Keyan Guo

keyanguo@buffalo.edu | +1 (917) 374-6606 | keyanub.github.io | linkedin.com/in/keyan96

Professional Summary

Ph.D. candidate in Computer Science & Engineering at UB, specializing in AI security and generative AI safety. Author of 10+ peer-reviewed publications, including work at USENIX Security, IEEE S&P, NDSS, and EMNLP. Experienced in multimodal content safety, LLM jailbreak defense, and secure code generation, with a proven record of collaboration, scalable experimentation on GPU clusters, and mentoring junior researchers.

Education

Ph.D. in CSE, University at Buffalo | GPA: 4.0 | Advisor: Prof. Hongxin Hu | *Exp. December 2026*
M.S. in Engineering Science, University at Buffalo, Buffalo, NY | *GPA: 3.8* | *June 2021*

Skills

Languages: Python (Advanced), C/C++, Java, SQL, Bash
AI/ML: PyTorch, TensorFlow, HuggingFace Transformers, OpenAI/Anthropic APIs
Data & Visualization: Pandas, NumPy, Matplotlib, MATLAB, MySQL/NoSQL
Tools & Platforms: Jupyter, Git, Docker, AWS, LaTeX

Selected Publications & Research Projects

Defending Large Language Models from Jailbreak Attacks (USENIX Security 2025)
- Developed the first activated-concept defense framework against jailbreak attacks, achieving an average detection accuracy of 0.95 and reducing attack success rate from 61% to 2% across diverse LLMs.

Classifying Hateful Videos with a Reasoning-based Detection Framework (EMNLP 2025)
- Proposed *HVGUARD*, the first reasoning-based hateful video detector, integrating multimodal features with model-generated rationales via a Mixture-of-Experts (MoE) network, achieving 0.86 accuracy and outperforming SOTA baselines.

Detecting Online Harmful Memes with Multimodal LLMs (NDSS 2025)
- Designed a novel in-context learning approach for harmful meme detection that reached 92% accuracy on challenging multimodal benchmarks.

Moderating New Waves of Online Hate with LLM Reasoning (IEEE S&P 2024)
- Built a chain-of-thought-based moderation system to address evolving hate speech triggered by emerging global events, achieving >90% accuracy and F1.

Moderating Unsafe Image Promotion in UGC Games (USENIX Security 2024)
- Identified a new illicit promotion threat from user-generated content games, such as Roblox; curated a large-scale dataset and proposed a zero-shot detection method, setting new SOTA.

Experience

Research Assistant, UB Security Lab Jan 2022 – Present
- Drive multiple research projects in AI safety and security from idea to publication, with results appearing at USENIX Security, IEEE S&P, NDSS, and EMNLP.
- Coordinated cross-project efforts within the lab, managing compute resources, experiment pipelines, and collaborative codebases to ensure reproducibility.
- Mentored junior Ph.D. and Master's students in research methods, paper writing, and experimental design, contributing to a sustained pipeline of high-quality publications.

Teaching Assistant, University at Buffalo Jul 2020 – Nov 2024
- Taught and supported 100+ students/semester in Machine Learning, Computer Security, and Database Systems.
- Designed labs and tutorials integrating real-world AI security cases, improving student comprehension scores.
- Recognized with the **CSE Best Graduate Teaching Award (2022)** for outstanding teaching and mentorship.

Honors & Awards

Internet Society Fellowship, NDSS, San Diego, CA *2025*
CSE Best Research Project (PhD) Award, University at Buffalo *2024*
Student Grant, USENIX Security (sponsored by Google & NSF), Philadelphia, PA *2024*
CSE Best Graduate Teaching Award, University at Buffalo *2022*