Keyan Guo

Ph.D. Candidate
Department of Computer Science and Engineering
University at Buffalo, The State University of New York
keyanguo@buffalo.edu | keyanub.github.io | Google Scholar | (917) 374-6606

CAREER OBJECTIVE

AI is reshaping both the threats facing our digital society and the defenses available to counter them. My career objective is to play a leading role in **CyberAI**, the intersection of AI and cybersecurity, as a researcher and educator: advancing research that makes AI systems trustworthy and uses AI to protect people, preparing students to build and secure AI systems, and translating discoveries into real-world defenses.

Research interests: **AI security and safety**, **AI for cybersecurity and online safety**, **human-centered security**

EDUCATION

Ph.D. in Computer Science and Engineering	University at Buffalo, Buffalo, USA	2022–2027
Toward Trustworthy Generative AI: Resisting Manipulation, Securing Outputs, and Mitigating Online Harms		
M.S. in Engineering Science	University at Buffalo, Buffalo, USA	2019–2021
B.S. in Information Engineering	Qingdao University, Qingdao, China	2014–2018

RESEARCH EXPERIENCE

Research Assistant, University at Buffalo — Jan 2022–Present
Advisor: Prof. Hongxin Hu

- **Securing LLMs, AI Agents, and AI-Generated Code**
  - Investigated why LLMs fail under jailbreak and multimodal attacks, and developed defenses that detect and suppress unsafe behaviors [SEC'25, CCS'26].
  - Guided student research on defending LLM agents against indirect prompt injection through temporal causal diagnostics and context purification [NDSS'27].
  - Co-led DualGauge, a benchmark that jointly evaluates the security and correctness of code generated by LLMs and agents; proposed a framework for secure code generation [EMNLP'26 (Findings), preprint].
- **Detection and Moderation of Online Harms**
  - Developed a reasoning approach that enables LLMs to recognize and moderate newly emerging waves of online hate [S&P'24].
  - Proposed a multimodal LLM framework for understanding and detecting harmful memes [NDSS'25].
  - Investigated and moderated illicit image promotion for unsafe user-generated content games (UGCGs) using large vision-language models [SEC'24].
  - Guided student research on hateful video detection and on cyberbullying detection across social media and video games [EMNLP'25, ICMLA'25].
  - Conducted empirical studies of how well hateful-meme detectors generalize and how capable LLMs are at real-world hate speech detection [ICMLA'22, ICMLA'23].
- **Children's and Users' Online Safety**
  - Compared parent and child perceptions of risk in user-generated content games, motivating approaches to children's online safety beyond age-based restrictions [CHI'26].
  - Analyzed the spread of COVID-19-related online hate through memes shared on Twitter [ASONAM'23].

ACHIEVEMENT HIGHLIGHTS AND HONORS

- **6** papers in BIG-4 security conferences (1 S&P, 2 SEC, 1 CCS, 2 NDSS); **2** papers in top AI/ML conferences (EMNLP); **1** paper in a top HCI conference (CHI); over 20 publications in **security**, **AI/ML**, and **HCI**
- NDSS **Internet Society Fellowship** 2025
- UB CSE Best Ph.D. Research Project Award 2024
- Student Travel Grants, NDSS and USENIX Security 2024
- UB Student Academic Excellence Showcase 2024
- UB CSE Best AI Poster Award 2023
- UB CSE Best Graduate Teaching Award 2022

PUBLICATIONS

- My work has been cited **390+** times with an h-index of 7 and an i10-index of 7 (Google Scholar, Oct 2026) (*equal contribution, †undergraduate and ‡high-school student under my supervision).

Conference Papers

NDSS'27 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." Accepted by *the Network and Distributed System Security Symposium* (**NDSS**), 2027.

CCS'26 Shenyi Zhang, Keyan Guo, Zihao Wang, Xuebin Li, Lingchen Zhao, Hongxin Hu, Chao Shen, Qian Wang. "MMAligner: Safeguarding Multimodal Large Language Models through Representation Calibration." Accepted by *the ACM Conference on Computer and Communications Security* (**CCS**), 2026.

EMNLP'26 (Findings) Rupam Patir*, Keyan Guo*, Suvadra Barua, Abhijeet Pathak, Dinesh Gudimetla, Jiawei Guo, Hongxin Hu, Haipeng Cai. "DualGauge: Automated Joint Security-Functionality Benchmarking of Specification-Only Code Generation by LLMs and Coding Agents." Accepted by *the Conference on Empirical Methods in Natural Language Processing* (**EMNLP Findings**), 2026.

CHI'26 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 CHI Conference on Human Factors in Computing Systems* (**CHI**), 2026.

SEC'25 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 the 34th USENIX Security Symposium* (**SEC**), 2025.

EMNLP'25 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.

NDSS'25 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.

ICMLA'25 David Cong[‡], Keyan Guo, Hongxin Hu. "Toward Unified Moderation of Cyberbullying Across Social Media and Video Games." In *Proceedings of the International Conference on Machine Learning and Applications* (ICMLA), 2025.

SEC'24 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 the 33rd USENIX Security Symposium* (**SEC**), 2024.

ICMLA'24 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 the International Conference on Machine Learning and Applications* (ICMLA), 2024.

S&P'24 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 the IEEE Symposium on Security and Privacy* (**S&P**), 2024.

ASONAM'23 Nishant Vishwamitra, Keyan Guo, Song Liao, 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 the International Conference on Advances in Social Networks Analysis and Mining* (**ASONAM**), 2023.

ICMLA'23 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 the International Conference on Machine Learning and Applications* (ICMLA), 2023.

SmartSP'23 Wenbo Ding, Song Liao, Keyan Guo, Fuqiang Zhang, Long Cheng, Ziming Zhao, Hongxin Hu. "Exploring vulnerabilities in voice command skills for connected vehicles." In *Proceedings of the International Conference on Security and Privacy in Cyber-Physical Systems and Smart Vehicles* (SmartSP), 2023.

SKM'23 Nishant Vishwamitra, Keyan Guo, Hongxin Hu, Ziming Zhao, Feng Luo. "Understanding and Measuring Robustness of Vision and Language Multimodal Models." In *Proceedings of the International Conference on Secure Knowledge Management* (SKM), 2023.

ICMLA'22 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 the International Conference on Machine Learning and Applications* (ICMLA), 2022.

Journal Papers

JCISSE'25 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." *Journal of the Colloquium for Information Systems Security Education* (JCISSE), vol. 12, no. 1, 2025.

JCISSE'25 Nishant Vishwamitra, Ebuka Okpala, Song Liao, Keyan Guo, Sandeep Shah, Hongxin Hu, Xiaohong Yuan, Long Cheng. "Enhancing AI-Centered Social Cybersecurity Education through Learning Platform Design." *Journal of the Colloquium for Information Systems Security Education* (JCISSE), vol. 12, no. 1, 2025.

Workshop and Position Papers

FutureG'25 Rupam Patir, Qiqing Huang, Keyan Guo, Wanda Guo, Guofei Gu, Haipeng Cai, Hongxin Hu. "Towards LLM-Assisted Vulnerability Detection and Repair for Open-Source 5G UE Implementations." In *Proceedings of the Workshop on Security and Privacy of Next-Generation Networks in the Network and Distributed Systems Security Symposium* (FutureG), 2025.

CHI EA'24 Keyan Guo, Guo Freeman, Hongxin Hu. "Moderating Embodied Cyber Threats Using Generative AI." Position paper in the CHI 2024 Workshop on *Novel Approaches for Understanding and Mitigating Emerging New Harms in Immersive and Embodied Virtual Spaces* (CHI EA), 2024.

AutoSec'22 Shaik Sabiha, Keyan Guo, Foad Hajiaghajani, Chunming Qiao, Hongxin Hu, Ziming Zhao. "Understanding the Effects of Paint Colors on LiDAR Perceptions." In *Proceedings of the International Workshop on Automotive and Autonomous Vehicle Security* (AutoSec), 2022.

Preprints

Preprint Rupam Patir, Keyan Guo, Haipeng Cai, Hongxin Hu. "SoK: AI Secure Code Generation: Progress, Pitfalls, and Paths Forward." arXiv:2606.25195, 2026.

Preprint Rupam Patir, Keyan Guo, Haipeng Cai, Hongxin Hu. "Fortifying LLM-Based Code Generation with Graph-Based Reasoning on Secure Coding Practices." arXiv:2510.09682, 2025.

GRANT PROPOSAL CONTRIBUTIONS

- **Collaborative Research: CyberAI Innovation: Educating the Future Workforce in Secure Agentic AI Software Development through an Open Learning Platform**
  Sponsoring Agency: *National Science Foundation (NSF)*
  PI: Hongxin Hu Aug. 2026–July 2029
  Award No.: 2623205 Award Amount (University at Buffalo): $250,000
  Project: Develops OpenSecCoder, an open-source experiential learning platform that prepares computing and non-computing students to develop, evaluate, and secure agentic AI software systems.
  Contributions: Contributed to writing the project implementation plan, particularly the design of eight hands-on labs to engage students in critical topics in secure AI-driven software development.

- **Collaborative Research: EAGER: SaTC-EDU: Learning Platform and Education Curriculum for Artificial Intelligence-Driven Socially-Relevant Cybersecurity**
  Sponsoring Agency: *National Science Foundation (NSF), SaTC-EDU*
  PI: Hongxin Hu June 2021–Sept. 2024
  Award No.: 2114982 Award Amount (University at Buffalo): $70,000
  Project: Developed an open learning platform that translates research on AI-driven social cybersecurity into hands-on educational materials for computing and non-computing students.
  Contributions: Co-developed and implemented course materials and hands-on labs translating AI-driven social cybersecurity research into practical learning activities. Contributed to drafting proposal sections on AI security course design and lab-based instruction.

TEACHING EXPERIENCE

- My teaching was recognized by students and faculty with the **UB CSE Best Graduate Teaching Award**, 2022.

Role	Course	Selected Contributions	Semester(s)
Guest Lecturer	**CSE 4/565** Computer Security	Delivered lectures on AI security, LLM safety, and adversarial machine learning.	Fall 2024–Fall 2026
	CPSC 4200 (Clemson University) Computer Security Principles	Developed and taught hands-on AI security labs.	Fall 2023
Graduate Teaching Assistant	**CSE 4/565** Computer Security	Delivered substitute lectures; designed assignments and hands-on labs; contributed to exam design.	Fall 2023; Fall 2024
	CSE 574 Introduction to Machine Learning	Led discussion sections on course concepts and problem solving.	Spring 2024
	CSE 460/560 Data Models and Query Languages	Designed course projects and contributed to exam design.	Spring 2022–Spring 2023
Undergraduate Teaching Assistant	**CSE 460/560** Data Models and Query Languages	Designed course projects and contributed to exam design.	Fall 2021
	CSE 368 Introduction to Artificial Intelligence	Provided course support through the TA duties listed below.	Fall 2020

All TA appointments included grading assignments and exams and holding office hours.

STUDENT MENTORING

Provided research mentoring under the supervision of Prof. Hongxin Hu.

- **Junior Ph.D. Students**
  - Rupam Patir, *AI-assisted Software Security*, University at Buffalo — Spring 2025–Present
- **Master's Students**
  - Shishir Hebbar, *MCP server security for AI agents*, University at Buffalo — Fall 2024–Spring 2025
  - Radhika Singh, *GenAI and multimodal risks*, University at Buffalo — Fall 2024–Fall 2025
  - Ayush Utkarsh, *online harassment and abuse moderation*, University at Buffalo — Fall 2023–Spring 2024
- **Undergraduate Students**
  - Isabelle Ondracek, *online social network security*, University at Buffalo — Fall 2023–Spring 2025
  - Shaik Sabiha, *autonomous driving perception safety*, University at Buffalo — Fall 2022
- **High School Students**

  Mentored **seven high school students** and published peer-review papers [1] [2] [3] [4].
  - Alicia Gu, Now at Stanford University — Summer 2025
  - David Cong, Williamsville East High School (Now at Duke University) — Summer 2024, 2025
  - Ishan Ajay, John Glenn School — Summer 2024
  - Helen Qin, Thomas Jefferson High School for Science and Technology — Summer 2024
  - Jaden Mu, East Chapel Hill High School (Now at Carnegie Mellon University) — Summer 2023, 2024
  - Alexander Hu, Union County Magnet High School (Now at UCLA) — Summer 2023
  - Wentai Zhao, Northville High School (Now at University of Michigan) — Summer 2023

TALKS AND PRESENTATIONS

Invited Talks and Tutorials

- *Year-End Mini Workshop of the SFS Scholars*, University at Buffalo 2025
  "Secure LLM-based Code Generation with Reasoning-based Methodology"
- *ICWSM Tutorial*, The 18th International AAAI Conference on Web and Social Media 2024
  "Machine Learning-Based Online Abuse Defense: Platform, Research, and Hands-on Labs"
- *Great Lakes Security Day*, Rochester Institute of Technology 2023
  "On Mitigating Online Hate in the Evolving Cyber Environment"
- *GenCyber Camp*, North Carolina A&T State University 2023
  "AI-related Cybersecurity Challenges and Cyberbullying Defense"

Research Presentations

- *NDSS* — Oral presentation, San Diego, CA, USA 2025
  "I know what you meme! Understanding and Detecting Harmful Memes with Multimodal LLMs"
- *CISSE Cybersecurity Education Conference* — Oral presentation, Online 2025
  "Enhancing AI-Centered Social Cybersecurity Education through Learning Platform Design"
  "AI-based Cyberharassment Detection Lab"
- *FutureG Workshop at NDSS* — Oral presentation, San Diego, CA, USA 2025
  "Towards LLM-Assisted Vulnerability Detection and Repair for Open-Source 5G UE Implementations"
- *USENIX Security* — Oral presentation, Philadelphia, PA, USA 2024
  "Moderating Illicit Online Image Promotion for UGCGs Using Large Vision-Language Models"
- *ICMLA* — Oral presentation, Jacksonville, FL, USA 2023
  "An Investigation of Large Language Models for Real-World Hate Speech Detection"
- *ICMLA* — Oral presentation, Online 2022
  "Understanding Hateful Memes Detection Models Against COVID-19-related Hateful Memes"

PROFESSIONAL ACTIVITIES AND SERVICES

- **Technical Program Committee:**
  - USENIX Security Symposium (SEC) 2027
  - IEEE Int'l Conf. on Trust, Privacy, and Security in Intelligent Systems and Applications (TPS) 2026
  - [Artifacts TPC] IEEE Symposium on Security and Privacy (S&P) 2026
  - [Artifacts TPC] USENIX Security Symposium (SEC) 2025, 2026
  - [Artifacts TPC] Network and Distributed System Security Symposium (NDSS) 2025, 2026
  - ACM Workshop on Security Implications of Deepfakes and Cheapfakes (WDC) 2026
  - Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM) 2024, 2026
  - IEEE Int'l Conf. on Machine Learning and Applications (ICMLA) 2024, 2025
- **Reviewer:**
  - *Conferences:* NDSS (2025), ACSAC (2023, 2024), ACL (2025), COLM (2026)
  - *Journals:* IEEE TDSC (2024–2026), ACM TCPS (2025), ACM TALLIP (2024, 2025), Expert Systems with Applications (2025), Information Systems Frontiers (2022–2024)
- **Organizing Committee:**
  - [Session Chair] IEEE Int'l Conf. on Machine Learning and Applications (ICMLA) 2023
- **Department Service:**
  - [Student Host] UB CSE Faculty Hiring—Graduate Student Roundtable 2025
  - [Review Committee] UB CSE Student Poster Demo Day Competition 2024, 2025