Trenton Chang

Website: https://trentonchang.org
Email: ctrenton@umich.edu
LinkedIn: trenton-chang

Research Interests

My research interests are in the alignment and misalignment of machine learning towards high-level principles and values, in real-world, high-stakes domains such as public policy and healthcare. To date, my contributions include studying bias mitigation in ML clinical risk-stratification models in the presence of biases in laboratory testing decisions, steerability of large language models with respect to multi-dimensional attributes, and detecting gaming/abuse of ML-based risk-adjustment in public health insurance.

Education

University of Michigan — Ann Arbor, MI
Ph.D. Candidate in Computer Science & Engineering, GPA: 4.00 — 2021–present
Advisor: Jenna Wiens

Stanford University — Stanford, CA
M.S. in Computer Science—Artificial Intelligence track, GPA: 4.05 — 2020–2021

Stanford University — Stanford, CA
B.A. in American Studies, *with distinction* & Phi Beta Kappa, GPA: 3.98 — 2016–2020

Experience

Microsoft Research — Redmond, WA
Mentors: Adith Swaminathan & Tobias Schnabel, Augmented Learning & Reasoning Group — May 2024-Aug 2024

Publications & Preprints

[P1] **T. Chang**, Tobias Schnabel, Adith Swaminathan, and Jenna Wiens. "A Course Correction in Steerability Evaluation: Revealing Miscalibration and Side Effects in LLMs". May 2025. arXiv: 2505.23816 [cs.CL].

[P2] Dylan Zapzalka, **T. Chang**, Lindsay Warrenburg, Sae-Hwan Park, Daniel K. Shenfeld, Ravi B. Parikh, Jenna Wiens, and Maggie Makar. "Estimating Misreporting in the Presence of Genuine Modification: A Causal Perspective". May 2025. arXiv: 2505.23954 [cs.LG].

[P3] Winston Chen, **T. Chang**, and Jenna Wiens. "LobsterNet: Estimating Conditional Average Treatment Effects Under Treatment Non-compliance". In: *Conference on Health, Inference, and Learning*. May 2025.

[P4] Sarah Jabbour*, **T. Chang***, Anindya Das Antar*, Joseph Peper, Insu Jang, Jiachen Liu, Jae-Won Chung, Shiqi He, Michael Wellman, Bryan Goodman, Elizabeth Bondi-Kelly, Kevin Samy, Rada Mihalcea, Mosharaf Chowhury, David Jurgens, and Lu Wang. "Evaluation Framework for AI Systems in "the Wild"". May 2025. arXiv: 2504.16778 [cs.CL].

[P5] **T. Chang**, Jenna Wiens, Tobias Schnabel, and Adith Swaminathan. "Measuring Steerability in Large Language Models". In: *Workshop on Safe Generative AI, Thirty-eighth Annual Conference on Neural Information Processing Systems*. Dec. 2024.

[P6] **T. Chang**, Lindsay Warrenburg, Sae-Hwan Park, Ravi B. Parikh, Maggie Makar, and Jenna Wiens. "Who's Gaming the System? A Causally-Motivated Approach for Detecting Strategic Adaptation". In: *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems*. Dec. 2024.

[P7] **T. Chang**, Mark Nuppnau, Ying He, Keith Kocher, Thomas S. Valley, Michael W. Sjoding, and Jenna Wiens. "Racial differences in laboratory testing as a mechanism for bias amplification for AI models in healthcare: the emergency department as a case study ". In: *PLOS Global Public Health*. Oct. 2024.

[P8] **T. Chang** and Jenna Wiens. "From Biased Selective Labels to Pseudo-Labels: An Expectation-Maximization Framework for Learning from Biased Decisions". In: *Proceedings of the 41st International Conference on Machine Learning*. July 2024.

[P9] Ethan A. Chi, Ashwin Paranjape, Abigail See, Caleb Chiam, **T. Chang**, Kathleen Kenealy, Swee Kiat Lim, Amelia Hardy, Chetanya Rastogi, Haojun Li, Alexander Iyabor, Yutong He, Hari Sowrirajan, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, Jillian Tang, Avanika Narayan, Giovanni Campagna, and Christopher Manning. "Neural Generation Meets Real People: Building a Social, Informative Open-Domain Dialogue Agent". In: *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Sept. 2022.

[P10] **T. Chang**, Michael W. Sjoding, and Jenna Wiens. "Disparate Censorship: A Plausible, Underexplored Mechanism for Model Performance Gaps in Clinical Machine Learning". In: *7th Machine Learning for Healthcare Conference*. Proceedings of Machine Learning Research, Aug. 2022.

[P11] **T. Chang** and Daniel Y. Fu. "Lost in Transmission: On the Impact of Networking Corruptions on Video Machine Learning Models". June 2022. arXiv: 2206.05252 [cs.CV].

[P12] Ethan A. Chi, Caleb Chiam, **T. Chang**, Swee Kiat Lim, Chetanya Rastogi, Alexander Iyabor, Yutong He, Hari Sowrirajan, Avanika Narayan, Jillian Tang, Haojun Li, Ashwin Paranjape, and Christopher D. Manning. "Neural, Neural Everywhere: Controlled Generation Meets Scaffolded, Structured Dialogue". In: *Alexa Prize Socialbot Grand Challenge 4 Proceedings*. July 2021.

[P13] **T. Chang**, Daniel Y. Fu, Yixuan Li, and Christopher Ré. "Beyond the Pixels: Exploring the Effect of Video File Corruptions on Model Performance". In: *2020 European Conference in Computer Vision, Workshop on Adversarial Robustness in the Real World*. Aug. 2020.

Press Appearances & Media Outreach

[M1] Derek Smith. "Accounting for bias in medical data helps prevent AI from amplifying racial disparity". In: *Michigan Engineering News* (Oct. 2024).

[M2] Casey Ross, Brittany Trang, and Mario Aguilar. "What does generative AI mean for health care? We asked the experts". In: *STAT+* (May 2023).

[M3] Michigan AI Lab [@michigan_AI]. "AI, Healthcare, and Humanities with Trenton Chang". Aug. 2022.

[M4] Truly Render. "Decisive Differences in Healthcare AI". In: *Discover Rackham* (Oct. 2022).

Invited Talks & Presentations

[T1] **T. Chang**. "Bias in, bias out: analyzing sources of bias in machine learning for healthcare". In: *Borgwardt Group, Max-Planck-Institut für Biochemie, internal talk*. July 2024.

[T2] **T. Chang**. "Building Bridges: University of Michigan". In: *Grand Valley State University, Women in Computing*. Nov. 2024.

[T3] **T. Chang**. "Machine learning in a world of messy data". In: *University of Michigan CSE Honors Competition*. Nov. 2024.

[T4] **T. Chang**. "Measuring and Mitigating the Impact of Biases in Laboratory Testing on Machine Learning Models". In: *NIH Office of Data Science Strategy AI Supplement Program PI Meeting*. Feb. 2024.

[T5] **T. Chang**. "Mitigating the Effects of Label-Bias: An Expectation-Maximization Approach". In: *Michigan AI Symposium*. Oct. 2023.

[T6] **T. Chang**. "Recognizing and Addressing Biases in Machine Learning for Healthcare". In: *Ann Arbor Machine Learning Meetup (Ann Arbor SPARK)*. Oct. 2023.

[T7] **T. Chang**. "Disparate Censorship: A Plausible, Underexplored Mechanism for Model Performance Gaps in Clinical Machine Learning". In: *Michigan AI Symposium*. Dec. 2022.

[T8] **T. Chang** and D. Ganelin. "Machine Learning Bias in Criminal Justice". In: *Computer Science Teachers of America Conference*. July 2021.

Teaching

- Graduate Student Instructor, EECS 598-009 (Causality and machine learning), University of Michigan (2023)

  - Delivered 80-min. lecture on fairness in machine learning from a causal perspective.
  - Assisted in writing homework and grading for causal inference seminar with 23 M.S. and Ph.D. stduents.
- Workshop Organizer, Discover Engineering, University of Michigan (2023)
  - Recruited 9 volunteer instructors and designed workshop introducing high school students to computer science and an interactive exploration of the limitations and capabilities of ChatGPT, reaching 4 cohorts of approx. 10 students each.
- Workshop Organizer, Xplore Engineering: "How do Computers Think?", University of Michigan (2023)
  - Recruited 12 volunteer instructors and designed workshop introducing 4th - 7th grade students to computer science and an activity analyzing the robustness of image classification models, reaching 6 cohorts of approx. 10 students each.
- Volunteer Instructor, AI4ALL, University of Michigan (2022)
  - Co-taught project on n-gram based text generation and sentiment analysis to 9 high school stduents.
- Instructor, Inspirit AI (2020, 2021)
  - Wrote and taught project on the usage of AI in criminal justice decisions for high school students.
- Residential Counselor, Artificial Intelligence Course, Stanford Pre-Collegiate Studies (2019)
  - Mentored projects in AI ranging from computer vision to price prediction for 2 cohorts of approx. 15 students each.

Mentees

- Zhiyi Hu (2024-present, SJTU/University of Michigan, **current:** incoming MS student, Computer Sci. & Eng., Michigan
  - Mentored student project on the robustness of matching estimators in high-dimensional, multi-treatment settings
  - Presentation at 2025 Michigan Undergraduate Research Symposium
- Ryan Chi & Ian Ng (2021), **current:** MS students, Computer Science, Stanford University
  - Mentored student benchmark submission for sarcasm detection by LLMs in social media comments with targeted edits
  - Task accepted to Google BIG-BENCH, later included in BIG-BENCH Hard

Service

- Outreach Chair, Machine Learning for Health (ML4H) Symposium (2025, *current*)
- University of Michigan AI Blog Co-Coordinator (2024, *current*)
- Workflow Chair, Machine Learning for Health (ML4H) Symposium (2024)
- Panelist, Senior Ph.D. Student Panel, EECS 498: Machine Learning Research Experience (2024)
- University Relations Chair, Comp. Sci. & Eng. Graduate Student Organization, University of Michigan (2023-2024)
- Panelist, Summer Research Opportunity Program, University of Michigan (2023)
- AI Lab Grad. Admissions Committee Volunteer, Division of Comp. Sci. & Eng., University of Michigan (2022, 2024)
- Reviewing: AISTATS, CHIL, ICML, ML4H, MLHC, NeurIPS, ICLR, TMLR, KDD (workshop only).

Awards

- NeurIPS Top Reviewer (2024)
- CSE Honors Competition Finalist (AI Lab Representative), University of Michigan (2024)
- NeurIPS Research2Clinics Workshop, Best Reviewer Award (2021)
- Team 2nd Prize (Stanford Chirpy Cardinal), Alexa Socialbot Grand Challenge (2021)