Soohyeon Choi

Research Scientist | AI and LLM Safety and Security

Singapore shchoi@smu.edu.sg soohyeonc.github.io Google Scholar
github.com/soohyeonc linkedin.com/in/soohyeon-choi ORCID 0009-0002-1252-2263

Profile

AI and LLM security researcher with a Ph.D. in Computer Science and more than five years of research experience. Investigates how advanced language models fail under adversarial use and how generated code can be detected, attributed, and provenance-marked. Research connects compositional jailbreaking, exposure-aware evaluation, human and AI-agent code attribution, and post-hoc watermarking around a common goal of reliable measurement under adversarial and deployment shifts. Builds reproducible, large-scale benchmarks and evaluation pipelines using fine-tuned code encoders, zero- and few-shot LLMs, matched controls, robustness tests, and statistical auditing across Python, Java, and C++. First-author publications include IEEE TDSC, IEEE ICDCS, ICICS, and CSoNet. Research has received 295 Google Scholar citations as of August 2026.

Research Experience

Singapore Management University Singapore
Research Scientist I *May 2025–Present*

- Conduct research on AI and LLM security, including compositional jailbreak attacks, exposure-aware generated-code detection, AI coding-agent attribution, and post-hoc code watermarking.
- Design large-scale evaluation pipelines using fine-tuned code encoders, zero-shot detectors, robustness tests, matched controls, and statistically audited datasets.
- Collaborate with academic and industry researchers across Singapore and the United States, and communicate findings through manuscripts, artifacts, and research presentations.

University of Central Florida Orlando, Florida, USA
Graduate Research Assistant, SEAL Lab *Aug. 2021–May 2025*

- Developed machine-learning methods to identify ChatGPT-generated and ChatGPT-transformed source code and evaluated their robustness to code transformation and distribution shift.
- Built a sequence-to-sequence code-transformation approach for authorship evasion and studied adversarial robustness in language and sensor-based classification systems.
- Published first-authored work in IEEE TDSC, IEEE ICDCS, ICICS, and CSoNet while completing a Ph.D. focused on machine learning, LLMs, and security.

A*STAR Institute for Infocomm Research Singapore
Research Officer *Dec. 2023–Dec. 2024*

- Developed zero-shot, few-shot, and tournament-style LLM methods for source-code authorship attribution, evaluating cross-language generalization and adversarial robustness across C++ and Java.

South Dakota State University Brookings, South Dakota, USA
Graduate Research Assistant, CCT Lab *Aug. 2019–Dec. 2020*

- Developed a low-computation authentication protocol for wireless body-area networks using elliptic-curve cryptography and two-dimensional hash chains.

Education

Ph.D. in Computer Science, University of Central Florida 2021–2025
Advisor: Prof. David Mohaisen. Focus: machine learning, LLMs, and security. Orlando, USA

M.S. in Computer Science, South Dakota State University 2018–2021
Advisor: Prof. Sung Shin. Focus: authentication and sensor-network security. Brookings, USA

B.E. in Computer Engineering, Keimyung University 2011–2017
Daegu, South Korea

Selected AI Safety and Security Research

CommitTrace: AI Coding-Agent Attribution First author; under review, 2026

- Diagnosed why prior detectors fail on repository-local changes, with zero-shot AUROC ranging from 0.398 to 0.528, then developed a code-only representation combining a fine-tuned encoder, character n-grams, and stylometry.
- Achieved binary macro F1 of 0.966 on file patches and 0.953 on hunks across five coding agents; evaluated fragmentation, context, robustness, cross-language transfer, and repository-disjoint generalization.

Multi-Channel Spread-Spectrum Code Watermarking First author; under review, 2026

- Developed a post-hoc, training-free method that embeds a 24-bit provider label through naming and structural channels with explicit erasure-oriented recovery.
- Evaluated 1,750 Python files with 100% clean target-label recovery and 94.1% recovery under 10% random per-site corruption in the tested setting; embedding and detection run on CPU without access to the generating model.

Talking to Themselves: Exposure-Aware Code Detection First author; under review, 2026

- Evaluated heuristic and LLM-based generated-code detectors across C++, Java, and Python; found heuristic AUROC falling from 0.97 on fresh GitHub data to 0.15 on exposed CodeNet data and GPT-4.1 accuracy falling from 93% to 2% after copyright-header insertion.

Compositional Jailbreaking of Aligned LLMs Corresponding author; under review, 2026

- Evaluated all ordered pairs of 12 jailbreak mutators against three LLMs to identify synergistic, destructive, and structurally incompatible attack compositions; developed completeness and validity measures for attack persistence and effectiveness.

I Can Find You in Seconds: LLM Code Authorship Attribution First author; under review, 2026

- Developed zero-shot, few-shot, and tournament-style LLM methods for code authorship attribution, achieving MCC up to 0.78 and scaling to 500 C++ and 686 Java authors with one reference sample per author.

Selected Peer-Reviewed Publications

- **S. Choi** and D. Mohaisen, "Attributing ChatGPT-Generated Source Codes," *IEEE Transactions on Dependable and Secure Computing*, 2025.
- **S. Choi** et al., "Attributing ChatGPT-Transformed Synthetic Code," *IEEE International Conference on Distributed Computing Systems*, 2025.
- M. F. Kharma, **S. Choi**, M. Alkhanafseh, and D. Mohaisen, "Security and Quality in LLM-Generated Code: A Multi-Language, Multi-Model Analysis," *IEEE Transactions on Dependable and Secure Computing*, 2026.
- A. Alkinoon, T. C. Dang, A. Alghuried, A. Alghamdi, **S. Choi**, et al., "A Comprehensive Analysis of Evolving Permission Usage in Android Apps: Trends, Threats, and Ecosystem Insights," *Journal of Cybersecurity and Privacy*, 2025.
- **S. Choi**, R. Jang, D. Nyang, and D. Mohaisen, "Untargeted Code Authorship Evasion with Seq2Seq Transformation," *CSoNet*, 2023.
- **S. Choi**, M. Mohaisen, D. Nyang, and D. Mohaisen, "Revisiting the Deep Learning-Based Eavesdropping Attacks via Facial Dynamics from VR Motion Sensors," *ICICS*, 2023.
- M. Omar, **S. Choi**, D. Nyang, and D. Mohaisen, "Quantifying the Performance of Adversarial Training on Language Models with Distribution Shifts," *CySSS*, 2022.
- M. Omar, **S. Choi**, D. Nyang, and D. Mohaisen, "Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions," *IEEE Access*, 2022.

Technical Skills and Professional Service

Programming and tools: Python, C/C++, Java, SQL, Linux, Git, LaTeX.
Machine learning: PyTorch, TensorFlow, scikit-learn, Transformers, model fine-tuning, LLM evaluation, adversarial ML, benchmark design, statistical evaluation.
Reviewing: IEEE TIFS (2025–2026), IEEE TDSC (2023–2026), ACM AsiaCCS (2025), IEEE TMC (2022), and *Computer Networks* (2022).