# Keyan Guo

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

## Research Interests

- Trustworthy & Responsible AI
- Generative AI Safety and Cybersecurity
- Intelligent Agents and Assistants
- Multimodal Content Moderation
- Security & Privacy in Online Social Networks

## Education

**University at Buffalo, SUNY** Buffalo, NY
*Ph.D. in Computer Science and Engineering* *January 2022 – December 2026 (Expected)*
- Advisor: Prof. Hongxin Hu
- GPA: 4.0/4.0

**University at Buffalo, SUNY** Buffalo, NY
*M.S. in Engineering Science* *July 2019 – June 2021*
- Supervisor: Prof. Mingchen Gao
- GPA: 3.8/4.0

**Qingdao University** Qingdao, China
*Bachelor of Engineering in Information Engineering* *July 2014 – June 2018*
- GPA: 3.6/4.0

## Professional Publications

* indicts equal contributions

- 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) (Big 4 Security Conference)***, 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." Accepted by ***The Conference on Empirical Methods in Natural Language Processing (EMNLP)***, 2025
- 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) (Big 4 Security Conference, Acceptance rate: 16.1%)***, 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), Acceptance rate: 18.32%)***, 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/Oakland) (Big 4 Security Conference, Acceptance rate:14.9%)***, 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." Accepted by *28th Colloquium for Information Systems Security Education (CISSE), 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." Accepted by *28th Colloquium for Information Systems Security Education (CISSE), 2024*
- 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 (ICMLA)*, 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(TA)** Fall 2020 – Fall 2024
*University at Buffalo, SUNY* *Buffalo, NY*

- Teaching assistant for the following courses:
  - CSE 465/510/565: 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)
- My TA responsibilities include:
  - Designing and grading AI/ML course projects
  - Instructor in charge of the AI Security

* Designing and grading hands-on AI and cybersecurity labs
* Participating in the development of course-related learning materials
* Answering student questions about course knowledge, assignments and projects

**Guest Speaker**

*The 18th International AAAI Conference on Web and Social Media* *Buffalo, NY*

- ICWSM 2024 Tutorial.
  My team and I presented our tutorial at the 18th ICWSM conference. The tutorial introduces the topic of machine learning-based online abuse defense, including our designed platform, current research, and self-developed hands-on cybersecurity labs.

*University at Buffalo, SUNY* *Buffalo, NY*

- SEAS 2023 Lightning Talk.
  I was invited to give a lightning talk about AI-related safety issues and our ongoing research projects in the School of Engineering and Applied Sciences at University of Buffalo. 12 Ph.D. students were invited to this event.
- CSE 702 Seminar: Machine Learning and Cybersecurity.
  I was invited to talk about AI-related safety issues and adversarial machine learning (AML) knowledge in a seminar course at the University of Buffalo. Over 20 graduate students attended the seminar.

## Academic Experience

**Program Committee**

* International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2025
* [Artifacts Technical Program Committee] USENIX Security Symposium (USENIX Security), 2025
* [Session Chair] IEEE International Conference on Machine Learning and Applications (ICMLA), 2023
* IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2023, 2024
* Annual Computer Security Applications Conference (ACSAC), 2023

**Conference/Journal Reviewer**

* ACM Transactions on the Web (TWEB), 2025
* ACM Transactions on Cyber-Physical Systems (TCPS), 2025
* The Network and Distributed System Security (NDSS) Symposium, 2025
* SIGSEC Workshop on Information Security and Privacy (WISP), 2024
* ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2024, 2025
* IEEE Transactions on Dependable and Secure Computing (TDSC), 2024, 2025
* International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2023, 2024
* International Conference on Mobility, Sensing and Networking (MSN), 2023
* IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCon), 2023
* IEEE International Conference on Machine Learning and Applications (ICMLA), 2022, 2023
* IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA), 2022
* Information Systems Frontiers, 2022, 2023, 2024

**Invited Talk**

* *Great Lakes Security Day.*
  I was invited to present our research work about mitigating online hate in the evolving cyber environment at Great Lakes Security Day (GLSD) 2023 on April $21^{st}$.GLSD brings together premier practitioners, researchers, students, and funding partners in security to share the latest advances, debate roadmaps, and future directions, create new collaborations, and seek new opportunities in cybersecurity in and around Western and Upstate New York.
* *GenCyber Camp.*
  I was invited to speak on AI-related cybersecurity challenges and cyberbullying defense in the GenCyber Camp. Additionally, I presented our designed cybersecurity labs at two sessions at North Carolina State A&T University in 2022 and 2023.

## Honors and Awards

- Internet Society Fellowship. NDSS, 2025
- CSE Best Research Project (PhD) Award. University at Buffalo, SUNY, 2024
- USENIX Conference Student Grant. USENIX Security, 2024
- Student Academic Excellence Showcase. University at Buffalo, SUNY, 2023
- CSE Best AI Poster Award. University at Buffalo, SUNY, 2023
- CSE Best Graduate Teaching Award. University at Buffalo, SUNY, 2022

## Technical Skills

**Languages:** Python, C/C++, Java
**AI & Machine Learning Tools:** PyTorch, TensorFlow, HuggingFace, Jupyter
**Data Management & Visualization:** MySQL, NoSQL, Pandas, NumPy, Matplotlib
**Collaboration & Productivity:** Git/GitHub, Visual Studio, Amazon SageMaker, MATLAB