# Sihui (Sophie) Dai

: sophiedai98@gmail.com

: https://github.com/dash29

## Overview

Applied researcher working on AI safety, automated adversarial evaluation, and robust machine learning. My current work focuses on developing production safeguards for LLM applications and methods for systematically stress-testing models and defenses. During my Ph.D. at Princeton, I studied robustness to unforeseen, adaptive, and evolving adversaries. I am broadly interested in building AI systems that remain reliable as attacks, models, and deployment environments change.

## Work Experience

- **Capital One** — New York, NY
  *Applied Researcher* — *Apr 2025 – Present*
  - Develop production LLM safeguards and safety evaluation methods for moderation, input-level safety, and prompt-injection detection use cases.
  - Build automated adversarial evaluation and red-teaming workflows to stress-test LLM applications and safeguard classifiers against adaptive, black-box, and obfuscated inputs.
  - Improve safeguard development through synthetic data generation, evaluation pipeline design, robustness testing, dataset decontamination, and data-quality analysis.
  - Translate evaluation findings into model-development decisions, release documentation, and cross-functional workflows with engineering and annotation partners.
  - Mentor research work on adaptive adversarial evaluation, including project scoping, experiment design, and research direction.
- **IBM Research** — Yorktown Heights, NY
  *Research Scientist Intern* — *May–Aug 2024; May–Aug 2023*
  - Developed memory-augmented LLM architectures with IBM Trustworthy AI for episodic memory control, knowledge editing, forgetting, and context generalization.
  - Contributed to *Larimar: Large Language Models with Episodic Memory Control*, published at ICML 2024.
  - Extended memory-augmented LLM research toward multi-hop reasoning and reasoning-in-a-haystack tasks, contributing to MemReasoner and ImReasoner.
- **California Institute of Technology** — Pasadena, CA
  *Summer Undergraduate Research Fellow* — *Jun–Sep 2019*
  - Researched recurrent generative feedback mechanisms for neural network architectures inspired by feedback in the visual cortex.
  - Contributed to *Neural Networks with Recurrent Generative Feedback*, published at NeurIPS 2020.
  - Received best poster award in Caltech's Summer Undergraduate Research Fellowship program for this work.
- **RealNetworks** — New York, NY
  *Quality Assurance Intern* — *Jul–Sep 2017*
  - Designed and executed regression tests for spam text-message recognition software across new software builds.
  - Developed test-automation scripts to improve QA efficiency and provided robustness-focused feedback to the development team.

## Education

- **Princeton University** — Princeton, NJ
  *Ph.D., Electrical and Computer Engineering* — *Sep 2020 – Apr 2025*
  - Advisor: Prateek Mittal. Research focus: adversarial machine learning, robustness to unforeseen and evolving adversaries, and adaptive test-time threats.
- **California Institute of Technology** — Pasadena, CA
  *B.S. in Computer Science; Minor in Information and Data Sciences* — *Sep 2016 – Jun 2020*
  - Advisors: Anima Anandkumar and Yisong Yue. Research in neuro-inspired neural networks and Bayesian optimization.

## Publications

**Preprints:**

- Decomposing the Delta: What Do Models Actually Learn from Preference Pairs? 2026
  Chia-Hsuan Lee, Mingyang Zhou, Renkun Ni, Zelei Cheng, **Sihui Dai**, Supriyo Chakraborty, Shixiong Zhang, Sambit Sahu, William Campbell
  *Preprint*

- MemReasoner: A Memory-augmented LLM Architecture for Multi-hop Reasoning 2024
  **Sihui Dai***, Ching-Yun Ko*, Payel Das*, Georgios Kollias, Subhajit Chaudhury, Aurelie Lozano
  *Preprint*
- Position: Towards Resilience against Adversarial Examples 2024
  **Sihui Dai**, Chong Xiang, Tong Wu, Prateek Mittal
  *Preprint*

**Conference Papers:**

- ImReasoner: Improving Memory-based Language Models for Reasoning-in-a-Haystack Tasks 2026
  Ching-Yun Ko, Payel Das, **Sihui Dai**, Georgios Kollias, Subhajit Chaudhury, Aurelie C. Lozano, Pin-Yu Chen
  *Association for Computational Linguistics (ACL)*
- Adapting to Evolving Adversaries with Regularized Continual Robust Training 2025
  **Sihui Dai***, Christian Cianfarani*, Arjun Nitin Bhagoji, Vikash Sehwag, Prateek Mittal
  *International Conference on Machine Learning (ICML)*
- Larimar: Large Language Models with Episodic Memory Control 2024
  Payel Das, Subhajit Chaudhury, Elliot Nelson, Igor Melnyk, Sarath Swaminathan, **Sihui Dai**, Aurélie Lozano, Georgios Kollias, Vijil Chenthamarakshan, Jiří Navrátil, Soham Dan, Pin-Yu Chen
  *International Conference on Machine Learning (ICML)*
- PatchCURE: Improving Certifiable Robustness, Model Utility, and Computation Efficiency of Adversarial Patch Defenses 2024
  Chong Xiang, Tong Wu, **Sihui Dai**, Jonathan Petit, Suman Jana, Prateek Mittal
  *USENIX Security Symposium*
- Characterizing the Optimal 0-1 Loss for Multi-class Classification with a Test-time Attacker (Spotlight) 2023
  **Sihui Dai***, Wenxin Ding*, Arjun Nitin Bhagoji, Daniel Cullina, Ben Y. Zhao, Haitao Zheng, Prateek Mittal
  *Conference on Neural Information Processing Systems (NeurIPS)*
- MultiRobustBench: Benchmarking Robustness Against Multiple Attacks 2023
  **Sihui Dai**, Saeed Mahloujifar, Chong Xiang, Vikash Sehwag, Pin-Yu Chen, Prateek Mittal
  *International Conference on Machine Learning (ICML)*
- Formulating Robustness Against Unforeseen Attacks 2022
  **Sihui Dai**, Saeed Mahloujifar, Prateek Mittal
  *Conference on Neural Information Processing Systems (NeurIPS)*
- Robust Learning Meets Generative Models: Can Proxy Distributions Improve Adversarial Robustness? 2022
  Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, **Sihui Dai**, Chong Xiang, Mung Chiang, Prateek Mittal
  *International Conference on Learning Representations (ICLR)*
- Neural Networks with Recurrent Generative Feedback 2020
  Yujia Huang, James Gornet, **Sihui Dai**, Zhiding Yu, Tan Nguyen, Doris Y. Tsao, Anima Anandkumar
  *Conference on Neural Information Processing Systems (NeurIPS)*

**Workshop Papers:**

- What Do Safety-Aligned LLMs Learn From Mixed Compliance Demonstrations? 2026
  **Sihui Dai**, Mann Patel
  *ICML Hypothesis Testing Workshop*
- Parameterizing Activation Functions for Adversarial Robustness 2022
  **Sihui Dai**, Saeed Mahloujifar, Prateek Mittal
  *IEEE Deep Learning Security (DLS) Workshop*
- Robustness from Perception 2021
  Saeed Mahloujifar, Chong Xiang, Vikash Sehwag, **Sihui Dai**, Prateek Mittal
  *ICLR Workshop on Security and Safety in Machine Learning Systems*
- Multi-task Bayesian Optimization via Gaussian Process Upper Confidence Bound 2020
  **Sihui Dai**, Jialin Song, Yisong Yue
  *ICML Workshop on Real World Experiment Design and Active Learning*

## Teaching

Served as assistant instructor in the following courses:

**Princeton University:**

- COS/ECE 432: Information Security — Spring 2022, 2023

**California Institute of Technology:**

- CS/CNS/EE/IDS 165: Foundations of Machine Learning and Statistical Inference — Winter 2020
- CS156a: Learning Systems — Fall 2019
- CS1: Introduction to Computer Programming — Fall 2017, 2018
- CS11: Computer Language Lab (C++ track) — Spring 2018

## Honors and Awards

- NSF Graduate Research Fellowship (2020–2023)
- Bhansali Family Prize in Computer Science, Caltech (2020), awarded on the basis of outstanding research in computer science

## Peer Review

- Conference on Neural Information Processing Systems (NeurIPS) 2021, 2022, 2023, 2024, 2025, 2026
- International Conference on Machine Learning (ICML) 2022, 2023, 2024, 2025
- International Conference on Learning Representations (ICLR) 2022, 2023, 2024, 2025
- Transactions on Machine Learning Research (TMLR) 2022, 2023
- Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 2022, 2023
- International Conference on Artificial Intelligence and Statistics (AISTATS) 2022
- Conference on Uncertainty in Artificial Intelligence (UAI) 2024
- Asian Conference on Machine Learning (ACML) 2024