems and Smart Vehicles (EAI SmartSP)*, 2023
- **Keyan Guo**, Wentai Zhao, Jaden Mu, Nishant Vishwamitra, Ziming Zhao, Hongxin Hu. "Understanding the Generalizability of Hateful Memes Detection Models Against COVID-19-related Hateful Memes." In *Proceedings of 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2022
- **Keyan Guo**, Shaik Sabiha, Foad Hajiaghajani, Chunming Qiao, Hongxin Hu, Ziming Zhao. "Demo: Understanding the Effects of Paint Colors on LiDAR Point Cloud Intensities." In *Workshop on Automotive and Autonomous Vehicle Security (AutoSec)*, 2022

Teaching Experience

Graduate Teaching Assistant Fall 2020 – Fall 2024
University at Buffalo, SUNY *Buffalo, NY*

- Taught 200+ students across AI, machine learning, computer security, and database courses:
  * CSE 465/565: Computer Security (Fall 2023, Fall 2024)
  * CSE 574: Introduction to Machine Learning (Spring 2024)
  * CSE 460/560: Data Models and Query Languages (Fall 2021 – Spring 2023)
  * CSE 368: Introduction to Artificial Intelligence (Fall 2020)
- Designed course projects and 6+ hands-on labs on AI and computer security
- Served as instructor in charge of the AI Security module

Invited Talks & Tutorials

ICWSM 2024 Tutorial June 2024
The 18th International AAAI Conference on Web and Social Media *Buffalo, NY*

- Presented a tutorial on machine learning-based online abuse defense, covering our designed platform, current research, and hands-on cybersecurity labs.

Guest Lecturer: AI Security and Adversarial Machine Learning Fall 2023
CSE 465/565 Computer Security, University at Buffalo *Buffalo, NY*

- Delivered a guest lecture on AI security and adversarial machine learning to 140+ graduate students.

Great Lakes Security Day April 2023
Rochester Institute of Technology *Western New York*

- Presented research on mitigating online hate in the evolving cyber environment.

SEAS Lightning Talk 2023
School of Engineering and Applied Sciences, University at Buffalo *Buffalo, NY*

- Presented on AI-related safety issues and ongoing research projects. 12 Ph.D. students in the university were selected for this event.

GenCyber Camp 2022, 2023
North Carolina A&T State University *Greensboro, NC*

- Invited speaker on AI-related cybersecurity challenges and cyberbullying defense. Presented hands-on cybersecurity labs across two sessions.

Academic Service

Program Committee

- [Artifacts TPC] IEEE Symposium on Security and Privacy (S&P), 2026
- [Artifacts TPC] USENIX Security Symposium, 2025, 2026
- [Artifacts TPC] Network and Distributed System Security Symposium (NDSS), 2025, 2026
- ACM Workshop on Security Implications of Deepfakes and Cheapfakes (WDC), 2026
- Annual Computer Security Applications Conference (ACSAC), 2023, 2024
- IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2023–2026
- [Session Chair] IEEE International Conference on Machine Learning and Applications (ICMLA), 2023

Journal/Conference Reviewer

- IEEE Transactions on Dependable and Secure Computing (TDSC), 2024–2026
- ACM Transactions on the Web (TWEB), 2025
- ACM Transactions on Cyber-Physical Systems (TCPS), 2025
- NDSS Symposium, 2025
- ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2024, 2025
- Information Systems Frontiers, 2022–2024
- ASONAM, ICMLA, MSN, TrustCon, TPS-ISA, WISP (various years, 2022–2024)

Mentoring

High School Research Mentees

- Jaden Mu (now at Carnegie Mellon University) – Co-authored publications at ICMLA 2022, ICMLA 2023, and ICMLA 2024
- Alexander Hu (now at UCLA) – Co-authored publication at ICMLA 2023
- David Cong (now at Duke University) – Co-authored publication at ICMLA 2024
- Wentai Zhao (now at University of Michigan) – Co-authored publication at ICMLA 2022
- Helen Qin (now at Caltech), Ishan Ajay, Johnson Chen – Co-authored publications at ICMLA 2024

Undergraduate and Master's Research Mentees

- Ayush Utkarsh (now at Snap) – Co-authored publication at USENIX Security 2024
- Shaik Sabiha (now at Adobe) – Co-authored publication at AutoSec 2022
- Amardhruva Narasimha Prabhu (now at Google), Radhika Singh (now at Lineaje) – Mentored thesis projects

Honors and Awards

- Internet Society Fellowship, NDSS Symposium, 2025
- Best Research Project (PhD) Award, Department of Computer Science and Engineering, University at Buffalo, 2024
- Student Grant, USENIX Security Symposium, 2024
- Student Academic Excellence Showcase, University at Buffalo, 2023
- Best AI Poster Award, Department of Computer Science and Engineering, University at Buffalo, 2023
- Best Graduate Teaching Award, Department of Computer Science and Engineering, University at Buffalo, 2022

Open-Source Software & Datasets

UGCG-Guard [Code] [Dataset]

- A large vision-language model based framework for detecting illicit online image promotion for unsafe user-generated content games. Published at USENIX Security 2024.

HateGuard [Code] [Dataset]

- A chain-of-thought reasoning framework using large language models for moderating new waves of online hate speech. Published at IEEE S&P 2024.

JBShield [Code]

- A defense framework against LLM jailbreak attacks through activated concept analysis and manipulation. Published at USENIX Security 2025.

AI & Cybersecurity Education Labs [Platform]

- A collection of hands-on AI and cybersecurity labs for education, deployed at GenCyber Camp (2022, 2023) and presented as a tutorial at ICWSM 2024.