

Anna Yoo Jeong Ha

Ph.D. Candidate, Computer Science — University of Chicago
annaha.net • GoogleScholar • annaha@uchicago.edu • LinkedIn

Research Summary

AI security/safety researcher with hands-on experience in generative model security and adversarial ML. Combines research and engineering to examine the technical mechanisms of generative AI systems, assess robustness, and develop methods that improve the reliability and safety of AI systems. Experienced in taking projects end-to-end, from problem formulation and rapid prototyping to large-scale experiments and validation on deployed systems.

Education

University of Chicago — **Ph.D., Computer Science** Expected 2028

Advisors: Prof. Ben Y. Zhao and Prof. Heather Zheng (SandLab). GPA: 4.0 / 4.0.

Supported by Daniels Research Fellowship

Korea University — **M.S., Electrical and Computer Engineering** Mar. 2021 – Feb. 2023

Korea University — **B.S., Mechanical Engineering** Mar. 2017 – Feb. 2021

Awards & Patents

Distinguished Paper Award, CCS 2024 Oct. 2024

Video Processing System and Video Processing Method Using Split Learning US-11915477-B2

Experience

Ph.D. AI Researcher — SandLab, University of Chicago Jul 2023 – Present

Whiteout: protecting LLMs from leaking private data *in submission*

- **Protected deployed LLMs** from regurgitating personal data by breaking the model’s association between a person and their true data.
- Tested across 7 LLMs (3B–20B params) with as few as 10 samples per protected fact, while preserving utility and safety near baseline — eliminating the protection–utility tradeoff that breaks unlearning and refusal baselines.
- Validated end-to-end on **OpenAI’s production fine-tuning API** (GPT-4o-mini), navigating real deployment constraints including content-moderation rejections.
- **Red-teamed the defense** against 9 adaptive attacks — GCG jailbreaks, quantization, relearning, and a custom erasure attack using the exact defense data; resulted in 0% attack success.

Provenance: black-box lineage attribution for generative models *USENIX Security 2026*

- Built a **black-box forensic system** that determines whether a text-to-image diffusion model was trained from scratch or fine-tuned from another model — and identifies the parent — using only query access (no weights, architecture, or training data).
- Designed the statistical core by probing models with detailed prompts, extracting multi-scale CLIP features from outputs, and comparing distributions via Jensen-Shannon divergence under a calibrated hypothesis test.
- Attributed lineage with **100% accuracy across 8 base models** (SDXL, FLUX, PixArt- Σ) and their fine-tuned variants; stable across LoRA vs. full fine-tuning and robust to evasion via weight noise and adversarial perturbation.
- Deployed the system **in the wild** against a live commercial image service.

Content authenticity: human and machine detection of AI-generated content *ACM CCS 2024*

- Led a large-scale measurement study of how people and automated detectors distinguish human art from AI-generated images across 5 generative models and 7 visual styles; the work earned a **CCS Distinguished Paper Award** and informed platform policy.

Autonomous Mobility Researcher — Korea University

Jul. 2019 – Dec. 2022

- Built an **autonomous driving** system for ultra-dense, hyper-connected vehicle networks; optimized multimodal sensing and network scheduling to move large volumes of sensor data efficiently.
- Trained **deep reinforcement learning** agents to plan courses of action over hierarchical attack representation models in a large-scale, distributed network environment.
- Developed a **distributed / federated** deep-learning algorithm that trains across three institutions' data without sharing raw records; contributed to a patented split-learning pipeline.

Selected Publications

Full publication list (11 papers, including IEEE Access, Nature Scientific Reports) available on Google Scholar.

Anna Yoo Jeong Ha, Ronik Bhaskar, Haitao Zheng, Ben Y. Zhao. Mitigating Private Data Leakage in LLMs with *Whiteout*. *In Submission*. 2026.

Anna Yoo Jeong Ha, Wenxin Ding, Stanley Wu, Shawn Shan, Haitao Zheng, Ben Y. Zhao. Identifying Provenance of Generative Text-to-Image Models. *In 35th USENIX Security Symposium (USENIX Security)*. 2026.

Stanley Wu, Ronik Bhaskar, **Anna Yoo Jeong Ha**, Shawn Shan, Haitao Zheng, Ben Y. Zhao. On the Feasibility of Poisoning Text-to-Image AI Models via Adversarial Mislabeling. *Conference on Computer and Communications Security (CCS)*. 2025.

Anna Yoo Jeong Ha, Josephine Passananti, Ronik Bhaskar, Shawn Shan, Reid Southen, Haitao Zheng, Ben Y. Zhao. Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?. *Conference on Computer and Communications Security (CCS)*. 2024. **Distinguished Paper Award**

Service

Teaching

TA, CS35800 Advanced Adversarial Machine Learning, UChicago CS	Winter 2026
TA, CS25800 Adversarial Machine Learning, UChicago CS	Spring 2025
TA, Intro to CS, Korea University CS	Fall 2021

Reviewer

International Conference on Machine Learning (ICML)	2026
Neural Information Processing Systems (NeurIPS)	2026

Leadership

PSDSC, Board Member of UChicago PSDSC	Spring 2023
---------------------------------------	-------------

Technical Skills

Programming Languages: Python, C++, JavaScript, SQL

ML/AI Frameworks: PyTorch, TensorFlow, HuggingFace, Scikit-learn, NumPy, SciPy, Pandas

AI Security: adversarial ML & red-teaming (jailbreaks/GCG, quantization, relearning, evasion attacks), ML-system robustness & vulnerability analysis, data-leakage & privacy defenses, threat modeling

ML Methods: fine-tuning LLMs & diffusion models (SFT, LoRA), automated model evaluation pipelines, feature extraction & embeddings, statistical hypothesis testing, reinforcement learning, federated learning

Languages: Native in English and Korean