# Jaechul Roh

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

## EDUCATION

**University of Massachusetts Amherst** **September 2023 – May 2028 (expected)**
*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 focuses on the **Privacy & Security of AI models and agents**. Under the supervision of Prof. Amir Houmansadr, I study the trustworthiness of multimodal generative models across audio, text, and vision domains, while also collaborating closely with Prof. Eugene Bagdasarian on several projects and publications. Most recently, I completed a Summer Research Internship at Brave Software under the guidance of Dr. Ali Shahin Shamsabadi, where our work centered on advancing privacy in AI web agents.

## PUBLICATIONS

### Under Review

1. **SPILLage: Agentic Oversharing on the Web**
   **Jaechul Roh**, Eugene Bagdasarian, Hamed Haddadi, Ali Shain Shamsabadi
   *Under Review*

2. ***Bob's Confetti*: Phonetic Memorization Attacks in Music and Video Generation**
   **Jaechul Roh**, Zachary Novack, Yuefeng Peng, Niloofar Mireshghallah, Taylor Berg-Kirkpatrick, Amir Houmansadr
   *Preprint at arXiv (Under Review)*
   [paper] [demo page] [video]

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

4. **World-To-Image: Grounding Text-to-Image Generation with Agent-Driven World Knowledge**
   Moohyun Son, Jintaek Oh, Sunbin Mun, **Jaechul Roh**, Sehyun Choi
   *Preprint at arXiv (Under Review)*
   [paper]

5. **Throttling Web Agents Using Reasoning Gates**
   Abhinav Kumar, **Jaechul Roh**, Ali Naseh, Amir Houmansadr, Eugene Bagdasarian
   *Preprint at arXiv (Under Review)*
   [paper] [code] [demo page]

### 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]

### Preprints

1. **Chain-of-Code Collapse: Reasoning Failures in LLMs via Adversarial Prompting in Code Generation**
   **Jaechul Roh**, Varun Gandhi, Shivani Anilkumar, Arin Garg
   *Preprint at arXiv*
   [paper] [code]

2. **R1dacted: Investigating Local Censorship in DeepSeek's R1 Language Model**
   Ali Naseh, Harsh Chaudhari, **Jaechul Roh**, Mingshi Wu, Alina Oprea, Amir Houmansadr
   *Preprint at arXiv*
   [paper]

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

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

### Magazine Article

1. **Can You Trust What AI Hears (and Says)?**
   **Jaechul Roh**
   *XRDS: Crossroads, The ACM Magazine for Students 31 (4), 34-39*
   [article]

## INVITED TALKS & ACADEMIC SERVICES

**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]

**Program Committee Member (Reviewer)**

- **Main conference**: ICLR (2025, 2026)
- **Workshop**: COLM XLLM-Workshop (2025)

## 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, Supervisor: Cat Yung* 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

## SKILLS / LANGUAGES

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