Jihye Choi

jihye@cs.wisc.edu

https://jihyechoi77.github.io

Interests

Trustworthy machine learning ecosystem in the wild:
- Reliable LLM-powered agents with Retrieval Augmented Generation that adapt to evolving knowledge and user needs, with a focus on healthcare applications.
- Interpretable and robust decision-making with Foundation Models under distribution shifts.
- Security and privacy in machine learning, including jailbreaking LLMs, privacy-preserving federated learning, membership inference exploiting memorization, and adversarial robustness.

Education

University of Wisconsin-Madison, Madison, WI — 2019 - 2025 (expected)
Ph.D., Computer Sciences (Advisor: Prof. Somesh Jha)

Carnegie Mellon University, Pittsburgh, PA — 2016 - 2018
M.S., Electrical and Computer Engineering

Yonsei University, Seoul, Korea — 2013 - 2016
B.S., Electrical and Computer Engineering (Early graduation with Highest Honors)

Publications

(* denotes equal contribution)

Multi-user Personalization with Collaborative LLM-powered Agents
Christine P Lee*, Jihye Choi*, Bilge Mutlu
Under Submission to CHI 2025

SLVR: Securely Leveraging Client Validation for Robust Federated Learning
Jihye Choi, Rahul Rachuri, Ke Wang, Somesh Jha, Yizhen Wang
Under Submission to IEEE S&P 2025

Adaptive Concept Bottleneck for Foundation Models Under Distribution Shifts
Jihye Choi, Jayaram Raghuram, Yixuan Li, Somesh Jha
Under Submission to ICLR 2025, ICML 2024 Workshop on Foundation Models in the Wild

MALADE: Orchestration of LLM-powered Agents with Retrieval Augmented Generation for Pharmacovigilance
Jihye Choi*, Nils Palumbo*, Prasad Chalasani, Matthew M. Engelhard, Somesh Jha, Anivarya Kumar, David Page
MLHC 2024

PRP: Propagating Universal Perturbations to Attack Large Language Model Guard-Rails
Ashish Hooda*, Neal Mangaokar*, Jihye Choi, Shreyas Chandrashekaran, Kassem Fawaz, Somesh Jha, Atul Prakash
ACL 2024 (Main)

Identifying and Mitigating the Security Risks of Generative AI
Clark Barrett, Brad Boyd, Ellie Burzstein, Nicholas Carlini, Brad Chen, Jihye Choi, ..., Dawn Song, Ankur Taly, Diyi Yang
Foundations and Trends in Privacy and Security, Vol. 6: No. 1, pp 1-52

Why Train More? Effective and Efficient Membership Inference via Memorization
Jihye Choi, Varun Chandrasekaran, Shruti Tople, Somesh Jha
Preprint, Under Submission to CVPR 2025

Concept-based Explanations for Out-Of-Distribution Detectors
Jihye Choi, Jayaram Raghuram, Ryan Feng, Jiefeng Chen, Somesh Jha, Atul Prakash
ICML 2023

Stratified Adversarial Robustness with Rejection
Jiefeng Chen*, Jayaram Raghuram*, Jihye Choi, Xi Wu, Yingyu Liang, Somesh Jha
(* equal contribution)
ICML 2023

Rethink Diversity in Deep Learning Testing
Zi Wang, Jihye Choi, Ke Wang, Somesh Jha
Preprint

Revisiting Adversarial Robustness of Classifiers With a Reject Option
Jiefeng Chen*, Jayaram Raghuram*, Jihye Choi, Xi Wu, Yingyu Liang, Somesh Jha
AAAI 2022 Workshop (**Oral Presentation and Best Paper Award**)

Stochastic Doubly Robust Gradient
Kanghoon Lee*, Jihye Choi*, Moonsoo Cha, Jung-Kwon Lee and Tae Yoon Kim
arXiv 2018

Data-driven Approach to Aesthetic Enhancement
Jihye Choi, Sungjoon Koh, Jongwoo Kwack, Yonghun Kwon, Hyunjung Shim
SPIE Electronic Imaging 2016

Work Experience

University of Wisconsin-Madison — Madison, WI
Research Assistant with Prof. Somesh Jha — Since Jun 2020

Visa Research, Identity and AI Security — Palo Alto, CA
PhD Intern with Dr. Yizhen Wang, Dr. Ke Wang, Dr. Rahul Rachuri — Summer 2023, 2022

CyLab, Carnegie Mellon University — Pittsburgh, PA
Research Assistant with Prof. Lujo Bauer — May 2017 - Mar 2019

T-Brain, SK Telecom — Seoul, Korea
Research Intern with Dr. Kanghoon Lee — Apr 2018 - Sep 2018

Vision & Learning Lab., Yonsei University — Seoul, Korea
Undergraduate Research Assistant with Prof. Hyunjung Shim — Mar 2015 - Aug 2016

Service

- Program Committee of Deep Learning Security and Privacy Workshop with IEEE S&P 2025
- Reviewer for ICML 2024-2025, NeurIPS 2022-2024, SaTML 2024 (external), AISTATS 2025, CVPR 2025
- Student organizer of GenAI Risk Workshop 2023