Keyan Guo

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

Researcher and engineer specializing in the safety and reliability of large language models and agentic AI systems, with a proven track record in leading venues. Driven by a core belief that the most impactful AI systems must be not only powerful but trustworthy, and committed to bringing that standard to real-world deployment at scale.

Education

University at Buffalo, SUNY — Buffalo, NY
Ph.D. in Computer Science and Engineering, GPA: 4.0/4.0 — *Jan. 2022 – Dec. 2026 (Expected)*

University at Buffalo, SUNY — Buffalo, NY
M.S. in Engineering Science, GPA: 3.8/4.0 — *Jul. 2019 – Jun. 2021*

Qingdao University — Qingdao, China
B.E. in Information Engineering, GPA: 3.6/4.0 — *Jul. 2014 – Jun. 2018*

Technical Skills

Languages: Python (primary), C/C++, Java
AI / ML: PyTorch, HuggingFace Transformers, Multimodal LLMs, LLM Safety Evaluation & Red-teaming, Prompt & Context Engineering, SFT, RLHF
Tools: Git/GitHub, Docker, VS Code, Jupyter, Amazon SageMaker, MySQL, Matplotlib

Selected Projects & Publications — Google Scholar

Secure Agentic Code Generation | *Policy-Driven, Training-Free Approach* — Ongoing
- Designed and co-developed GRASP (paper under submission), a graph-based reasoning framework over Secure Coding Principles; lifted Security Rate from ∼50% to **70–85%** across GPT-4o, GPT-5, Gemini, Claude-3, and Llama-3, with ∼**2**× gains on zero-day CVEs.
- Actively extending GRASP to agentic coding environments (OpenHands, Aider, Cursor) to enforce real-time security reasoning in multi-step code generation workflows.

JBShield | *LLM Jailbreak Defense via Activated Concept Analysis* — USENIX Security 2025
- Designed the core defense mechanism and co-implemented the pipeline; reduced jailbreak attack success rate from **61% to 2%** across 5 LLMs (7B–70B params) with <2% utility degradation on benign queries.

Multimodal Understanding & Content Moderation | *Text → Image → Meme → Video* — 2024–2025
- Developed system for adaptive hate speech detection in text; zero-shot CoT approach outperformed supervised baselines by **10.6–88%** on **30K+ social media posts** across 4 real-world datasets. *(IEEE S&P 2024)*
- Solely developed UGCGGuard end-to-end: LVLM-based image moderation pipeline across 3 gaming platforms; **94% F1** on 4,000+ posts with <5% false-positive rate, directly informing platform policy decisions. *(USENIX Security 2024)*
- Designed and co-developed HMGuard for harmful meme understanding via cross-modal MLLM reasoning on **2,000+ memes**; achieved top-1 on 2 public benchmarks. *(NDSS 2025)*
- Conceived the approach and supervised development of HVGuard, first multimodal hate video system with MoE fusion and CoT reasoning; improved accuracy **6.88–13.13%** and M-F1 **9.21–34.37%** on 20K+ videos. *(EMNLP 2025)*

Platform Safety Analysis | *User Behavior Study on Roblox (380M+ users)* — ACM CHI 2026
- Independently conducted all data collection, processing, and statistical analysis: content-analyzed **2,000 user reviews** (inter-coder reliability κ=0.887); chi-square testing revealed significant parent-child risk perception gaps (p <0.001) across age groups, challenging blanket age-gate policies.

Experience

Graduate Research Assistant — UBSec Lab — Buffalo, NY
University at Buffalo, SUNY | NSF-funded | Advisor: Prof. Hongxin Hu — *Jan. 2022 – Present*
- Lab Manager: coordinated day-to-day research operations, mentored junior PhD and master's students, and managed project timelines across 3 concurrent research threads.
- Authored 15+ publications at Security, AI, and HCI venues including USENIX Security, IEEE S&P, NDSS, EMNLP, and ACM CHI; served as Program Committee member and Artifacts TPC reviewer at leading security and AI venues.

Graduate Teaching Assistant — Buffalo, NY
University at Buffalo, SUNY — *Fall 2020 – Fall 2024*
- TA for 4 courses over 8+ semesters (AI, Database Systems, Computer Security, Machine Learning); instructor-in-charge for AI Security module with 140+ graduate students; **CSE Best Graduate Teaching Award** (UB 2022).

Honors & Awards

Internet Society Fellowship (NDSS 2025) · CSE Best Research Project Award (UB 2024)
Student Academic Excellence Showcase (UB 2023) · CSE Best AI Poster Award (UB 2023)