

Chawin Sitawarin

☎ 510-423-2880
✉ chawins@meta.com
📄 <https://chawins.github.io/>

*Postdoctoral Researcher @ Meta
Machine Learning Security & Privacy Researcher*

Education

- 2018–2024 **PhD in Computer Science**, *UC Berkeley*, Berkeley CA.
Advisor: Professor David Wagner | GPA 3.86
- 2014–2018 **BSE in Electrical Engineering (High Honor)**, *Princeton University*, Princeton NJ.
Cumulative GPA: 3.90, Departmental GPA: 3.95 | Certificate in Applications of Computing

Research Interests

I am broadly interested in security and privacy aspects of machine learning with a recent focus on large language models. Most of my past works are in the domain of adversarial machine learning, particularly adversarial examples and robustness of machine learning algorithms.

Selected Publications

- 2024 **Jatmo: Prompt Injection Defense by Task-Specific Finetuning**, *J. Piet**, *M. Alrashed**, *C. Sitawarin*, et al., ESORICS 2024, [Paper](#), [Code](#).
- 2024 **Defending against Transfer Attacks from Public Models**, *C. Sitawarin*, *J. Chang**, *D. Huang**, *W. Altoyan*, *D. Wagner*, ICLR 2024 (Poster), [Paper](#), [Code](#).
- 2024 **SPDER: Semiperiodic Damping-Enabled Object Representation**, *K. Shah*, *C. Sitawarin*, ICLR 2024 (Poster), [Paper](#).
- 2024 **StruQ: Defending Against Prompt Injection with Structured Queries**, *S. Chen*, *J. Piet*, *C. Sitawarin*, *D. Wagner*, Preprint (Under Submission), [Paper](#).
- 2023 **Mark My Words: Analyzing and Evaluating Language Model Watermarks**, *J. Piet*, *C. Sitawarin*, *V. Fang*, *N. Mu*, *D. Wagner*, Preprint (Under Submission), [Paper](#), [Code](#).
- 2023 **OODRobustBench: Benchmarking and Analyzing Adversarial Robustness Under Distribution Shift**, *L. Li*, *Y. Wang*, *C. Sitawarin*, *M. Spratling*, Preprint (Under Submission), [Paper](#).
- 2023 **REAP: A Large-Scale Realistic Adversarial Patch Benchmark**, *N. Hingun**, *C. Sitawarin**, *J. Li*, *D. Wagner*, ICCV 2023 (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 (Short Presentation) and AAAI-22 AdvML Workshop (Best Paper), [Paper](#), [Code](#).
- 2021 **Adversarial Examples for k -Nearest Neighbor Classifiers Based on Higher-Order Voronoi Diagrams**, *C. Sitawarin*, *E. M. Kornaropoulos*, *D. Song*, *D. Wagner*, NeurIPS 2021 (Poster), [Paper](#), [Code](#).
- 2021 **Improving the Accuracy-Robustness Trade-Off for Dual-Domain Adversarial Training**, *C. Sitawarin*, *A. Sridhar*, *D. Wagner*, Workshop on Uncertainty & Robustness in Deep Learning (ICML 2021), [Paper](#), [Code](#).
- 2021 **Mitigating Adversarial Training Instability with Batch Normalization**, *A. Sridhar*, *C. Sitawarin*, *D. Wagner*, Workshop on Security and Safety in Machine Learning Systems (ICLR 2021), [Paper](#).
- 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](#).
- 2019 **Analyzing the Robustness of Open-World Machine Learning**, [V. Schwag](#), [A. N. Bhagoji](#), [L. Song](#), [C. Sitawarin](#), [D. Cullina](#), [M. Chiang](#), and [P. Mittal](#), AISeC 2019 (co-located with CCS), [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 **Not All Pixels are Born Equal: An Analysis of Evasion Attacks under Locality Constraints**, [V. Schwag](#), [C. Sitawarin](#), [A. N. Bhagoji](#), [A. Mosenia](#), [M. Chiang](#), [P. Mittal](#), CCS 2018 Poster, [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 Employment

- Summer 2022 **Google**, *Sunnyvale*, Research Intern.
Evaluate and mitigate machine learning security risks in a practical setting where a pair of public client-side and secret server-side models is deployed for a malware detection task. Hosted by Ali Zand and David Tao.
- Fall 2021 - **Google**, *Remote*, Student researcher (part-time).
- Spring 2022 Developed threat model and appropriate evaluation for adversarial robustness in new and practical settings (e.g., dynamic models, black-box model recovery). Hosted by Nicholas Carlini.
- Summer 2021 **Nokia Bell Labs**, *Remote*, Summer research intern.
Investigated relationships between causality and robustness in machine learning, focusing on leveraging causal relationship to improve robustness and generalization to unseen attacks/corruptions. Mentored by Anwar Walid.
- Summer 2019 **IBM Research**, *Yorktown Heights NY*, Summer research intern.
Studied the effectiveness of existing defenses against adversarial examples from a perspective of concentration bound and improved adversarial training through optimization techniques. Mentored by Supriyo Chakraborty.

Awards & Honors

- 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*

Activities and Services

- Program Committee**, *AISeC '22, '23*.
- Reviewer**, *ICLR '24 | ICML '22 (top reviewer), '24 | NeurIPS '22, '23 | BAND workshop '23*.
- 2019–2023 **DARE: Diversifying Access to Research in Engineering**, *Mentor*, I have mentored multiple students from DARE, a program that promotes diversity in EECS undergraduate research.
- 2018–2020 **CSGSA**, *Treasurer*, Computer Science Graduate Student Assembly at UC Berkeley.
- 2018–2019 **Security Seminar**, *Organizer*, Organized a biweekly seminar on security and privacy at UC Berkeley, hosting outside speakers from both industry and academia.