

Nathan Hu
nathu@cs.stanford.edu

Education

2025–Present	Stanford University Ph.D. Computer Science
2023–2024	Stanford University M.S. Computer Science
2019–2023	Stanford University B.S. Computer Science

Research Experience

MATS Program Research Scholar [with Lee Sharkey]	Jun 2025-Aug 2025
Anthropic Research Resident [with Evan Hubinger]	Sep 2024-Jan 2025
Google DeepMind Student Researcher [with Quoc V. Le]	Feb 2024-Apr 2024
Stanford IRIS Lab Student Researcher [with Eric Mitchell and Chelsea Finn]	Jan 2022-Feb 2024
Redwood Research Research Resident [with Alexandre Variengien]	Dec 2022-Jan 2023
Stanford CURIS Student Researcher [with Nima Anari]	Jun 2020-Aug 2020

Research Output

Nathan Hu, Benjamin Wright, Carson Denison, Samuel Marks, Johannes Treutlein, Jonathan Uesato, Evan Hubinger. Training on Documents About Reward Hacking Induces Reward Hacking. *Anthropic Alignment Science Blog*. 2025.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, **Nathan Hu**, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, Quoc V. Le. Long-form factuality in large language models. *NeurIPS*. 2024.

Z. Jonny Kong, **Nathan Hu**, Y. Charlie Hu, Jiayi Meng, Yaron Koral. High-Fidelity Cellular Network Control-Plane Traffic Generation without Domain Knowledge. *ACM IMC*. 2024.

Nathan Hu, Eric Mitchell, Christopher D. Manning, Chelsea Finn. Meta-Learning Online Adaptation of Language Models. *EMNLP*. 2023.

Nima Anari, **Nathan Hu**, Amin Saberi, Aaron Schild. Sampling Arborescence in Parallel. *ITCS*. 2021.

Additional Experience

Tower Research Capital Quantitative Trading Intern	Jun 2023-Aug 2023
Jane Street Quantitative Trading Intern	Jun 2021-Aug 2021
Apple Data Science Intern, OS Power and Performance Team	Jan 2021-May 2021