Hanbin Hong

Research Scientist @ TikTok (ByteDance)
hanbin.hong1@bytedance.com
https://youbin2014.github.io

Work Experience

Research Scientist, TikTok — San Jose, California
Data Trust and Safety — June 2024 – Present
Focus: Visual Language Model Post Training and Video Moderation

AI Security Research Scientist Intern, ByteDance — San Jose, California
AI Security Team — July 2023 – June 2024
Focus: Large Language Model Security, Adversarial Red-Teaming, and Detection

Research Assistant, University of Connecticut — Storrs, Connecticut
Department of Computer Science and Engineering — September 2022 – May 2025
Focus: Machine Learning Security, Adversarial Machine Learning, and Privacy

Research Assistant, Illinois Institute of Technology — Chicago, Illinois
Department of Computer Science — January 2021 – May 2022
Focus: Security in AI and Applied Cryptography

Education

Ph.D., University of Connecticut — Storrs, Connecticut
Department of Computer Science and Engineering — September 2022 – Aug 2025
Area: Computer Science

Ph.D. Student, Illinois Institute of Technology — Chicago, Illinois
Department of Computer Science — January 2021 – May 2022
Area: Computer Science

Research Internship, Rochester Institute of Technology — Rochester, New York
Golisano College of Computing and Information Sciences — September 2019 – December 2020
Area: Computer Vision

Graduate School, Xi'an Jiaotong University — Xi'an, China
School of Economics and Finance — September 2018 – July 2019
Major: Financial Engineering

Xi'an Jiaotong University — Xi'an, China
Qian Xuesen School — September 2014 – July 2018
Major: Honor Science Program (Physics)

Research Experience

LLMs for Security
Research Scientist Intern @ ByteDance — January 2025 – Present

- Fine-tuning LLMs (e.g., ChatGLM, LLaMA 3, Qwen2.5) using LoRA/P-tuning for DDoS and bot detection, aiming to achieve industrial-grade anomaly detection performance. Expected to be deployed in June.
- Training reinforcement learning-based reasoning LLMs for improved detection accuracy and interpretability.

LLM Prompt Security Analysis System
Research Scientist Intern @ ByteDance September 2024 – December 2024

- Led analysis on a 2M jailbreak prompt dataset.
- Designed a Auto-labeling system that achieve 97% acc on jailbreak detection.
- Developed a feature extraction system to efficiently identify 50+ heuristic features for characterizing jailbreak attacks efficiently.
- Surveyed and anaylysis over 250 papers in prompt security. Building a python platform for testing 20+ SOTA jailbreak attacks and 10+ defense on 50+ online and offline LLMs.
- **Outcome**: A Systematization of Knowledge (SoK) paper is in preparation.

Interpretable Attack Prompts Analysis & ***Mitigation for LLMs***
Research Scientist Intern @ ByteDance July 2024 – August 2024

- Independently developed a semantic analysis platform (Python) based on pre-trained Sparse Autoencoders (SAE) and LLMs to analyze key semantics contributing to jailbreak success.
- Designed platform components for data/model prep, annotation (HarmBench, Llama3 Guard, GPT-4o), multi-strategy feature selection, SAE-based feature validation, and interpretation modules for semantic visualization.
- **Outcome**: Developed a Python library for LLM vulnerability analysis.

Optimizable Text-to-Image Generation for Active Learning Without Data
Research Assistant @ University of Connecticut January 2024 – March 2024

- Proposed using text-to-image models to generate data for vision model training, achieving end-to-end model training from textual descriptions.
- Leveraged active learning acquisition strategies to optimize text embedding for generating more informative data samples.
- **Outcome**: Paper submitted to CVPR 2025.

Certifiable Attack — A New Paradigm of Adversarial Black-box Attack
Research Assistant @ University of Connecticut December 2022 – May 2023

- Established a new paradigm of adversarial black-box attacks called certifiable black-box attacks, modeling adversarial examples as random variables in the input space following a noise distribution (e.g., Gaussian) to guarantee attack success probability.
- Designed a randomized parallel query mechanism for attackers to inject noise into queries and aggregate feedback, along with a novel framework to craft and refine the adversarial distribution.
- Demonstrated that certifiable attacks are undetectable by state-of-the-art detection methods (e.g., Blacklight) and can bypass strong randomized pre(post)-processing defenses. Ensured the attack success probability of crafted adversarial examples before verification on the target model.
- **Outcome**: Paper published at ACM CCS 2024.

Certified Robustness via Randomized Smoothing
Research Assistant @ University of Connecticut & Illinois Institute of Technology January 2021 – November 2022

- Extended randomized smoothing theories to universal settings, including arbitrary noise distributions and perturbation types, to enhance certified adversarial robustness in machine learning.
- Introduced anisotropic noise to randomized smoothing, establish theories for certification and technique to optimize noise, providing a new dimension for randomized smoothing.
- Advanced randomized smoothing in the NLP domain by certifying four types of adversarial attacks in text with novel theoretical guarantees.

- **Outcomes**: Papers published at ECCV 2022 and S&P 2024; paper submitted to SatML 2025.

PUBLICATION

Hanbin Hong, Ali Payani, Ashish Kundu, Binghui Wang, Yuan Hong. Universally Amplifying Randomized Smoothing for Certified Robustness with Anisotropic Noise. (Conditionally Accepted by CSF), 2026

Jieren Deng, **Hanbin Hong**, Aaron Palmer, Xin Zhou, Jinbo Bi, Kaleel Mahmood, Yuan Hong, Derek Aguiar. Certifying Adapters: Enabling and Enhancing the Certification of Classifier Adversarial Robustness. (IJCNN), 2025

Shuya Feng, Meisam Mohammady, **Hanbin Hong**, Shenao Yan, Ashish Kundu, Binghui Wang and Yuan Hong. Universally harmonizing differential privacy mechanisms for federated learning: Boosting accuracy and convergence. ACM Conference on Data and Application Security and Privacy (CODASPY), 2025

Hanbin Hong, Xinyu Zhang, Binghui Wang, Zhongjie Ba, Yuan Hong. Certifiable Black-Box Attacks with Randomized Adversarial Examples: Breaking Defenses with Provable Confidence. Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS), 2024

Shenao Yan, Shen Wang, Yue Duan, **Hanbin Hong**, Kiho Lee, Doowon Kim, Yuan Hong. An LLM-Assisted Easy-to-Trigger Backdoor Attack on Code Completion Models: Injecting Disguised Vulnerabilities against Strong Detection. (USENIX Security), 2024

Hanbin Hong, Shenao Yan, Shuya Feng, and Yuan Hong. GALOT: Generative Active Learning via Optimizable Zero-shot Text-to-image Generation. Under review.

Xinyu Zhang, **Hanbin Hong**, Yuan Hong, Peng Huang, Binghui Wang, Zhongjie Ba, and Kui Ren. Text-CRS: A Generalized Certified Robustness Framework against Textual Adversarial Attacks. 45th IEEE Symposium on Security and Privacy (IEEE S&P), 2024

Han Wang, **Hanbin Hong**, Li Xiong, Zhan Qin, and Yuan Hong. PrivLBS: Local Differential Privacy for Location-Based Services with Staircase Randomized Response. Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS), 2022.

Hanbin Hong, Binghui Wang, and Yuan Hong. UniCR: Universally Approximated Certified Robustness via Randomized Smoothing. Computer Vision European Conference (ECCV), 2022

Hanbin Hong, Yuan Hong, and Yu Kong. An Eye for an Eye: Defending against Gradient-based Attacks with Gradients. arXiv preprint arXiv:2202.01117, 2022.

Hanbin Hong, Wentao Bao, Yuan Hong, and Yu Kong. Privacy Attributes-aware Message Passing Neural Network for Visual Privacy Attributes Classification. 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021.

Junwen Chen, Haiting Hao, **Hanbin Hong**, and Yu Kong, RIT-18: A Novel Dataset for Compositional Group Activity Understanding, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020.

RESEARCH INTERESTS

AI Security, LLM, Generative AI, Data Synthesis, Adversarial Attacks, Adversarial Robustness, Certifiable Robustness, and Privacy-preserving Machine Learning.

TEACHING EXPERIENCE

CSE 3140: Cybersecurity Lab — Fall 2023, Spring 2024, Fall 2024

Services

Program Committee

Association for the Advancement of Artificial Intelligence (AAAI)	2022, 2023

Reviewer

International Conference on Machine Learning (ICML)	2022, 2023, 2024
Computer Vision and Pattern Recognition Conference (CVPR)	2023, 2024
International Conference on Learning Representations (ICLR)	2024
International Conference on Computer Vision (ICCV)	2023
Conference on Neural Information Processing Systems (NeurIPS)	2022
International World Wide Web Conference (Web)	2023
IEEE Transactions on Dependable and Secure Computing (TDSC)	2023
IEEE Internet of Things Journal (IoTJ)	2023
European Conference on Computer Vision (ECCV)	2022

External Reviewer

ACM Conference on Computer and Communications Security (CCS)	2022, 2023
USENIX Security Symposium (USENIX)	2022, 2023, 2024
International Conference on Autonomous Agents and Multiagent Systems (AAMAS)	2023
International Symposium on Research in Attacks, Intrusions and Defenses (RAID)	2022
Special Interest Group on Knowledge Discovery and Data Mining (KDD)	2022

Software competencies

Programming: Python, Matlab, C\C++, Java
Software Library: PyTorch, Keras, Tensorflow
Operating Systems: Windows, Linux, macOS

Honors and Awards

Predoctoral Fellowship in Computer Science & Engineering	2025
Anthony W.DeSio Endowed Fellowship in Computer Science & Engineering	2024
General Electric Graduate Fellowship for Excellence	2023, 2024
Predoctoral Fellowship with Excellent Research Award	2023
Synchrony Financial Cybersecurity Graduate Fellowship	2023
Student Travel Grant for CCS'22	2022
Certificate of Honors Graduate awarded by Qian Xuesen School	2018
2nd Class Zhufeng Scholarship	2017
Si Yuan Scholarship awarded by Xi'an Jiaotong University	2016
The 2nd Prize in China Undergraduate Mathematical Contest in Modeling	2015