Simran Arora (`https://arorasimran.com/`)

Overview

My research in **AI systems** focuses on expanding the Pareto frontier between quality and efficiency, to unlock new AI capabilities. I study AI algorithms, hardware, and applications in lockstep.

Education

2019-2025	**PhD in Computer Science**, *Stanford University* Research advisor: Christopher Ré
2015-2019	**Jerome Fisher Management & Technology Program**, *University of Pennsylvania* Summa cum laude BSE in Computer Science, School of Engineering Summa cum laude BS in Economics (Finance Concentration), The Wharton School Minor in Mathematics Research advisors: Boon Thau Loo, Vijay Kumar, Vincent Liu

Experience

Current	**PhD Student**, *Stanford University*
Current	**Academic Partner**, *Together AI*
Current	**Advisor**, *Cartesia AI*
Current	**Academic Partner**, *Looma Education, AI*
2021-2022	**Research Scientist**, *Facebook AI Research* Collaborated with Jacob Kahn, Patrick Lewis, Angela Fan, and Ronan Collobert Work published at TACL
Summer 2018	**Technology Investment Banking**, *Morgan Stanley, Menlo Park* Worked on the sale of Acxiom AMS to IPG for $2.3Bn, sale of Cylance to Blackberry for $ 1.7Bn, and Sonos IPO on Nasdaq
Summer 2017	**Software Engineering**, *Google* JavaScript Open Source Compiler, GitHub closure-compiler

Awards

2025	**ICML ES-FoMo Oral Award (Top 5 of 146 papers)** Cartridges: Lightweight long context representations via self-study
2025	**ICLR DL4C Best Paper Award (Amongst 63 papers)** KernelBench: Can LLMs Write Efficient GPU Kernels?
2025	**ICLR Spotlight Award (Top 5.1% of 11.7K Papers)** ThunderKittens: Simple, Fast, and Adorable AI Kernels
2025	**Stanford Computer Science Graduate Fellowship, 1-year**
2024	**ICML ES-FoMo Best Paper Award (Amongst 83 Papers)** Simple linear attention language models balance the recall-throughput tradeoff
2024	**ICML Spotlight Award (Top 3.5% of 10K Submitted Papers)** Simple linear attention language models balance the recall-throughput tradeoff
2023	**NeurIPS Outstanding Paper Award (4 Papers in 12.3K Submissions)** DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models
2023	**NeurIPS Oral Award (Top 0.5% of Papers)** DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models
2023	**NeurIPS Oral Award (Top 0.5% of 12.3K Papers)** Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture

2023	**ICLR Spotlight Award (Top 5% of 5K Papers)** Ask Me Anything: A simple strategy for prompting language models
2023	**AAAI KnowledgeNLP Workshop: Oral Award (Top 20% of Papers)** Reasoning over Public and Private Data in Retrieval-Based Systems
2019	**Stanford Graduate Fellowship, 3 years**
2019	**Rhodes Scholarship National Finalist**
2019	**Marshall Scholarship National Finalist**
2019	**Penn Computer Science Academic Award** One graduating CS major per year
2019	**Michele Huber and Bryan D. Giles Memorial Award** One graduating Jerome Fisher student per year
2019	**Penn Computer Science Senior Engineering Capstone Project 2nd Place**
2019	**Wharton School Summa Cum Laude (Highest Honors)**
2019	**Penn Engineering Summa Cum Laude (Highest Honors)**
2017	**Best Paper Runner Up: IEEE MARSS Conference (Top 3 papers)** Control of multiple microrobots with multiscale magnetic field superposition
2017	**University of Pennsylvania Tau Beta Pi and Eta Kappa Nu**
2015	**International University Physics Competition (Top 20%, Link)**

Select Open Source Artifacts

ThunderKittens GitHub (Link, 2.7K+ stars); EVAPORATE GitHub (Link, 500 stars); Bootleg GitHub (Link, 200+ stars); ConcurrentQA GitHub (Link, first benchmark and system for multi-distribution retrieval); Based GitHub (Link, 200+ stars), LoLCATS GitHub (Link, 200+ stars), Zoology GitHub (Link, 200+ stars), Monarch Mixer GitHub (Link, 500+ stars); M2-BERT Retrieval Model Checkpoints (Link, 70K+ downloads), Ask Me Anything GitHub (Link, 500+ stars); KernelBench GitHub (Link, 500+ stars), Recall-intensive benchmarks for sub-quadratic architectures (Link, 24K+ downloads); Benchmarks for long-context retrieval (Link, 135K+ downloads); On the Opportunities and Risks of Foundation Models white paper (Link, 4000+ Citations)

Select Public Industry Use and Press

Bootleg (Apple, Link; ZDNet Link); Ask Me Anything (Forbes Link; Snorkel AI, Link; Samba Nova Link; Numbers Station, Link, Aleksa Gordić Link); EVAPORATE (LlamaIndex, Link); BASED / LoLCATS (Hugging Face Candle integration Link; NVIDIA, Link; Together AI, Link), Monarch Mixer (Mongo DB Link; LangChain; LlamaIndex; Together AI Link; Nomic AI, Link); ConcurrentQA (Meta, Link); Privacy (Venture Beat, Link); Data Wrangling (Numbers Station, Link); KernelBench (NVIDIA, Link; METR Link, Sakana AI Link, Cognition Link); ThunderKittens (Together AI, Link; Cruise, Link, Cursor, Link)

Publications

[1] Sabri Eyuboglu, Ryan Ehrlich, Simran Arora, Neel Guha, Dylan Zinsley, Emily Liu, Will Tennien, Atri Rudra, James Zou, Azalia Mirhoseini, Christopher Ré
Cartridges: Lightweight and general-purpose long context representations via self-study
ICML ES-FoMo 2025
Oral Award (Top 5 of 146 papers)
Paper Link / GitHub Link

[2] Anne Ouyang, Simon Guo, Simran Arora, Alex L. Zhang, William Hu, Christopher Ré, Azalia Mirhoseini
KernelBench: Can LLMs Write Efficient GPU Kernels?
International Conference on Machine Learning (ICML) 2025
ICLR DL4C 2025
Best Paper Award
Paper Link / GitHub Link

[3] Benjamin Spector, Simran Arora, Aaryan Singhal, Daniel Fu, Christopher Ré
ThunderKittens: Simple, Fast, and Adorable AI Kernels
International Conference on Learning Representations (ICLR) 2025
Spotlight Award (Top 5.1% of 11.6K papers)
Paper Link / GitHub Link

[4] Michael Zhang, Simran Arora, Rahul Chalamala, Benjamin Spector, Alan Wu, Krithik Ramesh, Aaryan Singhal, Christopher Ré
LoLCATS: Low-rank Linearization of Large Language Models
International Conference on Learning Representations (ICLR) 2025
Paper Link / GitHub Link

[5] Jerry Liu, Jessica Grogan, Owen Dugan, Simran Arora, Atri Rudra, and Christopher Ré
Towards Learning High-Precision Least Squares Algorithms with Sequence Models
International Conference on Learning Representations (ICLR) 2025
Paper Link

[6] Simran Arora, Aman Timalsina, Aaryan Singhal, Benjamin Spector, Sabri Eyuboglu, Xinyi Zhao, Ashish Rao, Atri Rudra, and Christopher Ré
Just read twice: closing the recall gap for recurrent language models
ICML ES-FoMo 2025
Paper Link / GitHub Link

[7] Megha Srivastava, Simran Arora, and Dan Boneh
Optimistic Verifiable Training by Controlling Hardware Nondeterminism,
Advances in Neural Information Processing Systems (NeurIPS) 2024
ICML ES-FoMo 2024
Paper Link / GitHub Link

[8] Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, Christopher Ré
Simple linear attention language models balance the recall-throughput tradeoff
International Conference on Machine Learning (ICML) 2024
Spotlight Award (Top 3.5% of 10K papers)
ICML ES-FoMo 2024
Best Paper Award (1 paper)
Paper Link / GitHub Link

[9] Jon Saad-Falcon, Daniel Y. Fu, Simran Arora, Neel Guha, Christopher Ré
Benchmarking and Building Long-Context Retrieval Models with LoCo and M2-BERT
International Conference on Machine Learning (ICML) 2024
ICML ES-FoMo 2024
Paper Link / GitHub Link

[10] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, Mennatallah El-Assady
RELIC: Investigating Large Language Model Responses using Self-Consistency
Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI) 2024
Paper Link

[11] Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, Christopher Ré
Zoology: Measuring and Improving Recall in Efficient Language Models
International Conference on Learning Representations (ICLR) 2024
Paper Link / GitHub Link

[12] Daniel Y. Fu, Simran Arora, Jessica Grogan, Isys Johnson, Sabri Eyuboglu, Armin W. Thomas, Benjamin F. Spector, Michael Poli, Atri Rudra, Christopher Ré
Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture
Advances in Neural Information Processing Systems (NeurIPS) 2023
Oral Award (Top 0.5% of 12.3K papers)

Daniel Y. Fu* and Simran Arora*, Revisiting BERT, Without Attention or MLPs, Link
Paper Link / GitHub Link

[13] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, Bo Li
DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models
Advances in Neural Information Processing Systems (NeurIPS) 2023
Oral Award (Top 1% of 1K papers)
Outstanding Paper Award (Top 2 papers)
Paper Link / Benchmark Link

[14] Simran Arora, Brandon Yang, Sabri Eyuboglu, Avanika Narayan, Andrew Hojel, Immanuel Trummer, Christopher Ré
Language Models Enable Simple Systems for Generating Structured Views of Data Lakes
Proceedings of the VLDB Endowment (PVLDB) 2023.
Paper Link / GitHub Link

[15] Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, Christopher Ré
Reasoning over Public and Private Data in Retrieval-Based Systems
Transactions of the Association for Computational Linguistics (TACL) 2023
AAAI KnowledgeNLP 2023
Oral Award (Top 15% of papers)
Paper Link / GitHub Link

[16] Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, Christopher Ré
Ask Me Anything: A simple strategy for prompting language models
International Conference on Learning Representations (ICLR) 2023
Spotlight Award (Top 5% of 5K papers)
Paper Link / GitHub Link

[17] Simran Arora, Sen Wu, Enci Liu, Christopher Re
Metadata shaping: A simple approach for knowledge-enhanced language models
Findings of the Association for Computational Linguistics (ACL) 2022
Paper Link / GitHub Link

[18] Avanika Narayan, Laurel Orr, Ines Chami, Simran Arora, Christopher Ré
Can Foundation Models Wrangle Your Data?
Proceedings of the VLDB Endowment (PVLDB) 2022
Paper Link / GitHub Link

[19] Laurel Orr, Megan Leszczynski, Simran Arora, Sen Wu, Neel Guha, Xiao Ling, Christopher Re
Bootleg: Chasing the Tail with Self-Supervised Named Entity Disambiguation
Conference on Innovative Data Systems Research (CIDR) 2021
Paper Link / GitHub Link

[20] Simran Arora, Avner May, Jian Zhang, Christopher Ré
Contextual Embeddings: When Are They Worth It?
Proceedings of the Association for Computational Linguistics (ACL) 2020
Paper Link / GitHub Link

[21] Qizhen Zhang, Akash Acharya, Hongzhi Chen, Simran Arora, Ang Chen, Vincent Liu, Boon Thau Loo
Optimizing Declarative Graph Queries at Large Scale
Proceedings of the International Conference on Management of Data (SIGMOD) 2019
Paper Link

[22] Edward Steager, Denise Wong, Jeremy Wang, Simran Arora, Vijay Kumar
Control of multiple microrobots with multiscale magnetic field superposition
International Conference on Manipulation, Automation and Robotics at Small Scales (MARSS) 2017
Best Paper Runner Up (Top 3 papers)
Paper Link

[23] B. P. Mason, M. Whittaker, J. Hemmer, Simran Arora, A. Harper, S. Alnemrat, A. McEachen, S. Helmy, J. Read de Alaniz, J. P. Hooper
A temperature-mapping molecular sensor for polyurethane-based elastomers
Applied Physics Letters (APL) 2016
Paper Link

Workshop

[1] Sabri Eyuboglu, Dylan Zinsley, Jon Saad-Falcon, Simran Arora, Atri Rudra, James Zou, Chris Ré
Towards smaller language models via layer looping
ICML ES-FoMo 2024
Paper Link

[2] Simran Arora and Christopher Ré
Can Foundation Models Help Us Achieve Perfect Secrecy?
AAAI PPAI Workshop 2023
Paper Link / GitHub Link

Invited Talks

2025	Scale ML x GPUMode Lecture (Link; Virtual)
2025	Stanford CS229S Guest Lecture (Stanford, CA)
2025	ML Sys YPS Keynote Speaker (Santa Clara, CA)
2025	University of Toronto (Toronto, Canada)
2025	Massachusets Institute of Technology (Cambridge, MA)
2025	New York University (NYC, NY)
2025	Princeton University (Princeton, NJ)
2025	University of Pennsylvania (Philadelphia, PA)
2025	University of Washington (Seattle, WA)
2025	UCLA CS & ECE (Los Angeles, CA)
2025	University of Buffalo (Buffalo, NY)
2025	Columbia CS & EE (NYC, NY)
2025	Cornell University (Ithaca, NY)
2025	Cornell Tech (NYC, NY)
2025	Caltech (Pasedena, California)
2025	Rice University (Houston, Texas)
2025	Northwestern University (Evanston, Illinois)
2025	University of Michigan (Ann Arbor, Michigan)
2025	NVIDIA GTC @ GPU Mode (San Jose, CA)
2025	NVIDIA Reading Group (Virtual)
2024	NeurIPS Hacker-Cup AI (Link; Workshop Keynote Speaker; Vancouver, BC)
2024	Simons Institute: Are Transformers the end game? (Link; Berkeley, CA) *Panel discussion with Jitendra Malik, Stella Biderman, Andrew Gordon Wilson*
2024	Simons Institute: Transformers as a Computational Model (Link; Berkeley, CA)
2024	Stanford NLP Group (Stanford, CA)
2024	UC Berkeley NLP Group (Berkeley, CA)
2024	CCAIM Summer School (Link; Virtual)
2024	Liquid AI (Vienna, Austria)
2024	Princeton University PLI Group (Link; Princeton, NJ)
2024	Cornell Tech (New York, NY)
2024	Microsoft AI Research (Virtual)
2024	56th Annual ACM Symposium on Theory of Computing (Link; Workshop Keynote Speaker; Vancouver, Canada)
2023	NeurIPS 3rd Table Representation Learning Workshop (Link; Workshop Keynote Speaker; New Orleans, LA)
2023	Snorkel Foundation Model Summit (Virtual)
2023	Apple Machine Learning Research Reading Group (Cupertino, CA)
2023	ICLR Spotlight Presentation (Kigali, Rwanda)
2023	Stanford CRFM Research Spotlight Talk (Stanford, CA)
2022	IBM AI Research Reading Group (Virtual)
2022	MIT Computational Social Science Reading Group (Virtual)
2022	Stanford HAI: AI and Society (Stanford, CA)
2022	Oral at KnowledgeNLP-AAAI (Washington DC)
2021	Facebook AI Research Reading Group (Virtual)
2021	Spotlight at Stanford HAI Data-Centric AI Workshop (Virtual)
2020	ACL Conference (Virtual)
2020	Stanford DAWN Retreat (Virtual)

Teaching

Summer 2025	**Course Assistant**, *CS 229: Machine Learning* Stanford University
Fall 2023	**Course Co-Creator and Co-Instructor**, *CS 229S: Systems for Machine Learning* Stanford University 3-Unit Undergrad-Graduate course, Taught 110+ students.
Fall 2023	**Instructor** *CS: 528: Machine Learning Systems Seminar* Stanford University
Spring 2019	**Course Co-Creator**, *MCIT 595: Computer Systems* University of Pennsylvania
Fall 2018	**Course Assistant**, *CIS 380: Operating Systems* University of Pennsylvania
Spring 2018	**Course Assistant**, *CIS 160: Discrete Mathematics* University of Pennsylvania
Fall 2017	**Course Assistant**, *CIS 160: Discrete Mathematics* University of Pennsylvania

Educational Notes

- CS 229s Systems for ML (Link, course lecture notes)
- Efficient architectures as arithmetic circuits (Link, blog)
- ThunderKittens: Bringing fp8 to theaters near you (Link, blog)
- ThunderKittens: Easier, better, faster, cuter (Link, blog)
- ThunderKittens: GPUs Go Brrr (Link, blog)
- Linearizing LLMs with LoLCATS (Link, blog)
- Long-Context Retrieval Models with Monarch Mixer (Link, blog)
- Announcing LoCoV1 and the Latest M2-BERT Models (Link, blog)
- Just read twice: closing the recall gap for recurrent language models (Link, blog)
- Based: Simple linear attention language models balance the recall-throughput tradeoff (Link, blog)
- Zoology: Measuring and Improving Recall in Efficient Language Models (Link, blog)
- Monarch Mixer: Revisiting BERT, Without Attention or MLPs (Link, blog)
- The Safari of Deep Signal Processing: Hyena and Beyond (Link, blog)
- Building Blocks for AI Systems (Link, GitHub 300+ stars)
- On the Opportunities and Risks of Foundation Models, White paper (Link, 4000+ Citations)

Service

Ongoing	ICML (Top Reviewer Award), NeurIPS, ACL, PPAI-AAAI, NeurIPS TRL, ICLR ME-FoMo, ICML ES-FoMo
Ongoing	Looma AI Volunteer
2025	ICML ES-FoMo Workshop Organizer (Link)
2025	NeurIPS Efficient Reasoning Workshop Organizer (Link)
2022-2023	East Palo Alto Academy Foundation Volunteer
2023	Department (Stanford NLP Group Summer meetings, CRFM Leadership)
2018-2021	Undergrad Mentor (Stanford Women in STEM, Penn Women in CS)
2015-2017	UPenn Women in Physics Group Co-founder and Leadership / President