Chawin Sitawarin

✉ chawin.sitawarin@gmail.com
https://chawins.github.io/

Machine Learning Security & Privacy Researcher

Experiences

2025–present **Research Scientist**, *Google DeepMind*.
AI Security & Privacy

2024–2025 **Postdoctoral Researcher**, *Meta*, Central Applied Science, Privacy-Preserving Machine Learning.
Memorization, privacy, and copyright risks in generative AI

2018–2024 **PhD in Computer Science**, *UC Berkeley*.
LLM security and adversarial robustness, advised by Prof. David Wagner

2014–2018 **BSE in Electrical Engineering (High Honor)**, *Princeton University*.
Certificate in Applications of Computing

Research Interests

I am broadly interested in security and privacy aspects of machine learning. My recent works are on **jailbreak** attacks and **prompt injection** defenses on large language models; my older works are on the **adversarial robustness** of machine learning algorithms. Currently, the problems I am excited about are:

(1) Better evaluation of memorization, privacy, and copyright risks (2) Safety, security, and prompt injection defenses

Selected Publications

2025 **The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against LLM Jailbreaks and Prompt Injections**, *M. Nasr*, N. Carlini*, C. Sitawarin*, S. V. Schulhoff*, J. Hayes, M. Ilie, J. Pluto, S. Song, H. Chaudhari, I. Shumailov, A. Thakurta, K. Y. Xiao, A. Terzis, F. Tramèr*, Preprint (Under Submission) paper.

2025 **How Much Do Language Models Memorize?**, *J. Morris, C. Sitawarin, C. Guo, N. Kokhlikyan, E. Suh, A. M. Rush, K. Chaudhuri, S. Mahloujifar*, Preprint (Under Submission) paper.

2025 **Stronger Universal and Transfer Attacks by Suppressing Refusals**, *D. Huang, A. Shah, A. Araujo, D. Wagner, C. Sitawarin*, NAACL 2025 paper.

2025 **Mark My Words: Analyzing and Evaluating Language Model Watermarks**, *J. Piet, C. Sitawarin, V. Fang, N. Mu, D. Wagner*, SaTML 2025 paper code.

2025 **Vulnerability Detection with Code Language Models: How Far Are We?**, *Y. Ding, Y. Fu, O. Ibrahim, C. Sitawarin, X. Chen, B. Alomair, D. Wagner, B. Ray, Y. Chen*, ICSE 2025 paper code.

2025 **StruQ: Defending Against Prompt Injection with Structured Queries**, *S. Chen, J. Piet, C. Sitawarin, D. Wagner*, USENIX Security 2025 paper code.

2024 **PAL: Proxy-Guided Black-Box Attack on Large Language Models**, *C. Sitawarin, N. Mu, D. Wagner, A. Araujo*, Preprint paper code.

2024 **Jatmo: Prompt Injection Defense by Task-Specific Finetuning**, *J. Piet*, M. Alrashed*, C. Sitawarin, et al.*, ESORICS 2024 paper code.

2024 **PubDef: Defending against Transfer Attacks from Public Models**, *C. Sitawarin, J. Chang*, D. Huang*, W. Altoyan, D. Wagner*, ICLR 2024 (Poster) paper code.

2023 **Preprocessors Matter! Realistic Decision-Based Attacks on Machine Learning Systems**, *C. Sitawarin, F. Tramèr, N. Carlini*, ICML 2023 (Poster) paper code.

2023 **Part-Based Models Improve Adversarial Robustness**, *C. Sitawarin, K. Pongmala, Y. Chen, N. Carlini, D. Wagner*, ICLR 2023 (Poster) paper code.

2022 **Demystifying the Adversarial Robustness of Random Transformation Defenses**, *C. Sitawarin, Z. Golan-Strieb, D. Wagner*, ICML 2022 and AAAI-22 AdvML Workshop (Best Paper) paper code.

2021 **SAT: Improving Adversarial Training via Curriculum-Based Loss Smoothing**, *C. Sitawarin, S. Chakraborty, D. Wagner*, AISec 2021 (co-located with CCS) paper.

2020 **Minimum-Norm Adversarial Examples on k-NN and k-NN-Based Models**, *C. Sitawarin, D. Wagner*, Deep Learning and Security Workshop (IEEE S&P 2020) paper.

2018 **On the Robustness of Deep k-Nearest Neighbors**, *C. Sitawarin, D. Wagner*, Deep Learning and Security Workshop (IEEE S&P 2019) paper.

2018 **DARTS: Deceiving Autonomous Cars with Toxic Signs**, *C. Sitawarin, A. N. Bhagoji, A. Mosenia, M. Chiang, P. Mittal*, Preprint paper.

2018 **Enhancing Robustness of Machine Learning System via Data Transformations**, *A. N. Bhagoji, D. Cullina, C. Sitawarin, P. Mittal*, CISS 2018 paper.

Prior Internships

Summer 2022 **Google**, *Sunnyvale CA*, Research intern.
Developed a defense against transfer adversarial attacks for a malware classification task with a pair of public client-side and secret server-side models. Hosted by Ali Zand and David Tao.

Fall 2021–Spring 2022 **Google**, *Remote*, Part-time student researcher.
Developed new query-based adversarial attack and model-stealing attack against a black-box image preprocessing and recognition pipeline. Hosted by Nicholas Carlini.

Summer 2021 **Nokia Bell Labs**, *Remote*, Research intern.
Investigated relationships between causality and robustness in machine learning, focusing on leveraging causal relationship to improve robustness and generalization to unseen corruptions. Hosted by Anwar Walid.

Summer 2019 **IBM Research**, *Yorktown Heights NY*, Research intern.
Studied the effectiveness of existing defenses against adversarial examples from a perspective of concentration bound and improved adversarial training through optimization techniques. Hosted by Supriyo Chakraborty.

Awards & Grants

2023 **Outstanding Graduate Student Instructor Award** *Teaching award*

2022 **Google-BAIR Commons Project** *Research grant*

2021–2022 **Center for Long-Term Cybersecurity (CLTC)** *Research grant*

2021 **Microsoft-BAIR Commons Project** *Research grant*

2018 **Phi Beta Kappa** *Academic honor society*

2018 **Sigma Xi** *Scientific research honor society*

2017 **The P. Michael Lion III Fund** *Summer research funding for Princeton engineering students*

2016 **Tau Beta Pi** *Engineering honor society*

2016 **Shapiro Prize for Academic Excellence** *Academic award at Princeton University*

2013 **King's Scholarship** *Prestigious scholarship awarded by Thai government for pursuing a bachelor's degree*

Services

Reviewer, ICLR '24, '25 | ICML '22 (top reviewer), '24, '25 | NeurIPS '22, '23, '24. '25 | TPAMI '24 | IEEE S&P '26 | AISec '22, '23, '24, '25.