Eric Anthony Mitchell

eric.mitchell@cs.stanford.edu / https://ericmitchell.ai / @ericmitchellai

Education

Stanford University, Stanford, CA 2019 - 2024 (Expected)
Ph.D. Candidate, Computer Science
Advisors: Chelsea Finn & Christopher D. Manning
Fellowship: Stanford Knight-Hennessy Graduate Fellowship

Princeton University, Princeton, NJ 2014 - 2018
B.S.E., Computer Science (Highest Honors)
Advisor: H. Sebastian Seung

Professional Experience

DeepMind (AlphaCode team) *London, England*
Research Scientist Intern [hosts: Junyoung Chung, Nate Kushman, Aäron van den Oord] 2022

Samsung Research America *New York City, New York*
Research Intern, part-time 2019 - 2020
Machine Learning Research Engineer 2018 - 2019

Princeton Neuroscience Institute *Princeton, New Jersey*
Research Engineer 2018

Google *Mountain View, California*
Software Engineering Intern 2017

TapTools LLC *Winston-Salem, North Carolina*
Founder, iOS Developer 2012 - 2016

Selected Works

**indicates equal contribution*

Katherine Tian*, **Eric Mitchell***, Huaxiu Yao, Christopher D. Manning, Chelsea Finn. Fine-tuning Language Models for Factuality. *Preprint, under review.* 2023

Eric Mitchell, Rafael Rafailov, Archit Sharma, Chelsea Finn, Christopher D. Manning. An Emulator for Fine-Tuning Large Language Models using Small Language Models. *Preprint, under review.* 2023

Rafael Rafailov*, Archit Sharma*, **Eric Mitchell***, Stefano Ermon, Christopher D. Manning, Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. ***Outstanding Paper, Runner-Up***, *Neural Information Processing Systems (NeurIPS).* 2023

Nathan Hu*, **Eric Mitchell***, Christopher D. Manning, Chelsea Finn. Meta-Learning Online Adaptation of Language Models. *Empirical Methods in Natural Language Processing (EMNLP).* 2023

Katherine Tian*, **Eric Mitchell***, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, Christopher D. Manning. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. *Empirical Methods in Natural Language Processing (EMNLP).* 2023

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, Chelsea Finn. DetectGPT: DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. ***Oral Presentation (2% of submissions)***, *International Conference on Machine Learning (ICML).* 2023

Peter Henderson*, **Eric Mitchell***, Christopher D. Manning, Dan Jurafsky, Chelsea Finn. Self-Destructing Models: Increasing the Costs of Harmful Dual Uses in Foundation Models. ***Honorable Mention, Best Student Paper.*** *AAAI/ACM Conference on Artificial Intelligence, Ethics, & Society.* 2023

Eric Mitchell, Joseph J Noh, Siyan Li, William S Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, Christopher D Manning. Enhancing Self-Consistency and Performance of Pretrained Language Models with NLI. ***Oral Presentation (4% of submissions)****, Empirical Methods in Natural Language Processing (EMNLP).* 2022

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, Chelsea Finn. Memory-Based Model Editing at Scale. *International Conference on Machine Learning (ICML).* 2022

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, Christopher D. Manning. Fast Model Editing at Scale. *International Conference on Learning Representations (ICLR).* 2022

Rishi Bommasani et al. On the Risks and Opportunities of Foundation Models. *Whitepaper, Center for Research on Foundation Models.* 2021

Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, and Chelsea Finn. Offline Meta-Reinforcement Learning with Advantage Weighting. *International Conference on Machine Learning (ICML).* 2021

Eric Mitchell, Selim Engin, Volkan Isler, Daniel D Lee. Higher-Order Function Networks for Learning Composable 3D Object Representations. *International Conference on Learning Representations (ICLR).* 2020

Tarik Tosun*, **Eric Mitchell***, Ben Eisner, Jinwook Huh, Bhoram Lee, Daewon Lee, Volkan Isler, H. Sebastian Seung, Daniel D Lee. Pixels to Plans: Learning Non-Prehensile Manipulation by Imitating a Planner. *IEEE International Conference on Intelligent Robots and Systems (IROS).* 2019

Invited Talks

Simplifying RLHF with Direct Preference Optimization

Massachusetts Institute of Technology (MIT)	Dec 2023
UC Berkeley CS285: Deep Reinforcement Learning	Nov 2023
Cornell Tech	Nov 2023
Princeton Language and Intelligence (PLI)	Nov 2023
Prof. Aditi Raghunathan Group Meeting (Carnegie Mellon University)	Nov 2023
Prof. Mark Dredze Group Meeting (Johns Hopkins University)	Nov 2023

A Gradient-Free Emulator for Fine-tuning Language Models

Imbue (formerly Generally Intelligent)	Nov 2023

Toward More Truthful Language Models

UNC CS 790-150: Reliable Machine Learning	Oct 2023
Stanford CS329T: Trustworthy Machine Learning	Oct 2023

Detecting Machine-Generated Text

Google	Mar 2023
Beijing Academy of Artificial Intelligence (BAAI) Community	Mar 2023
Stanford Computer Systems Colloquium	Feb 2023

Editing the Behaviors of Pre-Trained Neural Networks

Hudson River Trading	Jan 2023
Allen Institute for AI	Feb 2022
Cohere	Dec 2021
Stanford CS 330: Deep Multi-task and Meta-Learning	Nov 2021
RAIVN Lab, University of Washington	Nov 2021

Myths about life & AI research

InspiritAI AI Education Conference	Oct 2021

Academic Awards

Stanford Accelerator for Learning Seed Grant ($50k award)	2023
Stanford Knight-Hennessy Graduate Fellowship	2019 - 2022
Princeton CS Outstanding Undergraduate Thesis Prize	2018
Princeton CS Best Undergraduate Research Poster Prize	2017
Srixon/Cleveland Golf All-America Scholar	2017
Tau Beta Pi Engineering Honor Society Vice President	2016 - 2017
Princeton Shapiro Prize for Academic Excellence	2016

Service

NeurIPS Socially Responsible Language Modeling Workshop Reviewer	2023
NeurIPS Conference Reviewer	2020 - 2021, 2023
Stanford CS Ph.D. Admissions Committee Student Representative	2020 - 2022
ICLR Conference Reviewer	2020 - 2021
NeurIPS Meta-Learning Workshop Reviewer	2020 - 2021
AI4ALL Summer Camp Instructor	2018

Other Activities

PADI Open Water SCUBA certification	2023
Princeton NCAA Division I Varsity Golf Team	2014 - 2018
HackPrinceton (1st Place Team, Software)	2015
Guitar, Singing, Songwriting	