Euisuh Jeong

AI Safety Evaluation • Agent Security • Measurement Infrastructure

euisuh.jeong@gmail.com | +82 10-8068-6644 | Seoul, Korea | uiseoya.com | euisuh | euisuh

RESEARCH INTERESTS

How capable coding agents can be evaluated and controlled so that safety claims stay linked to concrete policy decisions, tool traces, and executable outcomes. My work sits between AI safety evaluation, systems security, and empirical software engineering, and centers on a recurring question: when an evaluation reports that a guard or policy changed an agent's behavior, what evidence would have to exist for that claim to be checkable? I am especially interested in measurement infrastructure that makes agent behavior auditable and reproducible, and in where human judgment must enter evaluation design.

EDUCATION

B.S. Computer Science

Carnegie Mellon University | School of Computer Science — Pittsburgh, PA

- Graduated with University Honors; Minor in Mathematical Sciences.

M.S. Data Science and Engineering

Hamad Bin Khalifa University | College of Science and Engineering — Doha, Qatar

- Graduated with University Honors.
- Thesis: *Evaluating the Capabilities and Limitations of Large Language Models in Code Vulnerability Detection* (2024).

RESEARCH EXPERIENCE

Research Engineer

Aug 2023 – Jan 2025

Qatar Computing Research Institute | Cybersecurity Department — Doha, Qatar

- Architected the benchmark, reporting, and visualization stack for aiXamine, an LLM safety evaluation platform spanning 40+ benchmarks and 8 security dimensions; evaluated 50+ models across 2,000+ exams as a founding team member.
- Co-authored the resulting paper, *aiXamine: Simplified LLM Safety and Security* (arXiv:2504.14985).
- Engineered CVEFixesUtil and vulnerable-code dataset tooling over CVE/CWE data in six languages for code-LLM security.
- Prepared exploratory analysis acknowledged in Appendix D of *Poison with Style* (arXiv:2605.27631), ICML 2026.

AI/ML Research Engineer & Computer Vision Technical Lead

Feb 2025 – Nov 2026

Republic of Korea Air Force, AI-Based Technology Team | Staff Sergeant & Squad Leader — Gyeryong, Korea

- Published as first author a 231,347-image runway defect dataset in KOSAP: 52,800 real images, 178,547 synthetic, 9 classes.
- Improved real-image model accuracy from 73.8% to 84.2% (+10.4 pp) by designing and running an 11-soldier, 1,848-person-hour relabeling protocol, an early version of the labeling-quality questions that motivate my current work.
- Led squad-level data, model, and backend delivery of a production detection pipeline inside a closed military network.

RESEARCH PROJECTS AND ARTIFACTS

SurfaceGuardBench — provenance-bound evaluation for coding-agent guard policies

2026 – Present

- Designed and built a solo evaluation protocol binding policy decisions, agent trajectories, and executable outcomes to shared provenance: a 110-case fixed-call battery, a 300-run matched campaign, and a 144-run prospective challenge, with every reported cell linked back to its configuration, task, and trace hashes.
- Introduced a decision-level predicate-fireability check: a necessary precondition establishing that two policies could have decided differently before any contrast between them is reported as a finding.
- Applied the check to the study's own prospective challenge, where the distinguishing decision fired in 20 of 48 matched trajectories, its divergence was one-directional, and it appeared outside the destination class the design intended — conditions invisible in task-success or block-count reporting, the usual summary statistics.
- Released a reviewer artifact supporting deterministic recomputation of every reported aggregate from released traces.

aiXamine — LLM safety and security evaluation platform

2023 – 2025

- Shipped benchmark and visualization infrastructure for black-box LLM safety scorecards across 40+ benchmarks.

RepOrbit / RMI — repository-signal validation

2026 – Present

- Built an event-sourced repository-mining visualization and a Repository Management Index quality score, asking whether repository signals credibly measure software quality — the same question as whether trace signals measure agent safety.

CVEFixesUtil — vulnerability dataset tooling

2024

- Published a CVE/CWE vulnerability-dataset processor for Python, C/C++, Java, Ruby, and Go; dataset on Hugging Face.

PUBLICATIONS AND MANUSCRIPTS

E. Jeong. *SurfaceGuardBench: A Provenance-Bound Protocol for Measuring Coding-Agent Guard Actions and Task Outcomes.* Under review, ACM AISec 2026.

Y. Park, J. Park, **E. Jeong**. *Music Tagging Graph Neural Network with Tag Labels.* KSMI 2026. OpenReview:Tq09T70Miu

E. Jeong, S. Ji, M. Kim, H. Lee. *ROKAF Runway Crack Dataset.* KOSAP, Dec 2025. DOI:10.23379/jkosap.1.2.114

S. Ji, **E. Jeong**, M. Kim, H. Lee. *Deep Learning for PCI Assessment.* KOSAP, Aug 2025. DOI:10.23379/jkosap.1.1.52

F. Deniz, D. Popovic, Y. Boshmaf, **E. Jeong**, et al. *aiXamine: Simplified LLM Safety and Security*. Preprint, 2025. arXiv:2504.14985

E. Jeong. *RepOrbit/RMI*. Manuscript in preparation; targeting MSR 2027 Data & Tool Showcase.

TEACHING

Teaching Assistant — 11-785 Introduction to Deep Learning
Nov 2021 – May 2022
Carnegie Mellon University | Language Technologies Institute
Pittsburgh, PA

- Planned and delivered lectures, homework, recitations, and quizzes for 350+ students.
- Mentored student research teams through project scoping and exploration of new deep-learning research directions.

SELECTED TECHNICAL WORK

Runway Evaluation System
2025 – Present

- Architected a YOLOv11/Co-DETR detection and PCI-scoring pipeline (FastAPI, Celery, Redis, Postgres, Docker, Kubernetes), cutting manual pavement assessment time 98.2% at 86.8% accuracy (dataset paper).

SKILLS

Safety & Security Evaluation: LLM safety evaluation, benchmark harness design, coding-agent trace evaluation, tool-use policy analysis, code-LLM security, CVE/CWE datasets, reproducible evaluation artifacts

ML/AI: PyTorch, Transformers, CUDA, computer vision, graph neural networks, labeling QA, synthetic-data validation

Languages & Infrastructure: Python, C/C++, SQL, FastAPI, Celery, Redis, Postgres, Docker, Kubernetes, AWS, Linux, Git

Spoken Languages: Korean & English (Native/Bilingual) | Japanese (JLPT N4; actively studying)