

Minkyoo Song

✉ Mail |  LinkedIn |  Google Scholar
last update: October 2025

RESEARCH INTEREST

LLM Security, AI for Security, Data-driven Security, Social Network Analysis

EDUCATION

- **Korea Advanced Institute of Science and Technology (KAIST)** March 2023 - Present
Ph.D. Student in Electrical Engineering, Network and System Security Lab (Advisor: [Seungwon Shin](#)) Daejeon, South Korea
- **Korea Advanced Institute of Science and Technology (KAIST)** March 2021 - February 2023
M.S. in Electrical Engineering, Network and System Security Lab (Advisor: [Seungwon Shin](#)) Daejeon, South Korea
- **Korea Advanced Institute of Science and Technology (KAIST)** March 2016 - February 2021
B.S. in Industrial and Systems Engineering, double majored in Electrical Engineering Daejeon, South Korea

PUBLICATIONS [C]: CONFERENCE, [J]: JOURNAL, [U]: UNDER REVIEW

- [C] J. Kim, S.H. Na, **M. Song**, S. Shin, S. Son. **MoEvil: Poisoning Expert to Compromise the Safety of Mixture-of-Experts LLMs**. 2025 Annual Computer Security Applications Conference (ACSAC 2025) (to appear)
- [C] **M. Song**, H. Kim, J. Kim, S. Shin, S. Son. **Refusal Is Not an Option: Unlearning Safety Alignment of Large Language Models**. 34th USENIX Security Symposium (USENIX Sec 2025)
- [C] H. Kim, **M. Song**, S.H. Na, S. Shin, K. Lee. **When LLMs Go Online: The Emerging Threat of Web-Enabled LLMs**. 34th USENIX Security Symposium (USENIX Sec 2025)
- [C] **M. Song**, H. Kim, J. Kim, Y. Jin, S. Shin. **Claim-Guided Textual Backdoor Attack for Practical Applications**. The 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NACCL 2025 Findings)
- [C] J. Kim, **M. Song**, S.H. Na, S. Shin. **Obliviate: Neutralizing Task-Agnostic Backdoors within the Parameter-Efficient Fine-Tuning Paradigm**. The 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NACCL 2025 Findings)
- [C] **M. Song**, E. Jang, J. Kim, S. Shin. **Covering Cracks in Content Moderation: Delexicalized Distant Supervision for Illicit Drug Jargon Detection**. 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2025)
- [C] J. Kim, **M. Song**, M. Seo, Y. Jin, S. Shin. **PassREfinder: Credential Stuffing Risk Prediction by Representing Password Reuse between Websites on a Graph**. 2024 IEEE Symposium on Security and Privacy (SP) (S&P 2024)
- [J] J. Kim, **M. Song**, M. Seo, Y. Jin, S. Shin, J. Kim. **PassREfinder-FL: Privacy-Preserving Credential Stuffing Risk Prediction via Graph-Based Federated Learning for Representing Password Reuse between Websites**. Elsevier Expert Systems with Applications (ESWA) (to appear)
- [J] J. Choi, J. Kim, **M. Song**, H. Kim, N. Park, M. Seo, Y. Jin, S. Shin. **A Large-Scale Bitcoin Abuse Measurement and Clustering Analysis Utilizing Public Reports**. IEICE Transactions on Information and Systems
- [U] K. Kim, J. Cui, **M. Song**, S. Shin. **Exploring the Familiar Taste of Toxicity: A Causal Influence Analysis of Toxic Comments on Internet Forums**. Invited to Major Revision at IEEE Transactions on Knowledge and Data Engineering (TKDE)
- [U] J. Kim, **M. Song**, S. Shin, S. son. **SafeMoE: Safe Fine-Tuning for MoE LLMs by Aligning Harmful Input Routing**. Submitted to International Conference on Learning Representations (ICLR) 2026
- [U] K. Kim, S.H. Na, **M. Song**, S. Shin. **Global Meta-path-level Counterfactual Explanation for Heterogeneous Graph Neural Networks by Path Exclusion**. Submitted to International Conference on Learning Representations (ICLR) 2026
- [U] J. Kim, M. Seo, **M. Song**, S. Shin, J. Kim. **To Make Each Account Count: Exploring Credential Data Breach Threats through Victim-driven Analysis**. Submitted to IEEE Transactions on Information Forensics and Security (TIFS)

EXPERIENCE

- **S2W [🌐]** July 2022 - Feb 2023
South Korea
Research Intern @ AI Team
 - **Illicit drug jargon detection:** Analyzed illicit drug-related discussions and developed an LLM-based content moderation framework, independently capturing contextual and lexical characteristics.
- **KAIST DI Lab** Jan 2020 - June 2020
South Korea
Undergraduate Research Intern
 - **Big data mining with covid-19 dataset**
- **KAIST DM Lab** July 2019 - Aug 2019
South Korea
Undergraduate Research Intern
 - **Abnormal node detection in bipartite network via butterfly counting**

HONORS AND AWARDS

- **2nd Prize, 2023 Cybersecurity Paper Competition** 2023
Korean Association of Cybersecurity Studies (KACS)
 - Graph-based Deep Learning Framework for Credential Stuffing Risk Prediction
- **4th Prize, 2023 Cybersecurity Paper Competition** 2023
Korean Association of Cybersecurity Studies (KACS)
 - Delexicalized Distant Supervision for Illicit Drug Jargon Detection
- **4th Prize, 2023 Cybersecurity Paper Competition** 2023
Korean Association of Cybersecurity Studies (KACS)
 - Understanding the Occurrence and Impact of Credential Data Breach
- **Cum Laude** 2021
Korea Advanced Institute of Science and Technology (KAIST)
- **Academic Achievement Award: Salutatorian** 2019 Spring
Korea Advanced Institute of Science and Technology (KAIST)
- **Dean's List** 2019 Spring
Industrial and Systems Engineering (ISysE, KAIST)

LANGUAGES

Korean (Native), English (Fluent)