Jaechul Roh

jrohsc.github.io · Github · Google Scholar
+1 (470) 915 - 1137 · jroh@umass.edu

EDUCATION

University of Massachusetts Amherst — **September 2023 – Present**
Ph.D. in Computer Science — Amherst, Massachusetts, USA
Advisor: Prof. Amir Houmansadr
GPA: 4.0/4.0

Hong Kong University of Science and Technology — **September 2017 – May 2023**
B.Eng. in Computer Engineering, School of Engineering — Clear Water Bay, Hong Kong
Final Year Thesis Advisor: Prof. Jun Zhang
*2 years of Compulsory Korean Military Duty

RESEARCH INTERESTS

My research is centered on **Privacy & Security in AI** and **Trustworthy ML**. I am investigating the trustworthiness of multi-modal generative models across various domains, including text, audio, and image modalities, under the supervision of Prof. Amir Houmansadr. I am currently working on privacy and security of AI agents at Brave as a research intern with Dr. Ali Shahin Shamsabadi.

PUBLICATIONS

Preprints

1. **OverThink: Slowdown Attacks on Reasoning LLMs**
   Abhinav Kumar, **Jaechul Roh**, Ali Naseh, Marezna Karpinska, Mohit Iyyer, Amir Houmansadr, and Eugene Bagdasarian
   Preprint at arXiv
   [paper] [code]

2. **FameBias: Embedding Manipulation Bias Attack in Text-to-Image Models**
   Jaechul Roh*, Andrew Yuan*, Jinsong Mao*
   *Equal Contribution**
   Preprint at arXiv
   [paper]

3. **Understanding (Un)Intended Memorization in Text-to-Image Generative Models**
   Ali Naseh, **Jaechul Roh**, Amir Houmansadr
   Preprint at arXiv
   [paper]

Conference

1. **Multilingual and Multi-Accent Jailbreaking of Audio LLMs**
   Jaechul Roh, Virat Shejwalkar, Amir Houmansadr
   COLM 2025
   [paper]

2. **Backdooring Bias (B^2) into Stable Diffusion Models**
   Ali Naseh, **Jaechul Roh**, Eugene Bagdasarian, Amir Houmansadr
   USENIX Security '25
   [paper] [code]

3. **OSLO: One-Shot Label-Only Membership Inference Attacks**
   Yuefeng Peng, **Jaechul Roh**, Subhransu Maji, Amir Houmansadr
   NeurIPS 2024
   [paper]

4. **Memory Triggers: Unveiling Memorization in Text-To-Image Generative Models through Word-Level Duplication**
   Ali Naseh, **Jaechul Roh**, Amir Houmansadr
   The 5th AAAI Workshop on Privacy-Preserving Artificial Intelligence
   [paper]

5. **Robust Smart Home Face Recognition under Starving Federated Data**
   Jaechul Roh, Yajun Fang
   IEEE International Conference on Universal Village (IEEE UV2022)
   Oral Presentation
   [paper][code][slides][video]

6. **MSDT: Masked Language Model Scoring Defense in Text Domain**
   Jaechul Roh, Minhao Cheng, Yajun Fang
   IEEE International Conference on Universal Village (IEEE UV2022)
   Oral Presentation
   [paper][code][slides][video]

7. **Impact of Adversarial Training on the Robustness of Deep Neural Networks**
   Jaechul Roh
   2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE)
   [paper][code]

INVITED TALKS

Google Speech Technologies Group **July 2025**
Paper presentation Google DeepMind

- Presented our *"Multilingual and Multi-Accent Jailbreaking of Audio LLMs"* paper to the NFM Reading Group led by the Speech Technologies Group at Google DeepMind.
  [slides]

WORK EXPERIENCE

Brave Software **June 2025 – September 2025**
Research Intern, Supervisor: Ali Shahin Shamsabadi London, United Kingdom (Remote)

- Working on privacy & security of AI agents.

Super Chain AI (Conard International) **June 2021 – August 2021**
NLP Engineer Intern Kowloon Bay, Hong Kong

- In charge of topic modeling and semantic analysis based on customer reviews and assigning specific semantics to the topics extracted.
- Competitors' analysis through web-scrapping customer reviews from other drop-shipping websites.

Military Service at Head Quarter of 12th Infantry Division **July 2018 – March 2020**
Sergeant of Republic of Korea Army Injae, Kang Won Do, Republic of Korea

- Officer Administrative Clerk Specialist
- Squad Leader of the Head Quarter

PROFESSIONAL SERVICES

Program Committee Member (Reviewer)

- Main conferences: ICLR (2025)

SKILLS / LANGUAGES

Programming Language: Python, C++
Languages: Korean (Native), English (Native), Chinese (Fluent)