JIANGRUI ZHENG

New York, NY jzheng36@stevens.edu | zjr769806@gmail.com
www.linkedin.com/in/jiangrui-zheng-34b718209 Phone: 551-342-0096

EDUCATION

Stevens Institute of Technology, New Jersey, USA, **09/2024-Now**
PhD of Computer Science
***Interest**: Machine Learning, NLP, LLM, Security*
Stevens Institute of Technology, New Jersey, USA, **09/2022-05/2024**
Master of Computer Science
American University, Washington, D.C., USA, **08/2016-05/2020**
Bachelor of Science with a Major in Mathematics and a Minor in Computer Science

TECHNICAL SKILLS

Programming Languages: Java, Python, SQL, R, Linux, Markdown
Other: PyTorch, Keras, scikit-learn, Git

PUBLICATION

HateModerate: Testing Hate Speech Detectors against Content Moderation Policies
***Jiangrui Zheng**, Xueqing Liu, Guanqun Yang, et. al.*
In Findings of NAACL 2024, Mexico City, Mexico, June 2024.
From Reviewers' Lens: Understanding Bug Bounty Report Invalid Reasons with LLMs
***Jiangrui Zheng**, Yingming Zhou, Ali Abdullah Ahmad, Hanqing Yao, Xueqing Liu.*
Workshop on Secure and Safe AI Agents for Big Data Infrastructure (*S2AI@BigData2025)*
Improving the Context Length and Efficiency of Code Retrieval for Tracing Security Vulnerability Fixes
*Xueqing Liu, **Jiangrui Zheng**, Guanqun Yang, et. al.*
Can Highlighting Help GitHub Maintainers Track Security Fixes?
*Xueqing Liu, Yuchen Xiong, Qiushi Liu, **Jiangrui Zheng***

PROJECTS

Policy-Aware Evaluation of Hate Speech Detectors (link)
- Created a benchmark aligned to Facebook's policies via a six-step process with 28 annotators, retrieval, data augmentation, and verification.
- Benchmarked SOTA detectors (Google, OpenAI, Facebook) and fine-tuned a top-downloaded Hugging Face model, revealing policy non-conformity and improving it without sacrificing original-test performance.

CVE Patch Retrieval with Hierarchical Embedding (link)
- Proposed a three-phase retrieval pipeline using hierarchical embedding, ElasticSearch pre-ranking, and learning-to-rank, enabling tracing of patching commits across full repositories with long diffs..
- Achieved 18 %–28 % higher Recall@10 and MRR than VoyageAI (SOTA commercial embedding model), and successfully added patch links for 35 CVEs to GitHub Advisory Database.

Explainable Patch Tracing with TfIdf-Highlight (link)
- Constructed a large dataset of 3,573 CVE–patch commit pairs and trained a multi-modal retrieval model (CodeBERT/UnixCoder) to rank patching commits given a CVE description.
- Developed TfIdf-Highlight, an explainable token-highlighting method that outperforms LIME in faithfulness by ~15%, aiding human security maintainers in interpreting model outputs.

Sensitive API Risk Analysis in AI Code Generation
- Built a framework to detect unexpected API usage by code-generation agents.
- Collected 800k+ AI-generated commits (Jules, Claude, Copilot), and found an average of 1.59 vulnerabilities per commit.

Evaluating and Attacking AI Code Reviewers
- Built capability test suites for security evaluation of real-world AI code reviewers (Copilot, CodeRabbit) by combining static analysis (GuardDog Semgrep rules), LLM-based classification, and benign–malicious code pairing.
- Designed and executed targeted attacks — long-context evasion and security regression poisoning — achieving successful attacks on Copilot and CodeRabbit.