Claudia Shi

Website: www.claudiashi.com
Email: claudia.j.shi@gmail.com
GitHub: github.com/claudiashi57

Education

Columbia University — New York, NY
Ph.D. in Computer Science, Advisor: David Blei — Jan 2020–May 2025

Columbia University — New York, NY
M.S. in Computer Science — Sep 2018–Dec 2019

Tulane University — New Orleans, LA
B.S. in Math and Economics, B.A. in Philosophy — Sep 2014–May 2018

- General course (2016 - 2017) at the London School of Economics

Employment

Research Scientist — June 2025 – Current
Meta Superintelligence Labs, FAIR

- Led efforts on collective alignment and personalization.

Research Scientist Intern — June 2024 – August 2024
Apple Foundational Model Pre-training Team

- Researched online-offline gap in continual pre-training

Founding Member/Technical Advisor — April 2022 – January 2025
FAR (https://far.ai/)

- Advised research interns on language model safety projects.

Applied Scientist Intern — June 2021 – September 2021
Amazon

- Developed a probabilistic model and inference procedure for counterfactual estimation.

Publications

1. S. Karlekar*, **C. Shi***, C. Zheng, D. Blei, M. Makar, A. Manas Puli, J. Bowlan, M. Kucer What's in an Environment? Learning Representations in Language Models Robust to Distribution Shifts *Under submission at AISTATS 2026*
2. K. Ullrich, J. Su, **C. Shi**, A. Subramonian, A. Bar, I. Evtimov, N. Tsilivis, R. Balestriero, J. Kempe, M. Ibrahim OpenApps: Simulating Environment Variations to Measure UI-Agent Reliability *Under submission at ICLR 2026*
3. C. Zheng, N. Beltran-Velez, S. Karlekar, **C. Shi**, A. Nazaret, A. Mallik, A. Feder, D.Blei Model Directions, Not Words: Mechanistic Topic Models Using Sparse Autoencoders *Under submission at TACL*
4. **C. Shi***, N. Beltran*, A.Nazaret*, C.Zheng*, A. Garriga-Alonso, A. Jesson, M. Makar, D. Blei. Hypothesis Testing the Circuit Hypothesis in LLMs (*equal contribution) *Neural Information Processing Systems (NeurIPS) 2024.*
5. D. Banks, C. Bosone, B. Carpenter, T. Shah, **C. Shi**. Large language models: Trust and regulation *Harvard Data Science Review, 2024.*
6. M. Harrington, M. Alkon, X. He, **C. Shi**, R. Kennedy, J. Kopas, and J. Urpelainen. Anthropogenic Influences Increasingly Predict Groundwater Depletion Across India. *Environmental Research Letters, 2024.*

7. A. Nazaret, **C. Shi**, D. Blei. On Misspecification in Synthetic Controls. *Conference on Artificial Intelligence and Statistics (AISTATS), 2024. (Oral)*
8. N. Scherrer*, **C. Shi***, A. Feder, and D. Blei. Evaluating the Moral Beliefs Encoded in LLMs (*equal contribution) *Neural Information Processing Systems (NeurIPS), 2023.(Spotlight)*
9. A. Feder*, Y. Wald*, **C. Shi**, S. Saria, and D. Blei. Causal-structure Driven Augmentations for Text OOD Generalization (*equal contribution). *Neural Information Processing Systems (NeurIPS), 2023*
10. S. Casper, X. Davies, **C. Shi,**, T. K. Gilbert, J.Scheurer, J.Rando, ... & D. Hadfield-Menell (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *TMLR, 2023.*
11. C. Zheng*, **C. Shi***, K. Vafa, A. Feder, and D. Blei. An Invariant Learning Characterization of Controlled Text Generation (*equal contribution) *ACL, 2023.*
12. M. Yin, **C. Shi**, Y. Wang, and D. Blei. Conformal Sensitivity Analysis for Individual Treatment Effect. *Journal of the American Statistical Association (JASA), 2022.*
13. **C. Shi**, D. Sridhar, V. Misra, and D. Blei. On the Assumptions of Synthetic Control Methods. *Conference on Artificial Intelligence and Statistics (AISTATS), 2022. (Oral)*
14. **C.Shi**, V. Veitch, and D.Blei. 2021. Invariant Representation Learning for Treatment Effect Estimation. *Uncertainty in Artificial Intelligence (UAI), 2021. (Long talk)*
15. **C.Shi**, D.Blei, and V. Veitch. 2019. Adapting Neural Networks for the Estimation of Treatment Effects. *Neural Information Processing Systems (NeurIPS), 2019.*

Invited Talks

- Towards Science of LLMs — ML Lunch Seminar, UMass Amherst 2024
- Evaluating Moral Beliefs Encoded in LLMs — JSM invited session on LLMs and Statistics, 2024
- On the Assumptions and Misspecification of Synthetic Control Methods — CMStatistics 2023
- Evaluating Moral Beliefs Encoded in LLMs — New York Academy of Sciences – AI and Society Seminar, 2023
- Evaluating Moral Beliefs Encoded in LLMs — NYU CDS Lunch Seminar , 2023
- Evaluating Moral Beliefs Encoded in LLMs — Columbia workshop on text and LLMs , 2023
- Evaluating Moral Beliefs Encoded in LLMs — Center for Human Compatible AI workshop, 2023
- On the Assumptions of Synthetic Control Methods. — ICSA Applied Statistics Symposium, 2023
- On the Assumptions of Synthetic Control Methods. — American Causal Inference Conference, 2023
- On the Assumptions of Synthetic Control Methods. — Center for Causal Inference Seminar, Jan 2022
- Synthetic Controls, When do They Work and Why? — Amazon Causal Inference Group, July 2021
- Nearly Invariant Causal Estimation — Lyft Causal Inference Seminar, Feb 2021
- Adapting Neural Networks for Causal Estimation — Oxford Machine Learning and Statistics Seminar, July 2019

Awards & Grants

- **Best paper prize (3rd place) ICML Mechanistic Interpretability Workshop** 2024
- **Avanessians Doctoral Fellowship (Declined)** 2024
- **Center of AI Technology Ph.D. Fellowship** 2024
- **SFF Grant on beneficial AI ($60k)** 2020
- **Columbia CA Fellowship ($15k)** 2019
- **Tulane Presidential Scholarship ($120k)** 2014

Professional Activities

- **Machine learning NYC Speaker series** organizer, 2023 - 2025
- **NeurIPS workshop on Causal Representation Learning:** organizer, area chair 2023
- **ICML workshop on Spurious correlations, Invariance, and Stability:** organizer 2023
- **Machine Learning NYC Speaker Series:** Organizer of a city-wide speaker series and happy hours (2022 - Now)
- **Columbia ML Reading Group:** Organizer of Columbia ML reading group (2020 - 2022)
- **NeurIPS Social:** Organized Human Compatible AI social at NeurIPS 2019
- **Conference Reviewing:**
  - International Conference on Learning Representations (2021, 2022, 2023, 2025)
  - International Conference on Artificial Intelligence and Statistics (2022)
  - Neural Information Processing Systems (2020, 2021, 2023)
  - International Conference on Machine Learning (2020, 2022, 2024)