Jonathn Chang

(805) 807-8516 · jc3683@cornell.edu · jchang153.github.io

Education

Cornell University | Ithaca, NY — August 2024 - present

- Ph.D. in Applied Mathematics with Minors in Mathematics and Computer Science, Cumulative GPA: **4.12/4.3**
- Advisors: Prof. Lionel Levine and Prof. Ziv Goldfeld

University of California, Los Angeles | Los Angeles, CA — September 2020 - June 2024

- B.S. in Applied Mathematics with Specialization in Computing, Cumulative GPA: **3.98/4.00** (Magna Cum Laude)

Publications

- PLOT: Progressive Localization via Optimal Transport in Neural Causal Abstraction
  Jonathn Chang, Arya Datla, Ziv Goldfeld
  International Conference on Machine Learning (ICML) Workshop for Mechanistic Interpretability, 2026
- Side Effects of Character Training: Quantifying Cross-Constitution Drift in LLMs
  Bhagyesh Kumar, Ananya Sutradhar, Saurav Panigrahi, **Jonathn Chang**, Lionel Levine
  International Conference on Machine Learning (ICML) Workshop for Pluralistic Alignment, 2026
- EigenBench: A Comparative Behavioral Measure of Value Alignment
  Jonathn Chang, Leonhard Piff, Suvadip Sana, Jasmine Li, Lionel Levine
  International Conference on Learning Representations (ICLR), 2026, Oral (top 1.13% of submissions)
- Hate speech and hate crimes: a data-driven study of evolving discourse around marginalized groups
  Malvina Bozhidarova*, **Jonathn Chang***, Aaishah Ale-rasool, Yuxiang Liu, Chongyao Ma, Andrea L. Bertozzi, P. Jeffrey Brantingham, Junyuan Lin, Sanjukta Krishnagopal
  IEEE International Conference on Big Data, 2023

Talks and Posters

- IPAM Foundations of Interpretability Workshop, UCLA — August 2026
- ICML Workshop for Mechanistic Interpretability, Seoul, South Korea — July 2026
- ICML Workshop for Pluralistic Alignment, Seoul, South Korea — July 2026
- ICLR Oral, Rio de Janeiro, Brazil — April 2026
- Student Colloquium in Applied Mathematics (SCAM), Cornell University — December 2025
- Media in the Digital Age, Seminar Series at the Alan Turing Institute — October 2024
- IEEE International Conference on Big Data, Sorrento, Italy — December 2023
- UCLA Computational and Applied Mathematics REU Oral — August 2023
- Texas Tech REU Mini-Symposium Oral and Poster Session — July 2022

Awards and Grants

- Siegel PiTech PhD Impact Fellowship — May 2025
- UCLA Daus Scholarship in Mathematics — June 2024
- UCLA L. and M. Gross Undergraduate Math Scholarship — October 2023

Research Experience

Supervised Program for Alignment Research (SPAR) — January 2026 - Present
Co-mentor

- Selected as co-mentor to lead a team of four undergraduate/masters students, and awarded $3000 funding for an alignment research project to quantify emergent misalignment and unintended generalizations induced by character training in LLMs

Cornell MATH + AI Lab — January 2026 - May 2026
Co-mentor

- Led a team of four undergraduate students in a semester-long research project on causal abstractions and optimal transport for mechanistic interpretability

Cornell Graduate Research Assistant
Probabilistic Causal Abstractions via Optimal Transport (advisor: Dr. Ziv Goldfeld) January 2026 - Present

- Developed an optimal transport framework for mechanistic interpretability that progressively localizes causal variables in neural networks, matching SOTA accuracies on tasks in the Mechanistic Interpretability Benchmark (MIB) suite while reducing runtime by several orders of magnitude

EigenBench (advisor: Dr. Lionel Levine) April 2025 - November 2025

- Developed a black-box framework for quantifying subjective traits and value alignment in LLMs without requiring ground truth labels, using ensemble model evaluations and PageRank-style aggregation

NYC Department of Environmental Protection May 2025 - August 2025
Siegel PiTech Impact Fellow

- Built a multi-camera vehicle re-identification system to automate summons issuance for NYC's Noise Camera Enforcement Program, processing 30,000+ vehicle noise events using audio and vision-based pipelines, reducing the number of manually-inspected events by 51%

UCLA Undergraduate Research Assistant June 2023 - August 2023
Predictors of Gun Violence using Knowledge Graphs (advisor: Dr. Andrea Bertozzi)

- Analyzed large-scale Twitter data to study correlations between online discourse and hate-crime incidents using VADER sentiment analysis, BERTopic topic modeling, and dynamic network/community detection

TTU Undergraduate Research Assistant June 2022 - August 2022
Recovering PDEs via Friedrichs Learning (advisor: Dr. Haizhao Yang)

- Studied inverse problems for high-dimensional PDEs using physics-informed neural networks, utilizing a variational form of the PDE to develop a minimax training approach for robustness to noisy and non-uniform data

Teaching Experience

MATH 1920 (Multivariable Calculus for Engineers) Recitation TA, Cornell University — Fall 2025, Spring 2026
MATH 2940 (Linear Algebra for Engineers) Recitation TA, Cornell University — Spring 2025
MATH 2930 (Differential Equations for Engineers) Recitation TA, Cornell University — Fall 2024

Professional Service

Conference reviewer: Conference on Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML)

Technical Skills

Python, PyTorch, TensorFlow, Hugging Face Transformers, Weights and Biases, pyvene, Git/GitHub, Jupyter, Conda, NumPy, CUDA, LaTeX