Eric Wong

Phone: +1 (339) 223-7159
Email: exwong@upenn.edu
Website: https://www.cis.upenn.edu/~exwong/
Google Scholar: https://scholar.google.com/citations?user=pWnTMRkAAAAJ

Last updated: 2026/09/25

Briefly

How can we make sure that deep learning models are actually doing what we want them to do? My research interests are centered around foundations of reliable machine learning systems: understanding, debugging, and guaranteeing the behavior of data-driven models. I created the first provable defenses that guarantee robustness to adversarial examples and real-world specifications, and currently work on securing modern foundation models. I am also interested in explaining models with provable certificates and scientific applications ranging from surgery to cosmology.

Education

University of Pennsylvania, Assistant Professor — Philadelphia, PA, *2022-current*

Massachusetts Institute of Technology, Post-Doctoral Associate — Cambridge, MA, *2020-2022*
Advisor: Aleksander Mądry

Carnegie Mellon University, Ph. D. in Machine Learning — Pittsburgh, PA, *2015-2020*
Thesis: Provable, structured, and efficient methods for robustness of deep networks to adversarial examples; SCS Dissertation Award — Honorable Mention
Advisor: Zico Kolter

Carnegie Mellon University, B. S. in Computer Science — Pittsburgh, PA, *2011-2015*
Double major in Mathematics, minor in Machine Learning

Work experience

2019-2020 **Bosch Center for Artifical Intelligence (Renningen, Germany and Pittsburgh, PA)**
Created a virtual sensor based on neural networks for a fuel injection system in truck engines; formally verified the worst-case error of the system under conversative estimates of physical sensor noise.

2012-2015 **CERT Program (Pittsburgh, PA)**
Migrated secure coding rules from POSIX to C11; analyzed security reports for Java android applications; developed an analysis toolfor security vulnerabilities in source code.

Awards

2025	**AI2050 Early Career Award** *Towards Robust Generative AI with Adaptive Risk Evaluations*, Schmidt Sciences
2025	**NSF Early Career Award** *CAREER: Certified Explanations for Trustworthy Artificial Intelligence*, NSF
2024	**Amazon Research Award (AWS AI)** *Adversarial Manipulation of Prompting Interfaces*, Amazon
2023	**Area Chair Award (Interpretability and Analysis of Models for NLP)** *Faithful Chain-of-Thought Reasoning*, IJCNLP-AACL Conference
2020	**SCS Dissertation Award – Honorable Mention** *Provable, structured, and efficient methods for robustness of deep networks to adversarial examples*, Carnegie Mellon University
2020	**Siebel Scholar Fellowship** Carnegie Mellon University
2017	**Best Defense Paper** *Provable defenses against adversarial examples via the convex outer adversarial polytope*, NeurIPS 2017 ML & Security Workshop
2013	**Summer Undergraduate Research Fellowship** Carnegie Mellon University

Publications

NeurIPS 2026	**SuperActivators: Only the Tail of the Distribution Contains Reliable Concept Signals** Cassandra Goldberg, Chaehyeon Kim, Adam Stein, Eric Wong
NeurIPS 2026	**Position-aware eXplanation: A Model-Agnostic Framework for Positional Attributions** Chaehyeon Kim, Gary E. Weissman, Eric Wong
NeurIPS 2026	**The Road to Generalizable Neuro-Symbolic Learning Should be Paved with Foundation Models** Adam Stein, Aaditya Naik, Neelay Velingker, Mayur Naik, Eric Wong
COLM 2026	**Detecting Safety Violations Across Many Agent Traces** Adam Stein, Davis Brown, Hamed Hassani, Mayur Naik, Eric Wong
IJCARS 2026	**Sum-of-Checks: structured reasoning for surgical safety with large vision-language models** Weiqiu You, Cassandra Goldberg, Amin Madani, Daniel A. Hashimoto, Eric Wong
ICML 2026	**Interpretability Can Be Actionable** Hadas Orgad, Fazl Barez, Tal Haklay, Isabelle Lee, Marius Mosbach, Anja Reusch, Naomi Saphra, Byron C Wallace, Sarah Wiegreffe, Eric Wong, Ian Tenney, Mor Geva

ICLR 2026	**Missingness Bias Calibration in Feature Attribution Explanations** Shailesh Sridhar, Anton Xue, Eric Wong
ACL 2026	**NSF-SciFy: Mining the NSF Awards Database for Scientific Claims** Delip Rao, Weiqiu You, Eric Wong, Chris Callison-Burch
NeurIPS 2025	**Probabilistic Stability Guarantees for Feature Attributions** Helen Jin, Anton Xue, Weiqiu You, Surbhi Goel, Eric Wong
NeurIPS 2025	**CTSketch: Compositional Tensor Sketching for Scalable Neurosymbolic Learning** Seewon Choi, Alaia Solko-Breslin, Rajeev Alur, Eric Wong
NeurIPS 2025	**Once Upon an Input: Reasoning via Per-Instance Program Synthesis** Adam Stein, Neelay Velingker, Mayur Naik, Eric Wong
EMNLP 2025	**Probabilistic Soundness Guarantees in LLM Reasoning Chains** Weiqiu You, Anton Xue, Shreya Havaldar, Delip Rao, Helen Jin, Chris Callison-Burch, Eric Wong
EMNLP 2025	**Adaptively profiling models with task elicitation** Davis Brown, Prithvi Balehannina, Helen Jin, Shreya Havaldar, Hamed Hassani, Eric Wong
EMNLP Findings 2025	**Artifact or Flaw? Rethinking Prompt Sensitivity in Evaluating LLMs** Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, Yao Qin
DMLR 2025	**The FIX Benchmark: Extracting Features Interpretable to eXperts** Helen Jin, Shreya Havaldar, Chaehyeon Kim, Anton Xue, Weiqiu You, Helen Qu, Marco Gatti, Daniel A. Hashimoto, Bhuvnesh Jain, Amin Madani, Masao Sako, Lyle Ungar, Eric Wong
ACL 2025	**Towards Style Alignment in Cross-Cultural Translation** Shreya Havaldar, Adam Stein, Eric Wong, Lyle Ungar
ICML 2025	**Sum-of-Parts Models: Faithful Attributions for Groups of Features** Weiqiu You, Helen Qu, Marco Gatti, Bhuvnesh Jain, Eric Wong
ICML 2025	**DOLPHIN: A Programmable Framework for Scalable Neurosymbolic Learning** Aaditya Naik, Jason Liu, Claire Wang, Amish Sethi, Saikat Dutta, Mayur Naik, Eric Wong
NAACL-Findings 2025	**Avoiding Copyright Infringement via Machine Unlearning** Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, Eric Wong
ICLR 2025	**Logicbreaks: A Framework for Understanding Subversion of Rule-based Inference** Anton Xue, Avishree Khare, Rajeev Alur, Surbhi Goel, Eric Wong
TMLR 2025	**SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks** Alexander Robey, Eric Wong, Hamed Hassani, George J. Pappas
SaTML 2025	**Jailbreaking Black Box Large Language Models in Twenty Queries** Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, Eric Wong

IJCNLP-AACL, 2025	**Defending Large Language Models against Jailbreak Attacks via Semantic Smoothing** Jiabao Ji, Bairu Hou, Alexander Robey, George J. Pappas, Hamed Hassani, Yang Zhang, Eric Wong, Shiyu Chang
NeurIPS 2024	**AR-Pro: Counterfactual Explanations for Anomaly Repair with Formal Properties** Xiayan Ji, Anton Xue, Eric Wong, Oleg Sokolsky, Insup Lee
eBioMedicine	**Crowd-sourced machine learning prediction of long COVID using data from the National COVID Cohort Collaborative** Timothy Bergquist, Johanna Loomba, Emily Pfaff, Fangfang Xia, Zixuan Zhao, Yitan Zhu, Elliot Mitchell, Biplab Bhattacharya, Gaurav Shetty, Tamanna Munia, Grant Delong, Adbul Tariq, Zachary Butzin-Dozier, Yunwen Ji, Haodong Li, Jeremy Coyle, Seraphina Shi, Rachael V. Philips, Andrew Mertens, Romain Pirracchio, Mark van der Laan, John M. Colford Jr., Alan Hubbard, Jifan Gao, Guanhua Chen, Neelay Velingker, Ziyang Li, Yinjun Wu, Adam Stein, Jiani Huang, Zongyu Dai, Qi Long, Mayur Naik, John Holmes, Danielle Mowery, Eric Wong, Ravi Parekh, Emily Getzen, Jake Hightower, Jennifer Blase
NeurIPS 2024	**Data-Efficient Learning with Neural Programs** Alaia Solko-Breslin, Seewon Choi, Ziyang Li, Neelay Velingker, Rajeev Alur, Mayur Naik, Eric Wong
ICML 2024	**Towards Compositionality in Concept Learning** Adam Stein, Aaditya Naik, Yinjun Wu, Mayur Naik, Eric Wong
ICML 2024	**DISCRET: Synthesizing Faithful Explanations For Treatment Effect Estimation** Yinjun Wu, Mayank Keoliya, Kan Chen, Neelay Velingker, Ziyang Li, Emily J Getzen, Qi Long, Mayur Naik, Ravi B Parikh, Eric Wong
NeurIPS 2024	**JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models** Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, Eric Wong
ICLR 2024, Tiny Papers (Oral)	**Evaluating Groups of Features via Consistency, Contiguity, and Stability** Chaehyeon Kim, Weiqiu You, Shreya Havaldar, Eric Wong
ICLR 2024	**SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation** Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, Sijia Liu
CVPR 2024	**Initialization Matters for Adversarial Transfer Learning** Andong Hua, Jindong Gu, Zhiyu Xue, Nicholas Carlini, Eric Wong, Yao Qin
OOPSLA 2024	**TorchQL: A Programming Framework for Integrity Constraints in Machine Learning** Aaditya Naik, Adam Stein, Yinjun Wu, Eric Wong, Mayur Naik
EMNLP 2023	**Comparing Styles across Languages** Shreya Havaldar, Matthew Pressimone, Eric Wong, Lyle Ungar

NeurIPS 2023	**Stability Guarantees for Feature Attributions with Multiplicative Smoothing** Anton Xue, Rajeev Alur, Eric Wong
ICLR 2023, Tiny Papers	**TopEx: Topic-based Explanations for Model Comparison** Shreya Havaldar, Adam Stein, Eric Wong, Lyle Ungar
ICML 2023	**Do Machine Learning Models Learn Statistical Rules Inferred from Data?** Aaditya Naik, Yinjun Wu, Mayur Naik, Eric Wong
DLSP 2023 Keynote	**Adversarial Prompting for Black Box Foundation Models** Natalie Maus*, Patrick Chao*, Eric Wong, Jacob Gardner
IJCNLP-AACL, 2023	**Faithful Chain-of-Thought Reasoning** Qing Lyu*, Shreya Havaldar*, Adam Stein*, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, Chris Callison-Burch
CVPR 2023	**A data-based perspective on transfer learning** Saachi Jain*, Hadi Salman*, Alaa Khaddaj*, Eric Wong, Sung Min Park, Aleksander Madry
ICLR 2022	**Missingness bias in model debugging** Saachi Jain*, Hadi Salman*, Pengchuan Zhang, Vibhav Vineet, Sal Vemprala, Aleksander Madry
CVPR 2022	**Certified patch robustness via smoothed vision transformers** Hadi Salman*, Saachi Jain*, Eric Wong*, Aleksander Madry
OJCS 2022	**DeepSplit: Scalable verification of deep neural networks via operator splitting** Shaoru Chen*, Eric Wong*, J. Zico Kolter, Mahyar Fazlyab
ICML 2021 (Oral)	**Leveraging Sparse Linear Layers for Debuggable Deep Networks** Eric Wong*, Shibani Santurkar*, Aleksander Madry
ICLR 2021	**Learning perturbation sets for robust machine learning** Eric Wong, J. Zico Kolter
ICML 2020	**Overfitting in adversarially robust deep learning** Leslie Rice*, Eric Wong*, J. Zico Kolter
IEEE IV 2020	**Neural network virtual sensors for fuel injection quantities with provable performance specifications** Eric Wong, Tim Schneider, Joerg Schmitt, Frank R. Schmidt, J. Zico Kolter
ICLR 2020	**Fast is better than free: revisiting adversarial training** Eric Wong*, Leslie Rice*, J. Zico Kolter
ICML 2020	**Adversarial robustness against the union of multiple perturbation models** Pratyush Maini, Eric Wong, J. Zico Kolter
ICML 2019	**Wasserstein adversarial examples** Eric Wong, Frank R. Schmidt, J. Zico Kolter

NeurIPS 2018	**Scaling provable adversarial defenses** Eric Wong, Frank R. Schmidt, Jan Hendrik Metzen, J. Zico Kolter
ICML 2018	**Provable defenses against adversarial examples via the convex outer adversarial polytope** Eric Wong, J. Zico Kolter
ICML 2017	**A Semismooth Newton Method for Fast, Generic Convex Programming** Alnur Ali*, Eric Wong*, J. Zico Kolter
AAAI 2015	**An SVD and Derivative Kernel Approach to Learning from Geometric Data** Eric Wong, J. Zico Kolter

Preprints

2026	**Stateful Online Monitoring Catches Distributed Agent Attacks** Davis Brown, Samarth Bhargav, Arav Santhanam, Kasper Hong, Ivan Zhang, Matan Shtepel, Steffi Chern, Alexander Robey, Eric Wong, Hamed Hassani
2026	**Detecting and Correcting Reference Hallucinations in Commercial LLMs and Deep Research Agents** Delip Rao, Eric Wong, Chris Callison-Burch
2026	**CAMEL: An ECG Language Model for Forecasting Cardiac Events** Neelay Velingker, Alaia Solko-Breslin, Mayank Keoliya, Seewon Choi, Jiayi Xin, Anika Marathe, Alireza Oraii, Rajat Deo, Sameed Khatana, Rajeev Alur, Mayur Naik, Eric Wong
2026	**Semantics-Preserving Evasion of LLM Vulnerability Detectors** Luze Sun, Alina Oprea, Eric Wong
2025	**T-FIX: Text-Based Explanations with Features Interpretable to eXperts** Shreya Havaldar, Helen Jin, Chaehyeon Kim, Anton Xue, Weiqiu You, Marco Gatti, Bhuvnesh Jain, Helen Qu, Daniel A Hashimoto, Amin Madani, Rajat Deo, Sameed Ahmed M. Khatana, Gary E. Weissman, Lyle Ungar, Eric Wong
2025	**Instruction Following by Boosting Attention of Large Language Models** Vitoria Guardieiro, Adam Stein, Avishree Khare, Eric Wong
2025	**Benchmarking Misuse Mitigation Against Covert Adversaries** Davis Brown, Mahdi Sabbaghi, Luze Sun, Alexander Robey, George J. Pappas, Eric Wong, Hamed Hassani
2025	**The Road to Generalizable Neuro-Symbolic Learning Should be Paved with Foundation Models** Adam Stein, Aaditya Naik, Neelay Velingker, Eric Wong
2023	**Rectifying Group Irregularities in Explanations for Distribution Shift** Adam Stein, Yinjun Wu, Eric Wong, Mayur Naik
2023	**In-context Example Selection with Influences** Tai Nguyen, Eric Wong

2022 **When does bias transfer in transfer learning**
Hadi Salman*, Saachi Jain*, Andrew Ilyas*, Logan Engstrom*, Eric Wong, Aleksander Madry

Grants

2025-11-13 – 2026-11-13 **Collaborative Schmidt Proposal**
$30k, Schmidt Science collaboration grant with Chaowei Xiao (JHU)

2025-07-23 – 2026-07-23 **Collaborative Schmidt Proposal**
$30k, Schmidt Science collaboration grant with Aditi Raghunathan (CMU)

2025-11-25 – 2026-05-25 **Principled Inference-Time Guardrails via Projected Latent Rule Following**
$270k, UK AISI Challenge Fund

2025-06-01 – 2027-05-31 **Towards Robust Generative AI with Adaptive Risk Evaluations**
$450k, AI2050 Early Career Fellowship, Schmidt Sciences, PI

2025-06-01 – 2030-05-31 **CAREER: Certified Explanations for Trustworthy Artificial Intelligence**
$675k, NSF, PI

2024-12-01 – 2027-05-31 **Harnessing Artificial Intelligence and Language Modeling for Enhancing Innovation and Evaluating Research Claims (HAILMEIER-C)**
$5.9M, DARPA, Co-PI with Chris Callison-Burch, Hannaneh Hajishirzi, Andrew Head, Peter Jansen, Chinedum Osuji, Yulia Tsvetkov, Duncan Watts

2026-02-02 – 2027-07-31 **TIGER: Trustable Information Generation and Explanation Resilience**
$1.6M, IARPA, Co-PI with Rene Vidal, Chris Callison-Burch, Hamed Hassani, Mark Yatskar, Rama Chellappa, Vishal Patel

2024-05-01 – 2025-04-30 **Preventing Complications with Transparent Surgical AI Assistants**
$100K, ASSET-IBI, PI with Daniel Hashimoto

2024-07-01 – 2028-06-30 **Safe and Explainable AI-enabled Decision Making for Personalized Treatment**
$6.85M, ARPA-H, Co-PI with Rajeev Alur, Rajat Deo, Sameed Ahmed M. Khatana, Qi Long, Mayur Naik, Ravi Parikh, Gary Weissman

2023-10-01 – 2027-09-30 **SLES: SPECSRL: Specification-guided Perception-enabled Conformal Safe Reinforcement Learning**
$1.5M, NSF, Co-PI with Rajeev Alur, Osbert Bastani & Dinesh Jayaraman

2024-05-01 – 2025-04-30 **Adversarial Manipulation of Prompting Interfaces**
$70K (+$50K compute), Amazon Research Award, PI

2023-10-01 – 2027-09-30 **SHF: Medium: Scallop: A Neurosymbolic Programming Framework for Combining Logic with Deep Learning**
$1.2M, NSF, Co-PI with Mayur Naik & Rajeev Alur

Media

2026 **Using AI to help predict cardiac arrests** (Penn Today)
News article on our research for forecasting cardiac events from ECG signals
https://penntoday.upenn.edu/news/using-ai-help-predict-cardiac-arrests

2026 **We are losing the point of education** (The Daily Pennsylvanian)
News article on education in the age of AI and how courses like my class CIS 5200 (Machine Learning) are exploring new assessment strategies
https://www.thedp.com/article/2026/02/penn-education-point-ai

2026 **NBC10 Septa Story** (NBC10 Philadelphia)
News article on the use of AI cameras on SEPTA buses
https://www.nbcphiladelphia.com/investigations/ai-powered-cameras-on-septa-buses-have-led-to-thousands-of-tickets/4348922/

2025 **Experts warning about advancing AI voice technology being used in scams** (NBC10 Philadelphia)
News article on AI deepfakes
https://www.nbcphiladelphia.com/video/investigators/consumer/experts-warning-about-advancing-ai-voice-technology-being-used-in-scams/4327920/

2025 **Family says scammers used AI to impersonate South Philly woman's granddaughter** (NBC10 Philadelphia)
News article on AI phone scams
https://www.nbcphiladelphia.com/investigators/consumer/philadelphia-grandparent-scam-ai-theft-impersonation-elder-fraud/4254278/

2025 **Penn lab researches AI cyberbullying capabilities** (Penn Today)
News article about our cyberbullying research.
https://penntoday.upenn.edu/news/penn-seas-evaluating-large-language-models-cyberbullying-behavior

2024 **The Llama 3 Herd of Models** (Meta)
Meta technical report that used our PAIR red-teaming algorithm to develop their Llama3 models
https://arxiv.org/pdf/2407.21783

2024 **Gemini 1.5 Report** (Google Deepmind)
Google technical report that used our JailBreakBench evaluation to red-team Gemini models
https://arxiv.org/pdf/2403.05530

2024 **Anthropic Sleeper Agents** (Anthropic)
Google technical report that used our PAIR red-teaming algorithm to test the Claude family of models
https://arxiv.org/pdf/2401.05566

2023 **New method reveals how one LLM can be used to jailbreak another** (VentureBeat)
News article about our PAIR red-teaming algorithm.
https://venturebeat.com/ai/new-method-reveals-how-one-llm-can-be-used-to-jailbreak-another/

Invitations

2026	**A Mechanistic Theory of Safety: How Jailbreaking 1-Layer Transformers Taught us how to Steer LLMs** Seminar Speaker, Columbia Machine Learning and AI Seminar Series
2025	**Lessons from the Impossibility of Safety** Invited talk, University of Tubingen
2025	**Impossibility Results in Safety** Seminar speaker, Security & Privacy of LLMs Dagstuhl Seminar
2025	**Dynamic & Stateful Evals of Safety on the Frontier: What can Academics do?** Invited talk, Data in Generative Models, ICML 2025 Workshop
2025	**Explanations for Experts via Guarantees and Domain Knowledge: From Attributions to Reasoning** Invited talk, Actionable Interpretability, ICML 2025 Workshop
2025	**How Jailbreaking 1-Layer Transformers Taught us how to Steer LLMs** Invited talk, Methods and Opportunities at Small Scale (MOSS), ICML 2025 Workshop
2025	**Explanations for Experts** Seminar speaker, New Jersity Institute of Technology, AI & Data Science Stars Seminar Series
2025	**What does a Foundation Model (not) Know?** Keynote speaker, Towards Knowledgeable Foundation Models @ AAAI 2025 Workshop
2024	**Convincing Experts to (not) Trust ML Models** Seminar speaker, Cornell AI Seminar
2024	**Jailbreaking LLMs: Attack, Defense, and Theory** Seminar speaker, University of Maryland
2023	**Robustness of Adversarial Attacks for LLM** Distinguished speaker, Responsible Machine Learning Summit, UCSB
2023	**Adversarial Prompting: Return of the Adversarial Example** Keynote speaker, IEEE S&P 2023, 6th Deep Learning Security & Privacy Workshop
2023	**From Prompt Engineering to Prompt Science** Seminar speaker, Wayne State University
2022	**Robustness for the Real World** Invited talk, 6th Annual Conference on Information Sciences and Systems (CISS)
2022	**Debuggable Deep Networks** Invited talk, TrustML Young Scientist Seminar
2021	**Panel Discussion** Panelist, ATVA 2021 Workshop on Security and Reliability of Machine Learning

Patents

2024 **Methods, systems, and computer readable media for defending large language models (LLMs) against jailbreaking attacks** (18/907376)

2019 **Method, apparatus and computer program for generating robust automated learning systems and testing trained automated learning systems** (16/173698)

Teaching Experience

2026 **Mathematics of Machine Learning** - CIS 3333 (UPenn, Instructor)
Fourth iteration of the Mathematics of Machine Learning course, which has been successful as an undergraduate pipeline for theoretical ML research

2026 **Machine Learning** - CIS 5200 (UPenn, Instructor)
4th round teaching CIS 5200, with new oral assessments scaled to 220 students to foster learning in the presence of GenAI tools

2025 **Accelerating Research with Generative AI** - CIS 7000 (UPenn, Instructor)
New special topics course to prepare students for best research practices augmented with LLMs

2025 **Machine Learning** - CIS 5200 (UPenn, Instructor)
3rd round teaching CIS 5200

2024 **Mathematics of Machine Learning** - CIS 3333 (UPenn, Instructor)
Second iteration of the new Mathematics of Machine Learning course under an official course number as a SEAS mathematics elective.

2024 **Machine Learning** - CIS 5200 (UPenn, Instructor)
2nd round teaching CIS 5200, further consolidation and unification of the course material with the other AI faculty. 150 students.

2023 **Mathematics of Machine Learning** - CIS 3990 (UPenn, Instructor)
Created a new course that prepares undergraduates for technical research and a graduate level coursework in machine learning. Two students that took the course last semester are now doing ML theory research.

2023 **Machine Learning** - CIS 5200 (UPenn, Instructor)
Substantially overhauled and updated the Machine Learning course at UPenn to (a) fully autograded assignments using PennGrader, (b) brand new PyTorch-based programming assignments to replace dated Numpy notebooks, (c) expanded to a more balanced set of topics across all of Machine Learning (i.e. Duality/Lagrangian, MCMC, an entire theory module including PAC Learning & VC Theory, other learning paradigms like Online and Active Learning). 128 students.

2022 **Debugging Data & Models** - CIS 7000-005 (UPenn, Instructor)
Designed new special topics course in the seminar/lecture format on debugging machine learning (7000-005). Taught 25 enrolled students with average instructor/course scores of 3.47/4 and 3.54/4.

2016	**Practical Data Science** - 15-388/688 (CMU, TA) Designed new assignments, taught recitations, and prepared write-ups for the first iteration of CMU's Practical Data Science course (15-388/688). I was the head TA (out of two TAs) and managed over 300 enrolled students. Received 55 student reviews with an average rating of 4.85/5.
2016-2019	**Eberly Center for Teaching Excellence and Educational Innovation** - Teaching Seminars (CMU, Participant) Enrolled in teaching seminars at the Eberly Center in CMU to develop personal teaching skills; seminars include "Teaching Inclusively: Leveraging Diversity and Promoting Equity in Your Classroom" and "Helping Students Develop Mastery and Critical Thinking"
2015	**Advanced Introduction to Machine Learning** - 10-715 (CMU, TA) Taught recitations, held office hours, and created/graded assignments for the second iteration of the Advanced Introduction to Machine Learning course intended for doctoral students in CMU's Machine Learning Department.
2014	**Algorithm Design and Analysis** - 15-451 (CMU, TA) Taught recitations, conducted oral examinations/office hours, and graded assignments/exams for the computer science department's Algorithm Design and Analysis course (15-451) at CMU.
2014	**Pervasive and Mobile Computing Services** - 08-766/781 (CMU, TA) Held office hours and graded assignments/exams for the software engineering department's Pervasive and Mobile Computing Services course (08-766/781) at CMU.
2013	**Pervasive and Mobile Computing Services** - 08-766/781 (CMU, TA) Held office hours and graded assignments/exams for the software engineering department's Pervasive and Mobile Computing Services course (08-766/781) at CMU.
2013	**Mobile Development for iOS and Android** - 08-723 (CMU, TA) Held office hours and graded assignments/exams for the software engineering department's Mobile Deveopment course (08-723) at CMU.

Graduate Theses Supervised

Spring 2026 – Spring 2026	**Yushan Yan** (Masters) Independent study
Fall 2025 – Spring 2030	**Steffi Chern** (PhD) Thesis: Anticipated Spring 2030
Fall 2024 – Spring 2025	**Prithvi Balehannina** (Masters) Masters research project on LLM evaluations
Fall 2024 – Summer 2025	**Luze Sun** (Masters) Masters research project on LLM security
Fall 2024 – Spring 2026	**Cindy Xin** (Masters) Masters research project on AI-assisted discovery in cosmology

Spring 2024 – Summer 2025	**Rupkatha Hira** (Masters) Thesis: Anticipated Summer 2025
Fall 2024 – Spring 2029	**Cassandra Goldberg** (PhD) Thesis: Anticipated Spring 2029
Fall 2024 – Spring 2029	**Davis Brown** (PhD) Thesis: Anticipated Spring 2029, co-advised with Hamed Hassani
Fall 2024 – Spring 2029	**Vitoria Guardieiro** (PhD) Thesis: Anticipated Spring 2029
Spring 2024 – Spring 2026	**Adam Stein** (PhD) Thesis: On the Interface Between Foundation Models and Programs, co-advised with Mayur Naik
Fall 2023 – Spring 2028	**Chaehyeon Kim** (PhD) Thesis: Anticipated Spring 2028
Fall 2023 – Spring 2027	**Helen Jin** (PhD) Thesis: Anticipated Spring 2027
Summer 2023 – Spring 2025	**Anton Xue** (PhD) Thesis: Trustworthy Machine Learning: Verification, and Explanation
Spring 2023 – Spring 2026	**Weiqiu You** (PhD) Thesis: Structured Explanations for Scientific Domains Building on Foundation Models
Spring 2023 – Spring 2026	**Shreya Havaldar** (PhD) Thesis: Towards Explaining, Evaluating, and Mitigating Cultural Bias in LLMs, co-advised with Lyle Ungar
Fall 2022 – Spring 2024	**Ningyuan Li** (Masters) Independent study
Fall 2022 – Spring 2024	**Shailesh Sridhar** (Masters) Thesis: Controlling for Missingness Bias in Feature Attribution Evaluation
Fall 2022 – Spring 2024	**Tai Nguyen** (Masters) Thesis: Attribute in-context learning examples with influences

Undergraduate Projects Supervised

Spring 2026 – Spring 2026	**Angela Zhang** (Undergraduate) Undergraduate research project on scoring credible of reasoning sources
Spring 2025 – Spring 2026	**Siri Nellutla** (Undergraduate) Undergraduate research with VIPER on reasoning over material science documents
Fall 2024 – Spring 2025	**Lyuxin (David) Zhang** (Undergraduate) Undergraduate research project on in-context learning

Spring 2024 – Summer 2024	**Faraz Rahman** (Undergraduate) Undergraduate research on understanding features generated during diffusion
Spring 2024 – Spring 2024	**Dora Wu** (Undergraduate) Thesis: Image Generative Artificial Intelligence: Theory, Applications, and Outlook
Fall 2022 – Spring 2023	**Gideon Tesfaye** (Undergraduate) Practicum: Using deep learning to compose music

Penn Service

2026 – 2026	**ASSET Student Seminar** Oragnizer of student seminar centered around reading group and presentation skill-building activities
2026 – 2026	**PhD Admit Weekend** Organizer
2025 – 2026	**ASSET Seminar** Co-organizer for inviting and scheduling speakers for the ASSET Center seminar series
2025 – 2025	**PhD Admit Weekend** Organizer, Co-Lead with Andrew Head
2024 – 2025	**IDEAS Search Committee** Faculty search commiteee member
2024 – 2024	**PhD Admit Weekend** Organizer, Co-Lead with Andrew Head
2023 – 2025	**BSE in AI** Curriculum Committee
2023 – 2026	**ML+FM Seminar** Organizer of a seminar series for researchers in the areas of formal methods and machine learning. Averages 18 attendees weekly.
2023 – 2023	**Adhoc Computing Cluster Committee** Commitee Member
2023 – 2023	**PhD Admit Weekend** Organizer, Co-Lead with Andrew Head
2022 – 2025	**Locust Cluster** Test driver for the SEAS cluster and working with CETS to iron out scalability of the system.

DEI Service

2025	**Mentorship Workshop** (Learning Theory Alliance) How-to: Using AI tools in research and learning.
2024-2026	**WiML@PennCIS** (CIS) Organizer of a new DEI event to help women in CIS form a community. Averages 20 attendees per monthly meeting for women in machine learning at CIS.
2023-2025	**WiML Workshop Mentoring** (NeurIPS) Volunteered as a mentor for the round table event at the annual Women in Machine Learning workshop at NeurIPS.
2023	**USABE fireside chat** (Penn Engineering) Participated in the Faculty Fireside Chat Series, where students may talk to professors and faculty in a small-group setting run by the Underrepresented Student Advisory Board in Engineering (USABE). USABE is an organization that works with SEAS leadership to promote diversity, equity, and inclusion through student advocacy. One of their initiatives includes student-faculty engagement and allowing students to interact in more casual settings with faculty.
2022-2025	**Mentorship for Underrepresented Masters/Undergraduates at Penn** (CIS) Direct mentorship in research experiences of masters and undergraduate students that are underrepresented in CS (2 woman and 1 Ethiopean)
2021	**Graduate Application Assistance Program** (MIT) Assisted applicants from under-represented groups with their graduate student applications to MIT's EECS PhD program.
2021-2022	**MIT Undergraduate Research Opportunities Program** (MIT) Directly supervised an undergraduate for the UROP program at MIT; provided an opportunity for a member of an under-represented group to learn about machine learning and tackle a challenging research project
2020-2022	**MEnTorEd Opportunities in Research (METEOR)** (MIT) Participated in the METEOR postdoc fellowship selection committee, an effort at CSAIL MIT to increase diversity, equity, and inclusion. Provided confidential technical feedback on candidates based on their application materials.
2019-2020	**CMU AI Mentoring Program** (CMU) Mentored undergraduate women and minorities in one-on-one meetings to provide career advice and discuss research/graduate school; the mentee is now a PhD student at UC Berkeley.
2019-2020	**Tecknowledge Mentor** (Obama Academy) Taught middle schoolers how to code as part of the Teknowledge outreach program at the Obama Academy; courses were intended to provide early exposure to computer science for under-represented students in low-income neighborhoods of Pittsburgh
2019	**Mental Health First Aid Certification** (CMU) Underwent training to recognize mental health issues and provide first aid assistance to those in need

Workshop Organizing

2024	**3rd New Frontiers in Adversarial Machine Learning** Organizer for the 3nd workshop on new directions in adversarial machine learning held at NeurIPS 2024 Website: https://advml-frontier.github.io/
2023	**2nd New Frontiers in Adversarial Machine Learning** Organizer for the 2nd ICML 2023 workshop on new directions in adversarial machine learning Website: https://advml-frontier.github.io/
2022	**Workshop on Adversarial Machine Learning and Beyond** Organizer for an AAAI 2022 workshop broadly themed around adversarial machine learning Website: https://advml-workshop.github.io/aaai2022/
2022	**New Frontiers in Adversarial Machine Learning** Organizer for an ICML 2022 workshop on new directions in adversarial machine learning Website: https://advml-frontier.github.io/
2021	**A Blessing in Disguise: The Prospects and Perils of Adversarial Machine Learning** Organizer for an ICML 2021 workshop themed around the dangers and benefits of adversarial machine learning adversarial machine learning. Website: https://advml-workshop.github.io/icml2021/
2021	**Robust and reliable ML in the real world** Main organizer for an ICLR 2021 workshop on real world robustness. Website: https://sites.google.com/connect.hku.hk/robustml-2021/home

Research Community Service

2024-Current	**ICML, NeurIPS, ICLR** Area Chair
2025	**COLM** Area Chair
2024-2025	**NSF CISE/CNS** Panelist
2023	**WiML** Workshop Mentor
2023	**SatML** Program Committee
2022	**Principles of Distribution Shift Workshop at ICML** Program Committee Website: https://sites.google.com/view/icml-2022-pods
2022	**AAAI 2023 Doctoral Consortium** Program Committee

2021-2022	**14th ACM workshop on Artificial Intelligence and Security** Program Committee Website: https://aisec.cc/
2020	**Towards Trustworthy ML: Rethinking Security and Privacy for ML** Program Committee Website: https://trustworthyiclr20.github.io/
2020	**AAAI** Program Committee
2020-2024	**ICML, NeurIPS, ICLR** Reviewer
2019	**Human-Centric Machine Learning Workshop** Program Committee Website: https://sites.google.com/view/hcml-2019
2019	**Security and Privacy of Machine Learning Workshop at ICML** Program Committee Website: https://icml2019workshop.github.io/
2019	**Adversarial Machine Learning in Real-World Computer Vision Systems at CVPR** Technical Program Commitee
2019	**1st Workshop on Adversarial Learning Methods for Machine Learning and Data Mining at KDD** Technical Program Commitee Website: https://sites.google.com/view/advml
2019	**Safe Machine Learning Workshop at ICLR** Program Committee Website: https://sites.google.com/view/safeml-iclr2019/