Eric W. Todd

github.com/ericwtodd | ericwtodd.github.io | todd.er@northeastern.edu

Education

PhD Candidate, Computer Science — Sep 2022–Present
MS, Computer Science — Aug 2024
Northeastern University - Khoury College of Computer Sciences — Boston, MA
Advisor: David Bau
Research Area: Machine Learning, Interpretability

BS, Applied and Computational Mathematics — Apr 2020
Brigham Young University - GPA: 4.00/4.00, *Summa Cum Laude* — Provo, UT
Minors: Computer Science, Statistics

Publications

Conference Papers

1. Sheridan Feucht, **Eric Todd**, Byron C. Wallace, David Bau. "The Dual Route Model of Induction." *The Second Conference on Language Modeling.* (COLM 2025)
2. Jaden Fiotto-Kaufman, Alexander R. Loftus, **Eric Todd**, Jannik Brinkmann, Caden Juang, Koyena Pal, Can Rager, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Michael Ripa, Adam Belfki, Nikhil Prakash, Sumeet Multani, Carla Brodley, Arjun Guha, Jonathan Bell, Byron C. Wallace, David Bau. "NNsight and NDIF: Democratizing Access to Open-Weight Foundation Model Internals." *The Thirteenth International Conference on Learning Representations.* (ICLR 2025)
3. **Eric Todd**, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, David Bau. "Function Vectors in Large Language Models." *The Twelfth International Conference on Learning Representations.* (ICLR 2024)
4. **Eric Todd**, Mylan R. Cook, Katrina Pedersen, David S. Woolworth, Brooks A. Butler, Xin Zhao, Colt Liu, Kent L. Gee, Mark K. Transtrum, Sean Warnick. "Automatic detection of instances of focused crowd involvement at recreational events." *Proceedings of Meetings on Acoustics* **39** (1). (2019)
5. Brooks A. Butler, Katrina Pedersen, Mylan R. Cook, Spencer G. Wadsworth, **Eric Todd**, Dallen Stark, Kent L. Gee, Mark K. Transtrum, Sean Warnick. "Classifying crowd behavior at collegiate basketball games using acoustic data." *Proceedings of Meetings on Acoustics* **35** (1). (2018)

Preprints & In Submission

1. Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, **Eric Todd**, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, Tom McGrath "Open Problems in Mechanistic Interpretability." arxiv.org/abs/2501.16496 (2025)
2. Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, **Eric Todd**, David Bau, Yonatan

Belinkov. "The Quest for the Right Mediator: A History, Survey, and Theoretical Grounding of Causal Interpretability." arxiv.org/abs/2408.01416 (2024)

Employment

Research Assistant Sep 2022–Present
Northeastern University - Bau Lab (Interpretable Neural Networks) Boston, MA

– Researching the mechanisms large neural networks use to learn from context in order solve different tasks, and how their internal representations enable their impressive generalization capabilities.

– "Function Vectors in Large Language Models" (ICLR 2024) investigates whether LLMs contain function representations through the lens of in-context learning. We found that a small set of attention heads transport extractable task-representative information that is robust to different contexts.

Research Assistant Sep 2020–Aug 2022
Brigham Young University - Computer Science Department, Advisor: Ryan Farrell Provo, UT

– Researched unsupervised methods for fine-grained image segmentation

Research Intern May–Aug 2022
Air Force Research Lab/Wright State University, AFRL Advisor: Oliver Nina Remote

– Investigated self-supervised learning methods for fine-grained image classification and presented my research to other interns and AFRL research advisors.

Machine Learning Intern May–Aug 2019
Brigham Young University - Enrollment Services, Manager: Kristine Manwaring Provo, UT

– Helped develop Early Alert, a machine learning system that identifies students struggling academically and enables personalized outreach from campus support offices.

– Early Alert is deployed and in active use by most academic advisors and support offices at BYU, and was highlighted in this campus news article.

Research Assistant Feb 2018–Aug 2020
Brigham Young University - Physics Department, Advisor: Mark Transtrum Provo, UT

– Researched crowd noise classification using machine learning methods. My work focused on crowd noise data from basketball games and a Mardi Gras parade float, resulting in 2 publications [4, 5].

Teaching

Northeastern University
Khoury College of Computer Sciences

– Deep Learning (CS 7150), Guest Lecture — April 2025
  Instructor: David Bau, Lecture Topic: In-Context Learning
– Practical Neural Networks (DS 4440), Teaching Assistant — Spring 2024
  Instructor: David Bau

Brigham Young University
Department of Computer Science

– Computer Vision (CS 450), Teaching Assistant — Winter 2022
  Instructor: Ryan Farrell
– Introduction to Machine Learning (CS 472), Teaching Assistant — Winter 2021, Summer 2021
– Deep Learning (CS 474), Teaching Assistant — Fall 2021

Department of Mathematics

– Algorithm Design and Optimization 2 Lab (Math 323), Teaching Assistant — Winter 2020
– Mathematical Analysis 2 Lab (Math 347), Teaching Assistant — Winter 2020
– Algorithm Design and Optimization 1 Lab (Math 321), Teaching Assistant — Fall 2019
– Mathematical Analysis 1 Lab (Math 345), Teaching Assistant — Fall 2019
– Introduction to Mathematical Python (Math 495R), Teaching Assistant — Winter 2019

Presentations

Invited Talks
– *Function Vectors in Large Language Models.* Invited talk at the Princeton Neuroscience Institute. Princeton, NJ. May 2024
– *Opening AI's Black Box with Prof. David Bau, Koyena Pal, and Eric Todd of Northeastern University*. The Cognitive Revolution Podcast. Boston, MA. April 2024.

Conference Presentations
– *Showing vs. Telling in LLMs.* Poster at New England Mechanistic Interpretability Workshop. Boston, MA. August 2024.
– *Detecting instances of focused crowd involvement at recreational events*. Acoustical Society of America Meeting. San Diego, CA. December 2019. (Presenter: Mylan R. Cook)
– *Feature reduction of crowd noise used for machine learning classification*, Acoustical Society of America Meeting. San Diego, CA. December 2019. (Presenter: Brooks Butler)
– *Improved automated classification of basketball crowd noise*, Acoustical Society of America Meeting. Louisville, KY. May 2019. (Presenter: Mylan R. Cook)
– *Unsupervised classification of crowd noise at BYU basketball games.* BYU CPMS Student Research Conference. Provo, UT. March 2019. (Co-Presenter: Brooks Butler)
– *Clustering analysis of crowd noise from collegiate basketball games*, Acoustical Society of America Meeting. Victoria, BC, Canada. November 2018. (Presenter: Brooks Butler)
– *Modeling Crowd Noise with Machine Learning.* BYU CPMS Student Research Conference. Provo, UT. March 2018.

Awards

Northeastern Graduate Assistantship, *Northeastern University* 2022–Present
Khoury College Start-Up Fund, *Northeastern University* ($5,000) 2022
President's Leadership Council Presentation, *Brigham Young University* 2020
– Selected to represent my college's 3000+ students by presenting my internship work on Early Alert to BYU's $1M+ donors and top university administration.
Outstanding Performance in Mathematics Award, *Brigham Young University* 2020
– Awarded to the top performing mathematics majors of my graduating class as voted on by faculty
Warren Rollins and Murdell Hull Scholarship, *Brigham Young University* ($1,000) 2020
Brigham Young Scholarship (Full Tuition), *Brigham Young University* ($34,175) 2014–2020
CPMS Dean's List (Top 5% of College), *Brigham Young University* 2017-2020
New Century/Regents Scholarship, *Utah System of Higher Education* ($6,000) 2014-2018

Service

Program Committee Reviewer:

2025
– Mechanistic Interpretability Workshop (NeurIPS Workshop) - 4 Papers 2025
– 1st Workshop on the Interplay of Model Behavior and Internals (COLM Workshop) - 3 papers 2025
– Conference on Neural Information Processing Systems (NeurIPS) - 5 papers 2025
– Workshop on Actionable Interpretability (ICML Workshop) - 3 papers 2025
– Conference on Language Modeling (COLM) - 2 papers 2025
– International Conference on Machine Learning (ICML) - 7 papers 2025
– 1st Workshop on Mechanistic Interpretability for Vision (CVPR Workshop) - 2 papers 2025
– International Conference on Learning Representations (ICLR) - 3 papers 2025

2024
– Interpretable AI: Past, Present, and Future (NeurIPS Workshop) - 2 papers 2024
– Conference on Neural Information Processing Systems (NeurIPS) - 6 papers 2024
– 1st ICML Workshop on In-Context Learning (ICML Workshop) - 2 papers 2024