

### Education

2019 – **Stanford University**, *PhD Candidate in Computer Science*

- Advisor: Christopher Ré

2015 – 2019 **Princeton University**, *B.S.E in Operations Research and Financial Engineering (ORFE), certificate in Applications of Computing*, GPA: 3.962/4

- Graduated Summa Cum Laude
- Senior Thesis: *A Quantum Version of the Multiplicative Weights Algorithm* (recipient of the Ahmet S. Çakmak Thesis Prize)
- Thesis advisor: Elad Hazan

### Research Interests

I'm interested in studying and improving the fundamentals of modern machine learning through data (data-centric AI). On the model training side, I work on data mixing, curriculum learning, synthetic data, and data labeling. On the inference side, I work on techniques such as ensembling, routing, and verification. Currently, I am thinking about how to develop and operationalize a more principled understanding of how models learn from data.

### Publications

[20] **Shrinking the Generation-Verification Gap with Weak Verifiers.**

Jon Saad-Falcon\*, E. Kelly Buchanan\*, Mayee F. Chen\*, Tzu-Heng Huang, Brendan McLaughlin, Tanvir Bhathal, Shang Zhu, Ben Athiwaratkun, Frederic Sala, Scott Linderman, Azalia Mirhoseini, Christopher Ré. *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[19] **Archon: An Architecture Search Framework for Inference-Time Techniques.**

Jon Saad-Falcon, Adrian Gamarra Lafuente, Shlok Natarajan, Nahum Maru, Hristo Todorov, Etash Kumar Guha, E. Kelly Buchanan, Mayee F. Chen, Neel Guha, Christopher Ré, Azalia Mirhoseini. *International Conference on Machine Learning (ICML)*, 2025.

[18] **Aioli: A Unified Optimization Framework for Language Model Data Mixing.**

Mayee F. Chen\*, Michael Y. Hu\*, Nicholas Lourie, Kyunghyun Cho, Christopher Ré. *International Conference on Learning Representations (ICLR)*, 2025.

[17] **Smoothie: Label Free Language Model Routing.**

Neel Guha\*, Mayee F. Chen\*, Trevor Chow, Ishan S. Khare, Christopher Ré. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[16] **DataComp-LM: In search of the next generation of training sets for language models.**

Jeffrey Li, . . . , Ludwig Schmidt, Vaishaal Shankar (59 authors including Mayee F. Chen). *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

[15] **Cookbook: A framework for improving LLM generative abilities via programmatic data generating templates.**

Annika Narayan\*, Mayee F. Chen\*, Kush Bhatia, and Christopher Ré. *Conference on Language Modeling (COLM)*, 2024.

[14] **Skill-it! A Data-Driven Skills Framework for Understanding and Training Language Models.**

Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. **Spotlight (top 3.1% of submissions).**

[13] **Embroid: Unsupervised Prediction Smoothing can Improve Few-Shot Classification.**

Neel Guha\*, Mayee F. Chen\*, Kush Bhatia, Azalia Mirhoseini, Frederic Sala, and Christopher Ré. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[12] **A Case for Reframing Automated Medical Image Classification as Segmentation.**

Sarah M. Hooper, Mayee F. Chen, Khaled Saab, Kush Bhatia, Curtis Langlotz, and Christopher Ré. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

- [11] **Anomaly Detection with Multiple Reference Datasets in High Energy Physics.**  
 Mayee F. Chen, Benjamin Nachman, and Frederic Sala.  
*Journal of High Energy Physics*, 2023.
- [10] **Ask Me Anything: A Simple Strategy for Prompting Language Models.**  
 Simran Arora\*, Avanika Narayan\*, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré.  
*International Conference on Learning Representations (ICLR)*, 2023. **Notable 25% of acceptances.**
- [9] **Reducing Reliance on Spurious Features in Medical Image Classification with Spatial Specificity.**  
 Khaled Saab, Sarah M. Hooper, Mayee F. Chen, Michael Zhang, Daniel Rubin, and Christopher Ré.  
*Machine Learning for Healthcare (MLHC)*, 2022.
- [8] **Shoring Up the Foundations: Fusing Model Embeddings and Weak Supervision.**  
 Mayee F. Chen\*, Daniel Y. Fu\*, Dyah Adila, Michael Zhang, Frederic Sala, and Christopher Ré.  
*Uncertainty in Artificial Intelligence (UAI)*, 2022. **Best Student Paper Runner-Up Award, Oral Presentation.**
- [7] **Perfectly Balanced: Improving Transfer and Robustness of Supervised Contrastive Learning.**  
 Mayee F. Chen\*, Daniel Y. Fu\*, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré.  
*International Conference on Machine Learning (ICML)*, 2022.
- [6] **TABi: Type-Aware Bi-Encoders for Open-Domain Entity Retrieval.**  
 Megan Leszczynski, Daniel Y. Fu, Mayee F. Chen, and Christopher Ré.  
*Findings of the Association for Computational Linguistics*, 2022.
- [5] **The Details Matter: Preventing Class Collapse in Supervised Contrastive Learning.**  
 Mayee F. Chen\*, Daniel Y. Fu\*, Michael Zhang, Kayvon Fatahalian, and Christopher Ré.  
*AAAI Workshop on Artificial Intelligence with Biased or Scarce Data*, 2022. **Best Paper Award.**
- [4] **An Adversarial Model of Network Disruption: Maximizing Disagreement and Polarization in Social Networks.**  
 Mayee F. Chen and Miklos Z. Racz.  
*IEEE Transactions on Network Science and Engineering (TNSE)*, 2021.
- [3] **Mandoline: Model Evaluation under Distribution Shift.**  
 Mayee F. Chen\*, Karan Goel\*, Nimit Sohoni\*, Fait Poms, Kayvon Fatahalian, and Christopher Ré.  
*ICML*, 2021.
- [2] **Comparing the Value of Labeled and Unlabeled Data in Method-of-Moments Latent Variable Estimation.**  
 Mayee F. Chen\*, Benjamin Cohen-Wang\*, Steve Mussmann, Frederic Sala, and Christopher Ré.  
*AISTATS*, 2021.
- [1] **Fast and Three-rious: Speeding Up Weak Supervision with Triplet Methods.**  
 Mayee F. Chen\*, Daniel Y. Fu\*, Frederic Sala, Sarah M. Hooper, Kayvon Fatahalian, and Christopher Ré.  
*International Conference on Machine Learning (ICML)*, 2020

## Awards and Honors

- 2023 Microsoft Accelerate Foundation Models Research Grant
- 2021 NSF GRFP Honorable Mention
- 2019 Ahmet S. Çakmak Prize, *Princeton University*, awarded for innovative research and an exceptional senior thesis.
- 2018 Phi Beta Kappa, *Princeton University*, one of 28 early inductees.
- 2017 Tau Beta Pi Engineering Honor Society, *Princeton University*
- 2017 Shapiro Prize for Academic Excellence, *Princeton University*, awarded to 2-3% of the class for exceptional academic record.

## Work and Teaching Experience

- 09/25 – **Course Assistant, Stanford University**
- CS229: Machine Learning. Advising multiple student teams on their final research project.

- 03/25 – **Research Intern**, *Allen Institute for Artificial Intelligence*, AllenNLP, Part-time
  - Driving pre-training data mixing efforts for OLMo open-source language models, building on Aioli (ICLR 2025) and Skill-It (NeurIPS 2023).
- 10/24 – **Research Consultant**, *Snorkel AI*
  - 03/25 ○ Developed reasoning evaluation benchmarks for frontier models.
- 06/23 – **Research Intern**, *Microsoft*, Redmond, WA, Office of Applied Research
  - 09/23 ○ Research on multi-objective prompt optimization.
- 01/23 – **Course Assistant**, *Stanford University*
  - 04/23 ○ CS228: Probabilistic Graphical Models
- 2016 – 19 **Grader for Computer Science Department**, *Princeton University*
  - COS226: Algorithms and Data Structures (lead grader), COS326: Functional Programming, COS340: Reasoning about Computation, COS324: Introduction to Machine Learning, and COS445: Economics and Computing
- 06/18–08/18 **Quantitative Trading Intern**, *IMC Trading*, Chicago, IL, Fixed Income, Currencies, and Commodities Desk
- 05/17–08/17 **Software Engineering Intern**, *Google*, Mountain View, CA, Advertiser Platform Team  
Worked on AdWords Next Overviews, frontpage data analytics for ads campaigns
- 05/16–08/16 **Engineering Practicum Intern**, *Google*, Mountain View, CA, Cloud/Cluster/Kernel team  
Worked on an infrastructure tool for pushing configuration and data updates to services within Google

## Talks

- Sept 27, 2025 Data Mixing Tutorial at Data@CoRL Workshop (Conference on Robot Learning)
- Jan 30, 2025 Morgan Stanley Research Machine Learning Seminar
- Dec 5, 2024 DatologyAI
- Dec 4, 2024 UChicago ChiData Lab
- Oct 31, 2024 New York University Computational Linguistics and Cognitive Science Seminar
- Oct 24, 2024 Salesforce AI Research FutureForum
- Oct 11, 2023 Stanford Social and Language Technologies Lab
- Aug 31, 2023 Allen Institute for AI
- April 2, 2023 Stanford Generative AI and Foundation Models Workshop
- Dec 5, 2022 NeurIPS Tutorial on Theory and Practice of Dataset Construction, Panelist
- April 30, 2022 Stanford-Berkeley Women in CS/EE Research Meetup
- April 8, 2022 Snorkel AI Machine Learning Whiteboard Talk
- Aug 5, 2021 Stanford MedAI Talk Series
- June 8, 2021 DAWN Research Workshop
- Nov 6, 2020 Google x Stanford Summit

## Coursework

Relevant graduate courses:

- Information Theoretic Lower Bounds in Data Science, Convex Optimization II, Randomized Algorithms

Relevant undergraduate courses:

- *ORFE Courses*: Probability Theory (graduate-level course), Optimization, High Frequency Trading, Decision Modeling for Business Analytics, Monte Carlo Simulation, Strategy and Information, Financial Mathematics, Analysis of Big Data, Probability and Stochastics, Microeconomic Theory, Statistics
- *Computer Science Courses*: Optimization for Machine Learning (graduate-level seminar), Computer Networks, Operating Systems, Economics and Computing, Introduction to Machine Learning, Information Security, Human-Computer Interfaces, Neural Networks, Functional Programming, Reasoning About Computation, Programming Systems, Algorithms and Data Structures

## Service

### Reviewing

I have served as **reviewer** for the following conferences:

- ICML (2021-2025)
- NeurIPS (2021-2025), NeurIPS Datasets and Benchmarks (2024)
- ICLR (2024-2026)
- COLM (2025)
- AISTATS (2023)
- UAI (2020, 2023-2024)
- KDD (2020)

the following workshops:

- ICML Machine Learning for Data: Automated Creation, Privacy and Bias (2021)
  - NeurIPS Interpolate: First Workshop on Interpolation Regularizers and Beyond (2022)
  - Mathematical and Empirical Understanding of Foundation Models (ICLR 2023-24)
  - Efficient Systems for Foundation Models (ICML 2023-2024)
  - Data-Centric Machine Learning Research (ICML 2023-2024, ICLR 2024)
  - Navigating and Addressing Data Problems for Foundation Models (ICLR 2024)
- and the following journals:
- Journal of Data-Centric Machine Learning Research

### Activities

At Stanford University:

- Computer Science PhD Admissions Committee (2020-2023)
- CS Student Applicant Support Program (Mentor 2020-2022, Organizer 2023-2025)
- WiML (Women in Machine Learning) PhD Application Mentorship Program (2022)
- Graduate WiCS (Women in Computer Science) Mentor (2021-2022)
- CS Undergraduate Mentorship Program (2021-2022)
- XTRM Kpop Cover Group: dance captain (2019–), Alliance Dance Team (2019–2021, 2023–)

Other:

- WiML PhD Application Mentorship Program (2022)
- Organizer for Data-Centric Machine Learning Research Workshop (ICLR 2024)