Jaehan Kim

Mail | LinkedIn | Google Scholar | Homepage

last update: January 2026

Research Interest

AI Security (LLM Safety), AI for Cybersecurity

Education

- **Korea Advanced Institute of Science and Technology (KAIST)** *March 2022 - February 2026 (expected)*
  Ph.D. (Candidate) in Electrical Engineering, Network and System Security Lab (Advisor: Seungwon Shin) Daejeon, South Korea
  Dissertation: Securing Efficient Techniques for Large Language Model Development (to appear)
- **Korea Advanced Institute of Science and Technology (KAIST)** *March 2020 - February 2022*
  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 2020*
  B.S. in Electrical Engineering, Minor in Computer Science Daejeon, South Korea

Publications [C]: Conference, [J]: Journal, [U]: Under Review

[C] **J. Kim**, M. Song, S. Shin, S. Son. **SafeMoE: Safe Fine-Tuning for MoE LLMs by Aligning Harmful Input Routing**. *The Fourteenth International Conference on Learning Representations (ICLR 2026) (to appear)*

[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)*

[C] **J. Kim**, S.H. Na, M. Song, S. Shin, S. Son. **MoEvil: Poisoning Experts to Compromise the Safety of Mixture-of-Experts LLMs**. *Proceedings of the 41th Annual Computer Security Applications Conference (ACSAC 2025)* ***(Distinguished Paper Award)***

[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, 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] 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 Security 2025)*

[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] M. Seo, M. You, **J. Kim**, T. Park, S. Shin, J. Kim. **MUFFLER: Secure Tor Traffic Obfuscation with Dynamic Connection Shuffling and Splitting**. *2025 IEEE International Conference on Computer Communications (INFOCOM 2025)*

[C] G. Park, **J. Kim**, J. Choi, J. Kim. **CryptoGuard: Lightweight Hybrid Detection and Prevention of Host-Based Cryptojackers**. *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security (ASIACCS 2025)*

[C] S. Kim, S.H. Na, **J. Kim**, S. Shin, H. Choi. **AVXProbe: Enhancing Website Fingerprinting with Side-Channel Assisted Kernel-Level Traces**. *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security (ASIACCS 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 (S&P 2024)*

[C] M. You, J. Nam, H. Seo, M. Seo, **J. Kim**, D. Choi, S. Shin. **HardWhale: A Hardware-Isolated Network Security Enforcement System for Cloud Environments**. *2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS 2024)*

[J] M. You, M. Seo, **J. Kim**, S. Shin, J. Nam. **Hyperion: Hardware-Based High-Performance and Secure System for Container Networks**. *IEEE Transactions on Cloud Computing 2024 (TCC 2024)*

[J] M. Seo*, **J. Kim***, M. You, S. Shin, J. Kim. **gShock: A GNN-based Fingerprinting System for Permissioned Blockchain Networks over Encrypted Channels**. *IEEE Access 2024 *equally contributed*

[J] S. Lee*, **J. Kim***, M. Seo, S.H. Na, S. Shin, J. Kim. **CENSor: Detecting Illicit Bitcoin Operation via GCN-Based Hyperedge Classification**. *IEEE Access 2024 *equally contributed*

[C] M. Seo, **J. Kim**, E. Marin, M. You, T. Park, S. Lee, S. Shin, J. Kim. **Heimdallr: Fingerprinting SD-WAN Control-Plane Architecture via Encrypted Control Traffic**. *Proceedings of the 38th Annual Computer Security Applications Conference (ACSAC 2022)*

[J] M. You, Y. Kim, **J. Kim**, M. Seo, S. Son, S. Shin, S. Lee. **FuzzDocs: An Automated Security Evaluation Framework for IoT**. *IEEE Access 2022*

[J] M. You, **J. Kim**, S. Shin. **Revisiting Security Landscape of Docker Hub Container**. *The Journal of Korean Institute of Communications and Information Sciences 2022*

[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 2022 *equally contributed*

[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 Elsevier Computers & Security*

[U] S. Song, D. Lee, **J. Kim**, J. Choi, J. Kim. **Assessing Hallucination in Large Language Models for Cyber Threat Intelligence: A First Measurement Study**. *Submitted to Elsevier Engineering Applications of Artificial Intelligence (EAAI)*

Experience

- **S2W [🌐]** *Dec 2020 - Feb 2021*
  Research Intern @ NLP Team South Korea
  - **Cyber Security Event Detection System**: Proposed a deep learning–based system for extracting security-related events from real-time, unstructured, and noisy data across platforms such as SNS, blogs, and the dark web.
- **SK hynix [🌐]** *Dec 2018 - Feb 2019*
  Software Engineer Intern @ NAND/Solution PE Team South Korea
  - **Storage Plan for Edge Computing Systems:** Developed a future plan for adopting appropriate storage interfaces based on the workload demands of emerging technologies such as AI.
- **KAIST NSS Lab** *Jun 2018 - Dec 2019*
  Undergraduate Research Intern South Korea
  - Developed a CLI extension for security assessment framework for software-defined networks.

Awards

- **Distinguished Paper Award** *2025*
  Annual Computer Security Applications Conference (ACSAC)
  - MoEvil: Poisoning Expert to Compromise the Safety of Mixture-of-Experts LLMs
- **4th Prize, 2025 Cybersecurity Paper Competition** *2025*
  Korean Association of Cybersecurity Studies (KACS)
  - Poisoning Expert to Compromise the Safety of Mixture-of-Experts LLMs
- **4th Prize, 2024 Cybersecurity Paper Competition** *2024*
  Korean Association of Cybersecurity Studies (KACS)
  - LLM Backdoor Defense within the Parameter-Efficient Fine-Tuning Paradigm
- **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
- **4th Prize, 29th Samsung Humantech Paper Awards** *2023*
  Samsung Electronics
  - Heimdallr: Fingerprinting SD-WAN Control-Plane Architecture via Encrypted Control Traffic
- **4th Prize, 2021 Cybersecurity Paper Competition** *2021*
  Korea Institute of Information Security and Cryptology (KIISC)
  - Fingerprinting Distributed SD-WAN Control-Plane Architecture via Encrypted Control Traffic