# Aahlad Manas Puli

aahlad@nyu.edu (929) 312-7360 aahladmanas Google Scholar

My research centers on **adaptation**: building AI systems that generalize beyond their training distribution, reason about the mechanisms that transfer across environments, and reveal what they rely on. As a Faculty Fellow at NYU, I develop the representations and algorithms that make adaptation reliable, working toward the broader goal of reliable world models. My work appears at NeurIPS, ICLR, ICML, and EMNLP, and I received NYU CS's Janet Fabri Prize for the best dissertation.

## Research

**Representations & mechanisms.** Generalization and shortcut learning [P2, P3, P4] • Faithfulness, introspection, and encoding [P1] • Causal inference [P10] • Meta-learning and black-box estimation [P9] • Multimodal representation alignment [P6]

**Modeling & architectures.** Long-context modeling [P7] • Transformer dynamics and retrieval [P8] • Flow maps [P5]

**Impact on science.** Equal learning for transportable clinical risk models [P8, P11] • LODE for genome-wide association studies [P10] • NURD for new-particle detection on LHC data [P12]

## Experience

**Center for Data Science, New York University** — **New York, USA**
*Faculty Fellow / Assistant Professor* — *Sep 2024 – Present*

- Proposed STRIPE-X, the first scalable detection method for introspection failures in LLMs, and proved the main results [P1]; in follow-up work, detected such failures in introspection traces from an 8B Llama model.
- Co-supervised the development of CAT, a controllably efficient architecture that matches 10 efficient alternatives from a single trained model at 1.4–3.7× the generation throughput of a dense transformer, and scales to 1B parameters [P7].
- Extended my margin control method [P2] to design equal learning for LLMs [P8] and improve clinical risk models [P11].
- Co-designed the first meta-learning approach to learn causal estimators from data, enabling black-box causal inference [P9].
- Contributed to the study of convergence for flow-map learning without distillation [P5], and the experimental design for Symile, a model-agnostic multimodal alignment method that outperforms CLIP-based objectives by 12.5% [P6].
- Delivered the NeurIPS 2024 tutorial on spurious correlations, and ran the CDS capstone project course (>100 students).

**Courant Institute, New York University** — **New York, USA**
*PhD Researcher — Advisor: Rajesh Ranganath* — *Sep 2018 – Sep 2024*

- Developed NURD, a method for out-of-distribution generalization with optimality guarantees [P3], and semantic corruptions, a data-augmentation approach that adjusts away *unknown* nuisances [P4].
- Developed *margin control* to mitigate shortcut learning that arises from gradient-based training with cross entropy [P2].
- Proposed LODE, a scalable causal estimation algorithm to handle functionally-specified confounding, and applied it to genome-wide association studies [P10].
- Led the experimental program for the first EHR-scale twinned survival study for coronary heart disease, cutting prediction error by 20% against the current AHA-standard risk model [P11].
- Helped adapt NURD to enable robust new-particle detection in Large Hadron Collider data [P12].

**Adobe Research, Data Science Lab** — **San Jose, USA**
*Research Intern* — *Summer 2019*

- Proposed a new Bayesian uplift modeling method for marketing attribution problems that makes use of order information.

**Biomedical Informatics, Columbia University** — **New York, USA**
*Software Developer* — *Jul 2017 – Jun 2018*

- Contributed to Columbia DBMI's data pipeline for the NIH *All of Us* research program on Google Cloud and BigQuery, enabling secure and scalable access to EHR data for >100 researchers.

## Education

**Ph.D. in Computer Science**, New York University 2018–2024
**M.S. in Computer Science**, New York University 2015–2017
**B.Tech. + M.Tech. in Electrical Engineering**, Indian Institute of Technology Madras 2010–2015

## Invited Talks

**Out-of-Distribution Generalization: Shortcuts, Spuriousness, and Stability** — Tutorial, NeurIPS 2024.
**The Spectre of Spurious Correlations in OOD Generalization and Interpretability** — Fermilab, 2026.
**Making the most of your data for Reliable AI** — Columbia DBMI, UCLA, UC Irvine, UC Riverside, 2026.
**Predictive Modeling under Nuisance-Induced Spurious Correlations** — Trustworthy ML Rising Star Seminar, 2021.

## Selected Work

[P1] **Explanations that Reveal All through the Definition of Encoding** — *NeurIPS 2024*
**Aahlad Puli**[*], Nhi Nguyen[*], Rajesh Ranganath. ([*] equal contribution)

[P2] **Don't Blame Dataset Shift! Shortcut Learning due to Gradients and Cross Entropy** — *NeurIPS 2023*
**Aahlad Puli**, Lily Zhang, Yoav Wald, Rajesh Ranganath.

[P3] **OOD Generalization in the Presence of Nuisance-Induced Spurious Correlations** — *ICLR 2022*
**Aahlad Puli**, Lily Zhang, Eric Oermann, Rajesh Ranganath.

[P4] **Nuisances via Negativa: Adjusting for Spurious Correlations via Data Augmentation** — *TMLR 2024*
**Aahlad Puli**, Nitish Joshi, Yoav Wald, He He, Rajesh Ranganath.

[P5] **Flow Map Learning via Non-Gradient Vector Flow** — *ICLR 2026*
Mark Goldstein, Anshuk Uppal, Raghav Singhal, **Aahlad Puli**, Rajesh Ranganath.

[P6] **Contrasting with Symile: Simple Model-Agnostic Representation Learning for Unlimited Modalities** — *NeurIPS 2024*
Adriel Saporta, **Aahlad Puli**, Mark Goldstein, Rajesh Ranganath.

[P7] **Attention and Compression is all you need for Controllably Efficient Language Models** — *In submission, 2026*
Jatin Prakash, **Aahlad Puli**, Rajesh Ranganath.

[P8] **Learning Is Not A Race: Improving Retrieval in Language Models via Equal Learning** — *EMNLP 2025*
Wanqian Yang, **Aahlad Puli**, Rajesh Ranganath.

[P9] **Black Box Causal Inference: Effect Estimation via Meta Prediction** — *arXiv:2503.05985, 2025*
Lucius E.J. Bynum[*], **Aahlad Puli**[*], Diego Herrero-Quevedo, Nhi Nguyen, Carlos Fernandez-Granda, Kyunghyun Cho, Rajesh Ranganath. ([*] equal contribution)

[P10] **Causal Estimation with Functional Confounders** — *NeurIPS 2020*
**Aahlad Puli**, Adler J. Perotte, Rajesh Ranganath.

[P11] **Why Risk It? Performant and Transportable Risk Models for Coronary Heart Disease from EHR Data** — *In preparation, 2026*
**Aahlad Puli**[*], Shreyas Bhave[*], Mark Goldstein[*], Mert Ketenci[*], Pierre Elias, Sumit Chopra, Glenn Fishman, Stuart Katz, John Dodson, Yindalon Aphinyanaphongs, Shalmali Joshi, Noémie Elhadad, Adler Perotte, Rajesh Ranganath.

[P12] **Robust Anomaly Detection for Particle Physics Using Multi-Background Representation Learning** — *Machine Learning: Science and Technology, 2024*
Abhijith Gandrakota, Lily Zhang, **Aahlad Puli**, Kyle Cranmer, Jennifer Ngadiuba, Rajesh Ranganath, Nhan Tran.

## Awards

- Janet Fabri Prize — Outstanding PhD Dissertation, NYU Computer Science (2025)
- Apple Scholars in AI/ML Fellowship (2022)
- Trustworthy ML Rising Star (2021)
- MacCracken Doctoral Fellowship (2018)

## Academic Service

**Organizer:** Spurious Correlations, Invariance and Stability (SCIS) Workshop, ICML 2022–2023.
**Reviewer:** NeurIPS, ICLR, ICML, AISTATS, JMLR, TMLR, MLHC, UAI (Top Reviewer, UAI 2022).

## Technical Skills

**Languages & tools:** Python, PyTorch
**Training & systems:** distributed and multi-GPU training, LLM pre-training and fine-tuning, inference efficiency
**Modeling:** transformers and sequence models, flow-based generative models