Stanford

Azalia Mirhoseini

Assistant Professor of Computer Science

Bio

BIO

Azalia Mirhoseini is an Assistant Professor of Computer Science at Stanford University where she directs Scaling Intelligence, a research lab focused on developing scalable and self-improving AI systems. She is a co-founder of Ricursive Intelligence, a frontier lab dedicated to recursive self-improvement through AI that designs the chips that fuel it. Previously, she spent several years in industry AI labs, including Google Brain, Anthropic, and Google DeepMind, working on Gemini and Claude, among other projects. Her past work includes Mixture-of-Experts (MoE) neural architectures, now predominantly used in leading frontier AI models; AlphaChip, a reinforcement learning approach for layout optimization used in the design of advanced chips like Google AI accelerators (TPUs) and data center CPUs; as well as pioneering research on LLM Test-Time Scaling. Her work has been recognized through the Okawa Research Grant, the Google ML and Systems Junior Faculty Award, MIT Technology Review's 35 Under 35 Award, the Best ECE Thesis Award at Rice University, publications in flagship venues such as Nature, and coverage by various media outlets, including WSJ, NYT, Forbes, MIT Technology Review, IEEE Spectrum, WIRED, and TechCrunch.

ACADEMIC APPOINTMENTS

- Assistant Professor, Computer Science

PROGRAM AFFILIATIONS

- Stanford SystemX Alliance

LINKS

- http://azaliamirhoseini.com/: http://azaliamirhoseini.com/
- Google Scholar: https://scholar.google.com/citations?user=I1nZzWsAAAAJ&hl=en&oi=ao

Teaching

COURSES

2026-27

- Self Improving AI Agents: CS 329A (Spr)

2025-26

- Self Improving AI Agents: CS 329A (Aut)

2024-25

- Self Improving AI Agents: CS 329A (Win)
- Systems for Machine Learning: CS 229S (Aut)

2023-24

- Systems for Machine Learning: CS 229S (Aut)

STANFORD ADVISEES

Doctoral Dissertation Reader (AC)

Avanika Narayan, Qizheng Zhang

Orals Evaluator

Avanika Narayan, Zhiqiang Xie

Doctoral Dissertation Advisor (AC)

Jon Saad-Falcon

Master's Program Advisor

Armaan Abraham, Gia Ancone, Sarah Barragan, Gabriel Bo, William Briger, Jett Carruth, Manat Kaur, Brandon Liu, Nora Menon, Sid Potti, Arnuv Tandon, Andy Wang, Berk Yalcinkaya, Daniel Yang, Shirley Yu

Doctoral Dissertation Co-Advisor (AC)

Hermann Kumbong, Jacky Kwok, Yuzhen Mao

Doctoral (Program)

Bradley Brown, Simon Guo, Jordan Juravsky, Anne Ouyang, Jon Saad-Falcon

Publications

PUBLICATIONS

- **Training of physical neural networks.** *Nature*
  Momeni, A., Rahmani, B., Scellier, B., Wright, L. G., McMahon, P. L., Wanjura, C. C., Li, Y., Skalli, A., Berloff, N. G., Onodera, T., Oguz, I., Morichetti, F., Del Hougne, et al
  2025; 645 (8079): 53-61

- **An Architecture Search Framework for Inference-Time Techniques**
  Saad-Falcon, J., Lafuente, A., Natarajan, S., Maru, N., Todorov, H., Guha, E., Buchanan, E., Chen, M., Guha, N., Re, C., Mirhoseini, A.
  edited by Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2025: 52475-52507

- **RoboMonkey: Scaling Test-Time Sampling and Verification for Vision-Language-Action Models**
  Kwok, J., Agia, C., Sinha, R., Foutter, M., Li, S., Stoica, I., Mirhoseini, A., Payone, M.
  edited by Lim, J., Song, S., Park, H. W.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2025: 3200-3217

- **Think, Prune, Train: Can Small Models Teach Themselves to Reason?**
  Costello, C., Guo, S., Goldie, A., Mirhoseini, A., IEEE COMPUTER SOC
  IEEE COMPUTER SOC.2025: 195-203

- **How Do Large Language Monkeys Get Their Power (Laws)?**
  Schaeffer, R., Kazdan, J., Hughes, J., Juravsky, J., Price, S., Lynch, A., Jones, E., Kirk, R., Mirhoseini, A., Koyejo, S.
  edited by Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2025: 53132-53176

- **KernelBench: Can LLMs Write Efficient GPU Kernels?**
  Ouyang, A., Guo, S., Arora, S., Zhang, A. L., Hu, W., Re, C., Mirhoseini, A.
  edited by Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2025: 47356-47415

- **Addendum: A graph placement methodology for fast chip design.** *Nature*
  Goldie, A., Mirhoseini, A., Yazgan, M., Jiang, J. W., Songhori, E., Wang, S., Lee, Y., Johnson, E., Pathak, O., Nova, A., Pak, J., Tong, A., Srinivasa, et al
  2024

- **Embroid: Unsupervised Prediction Smoothing Can Improve Few-Shot Classification**
  Guha, N., Chen, M. F., Bhatia, K., Mirhoseini, A., Sala, F., Re, C.
  edited by Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S.
  NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).2023

- **A Full-Stack Search Technique for Domain Optimized Deep Learning Accelerators**
  Zhang, D., Huda, S., Songhori, E., Prabhu, K., Quoc Le, Goldie, A., Mirhoseini, A.
  edited by Falsafi, B., Ferdman, M., Lu, S., Weinisch, T.
  ASSOC COMPUTING MACHINERY.2022: 27-42

- **A graph placement methodology for fast chip design.** *Nature*
  Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J. W., Songhori, E., Wang, S., Lee, Y., Johnson, E., Pathak, O., Nazi, A., Pak, J., Tong, A., Srinivasa, et al
  2021; 594 (7862): 207-212