Shayne Longpre

Email: shayne.r.longpre@gmail.com
Web: shaynelongpre.com

Summary Research scientist at Anthropic building AI models and studying their impacts on the world. Founder of the Data Provenance Initiative, MIT Technology Review Innovator Under 35 (2026), and recipient of four best/outstanding paper awards at major AI conferences; writes for broader audiences about how technology evolves and should be governed.

Education

Massachusetts Institute of Technology, Cambridge, Massachusetts — 2021 - 2026
Ph.D. in Media Arts & Sciences
Advisor: Prof. Sandy Pentland

Stanford University, Palo Alto, California — 2012 - 2018
M.S. in Computer Science, Artificial Intelligence
B.A. in Economics, minor in History
Advisors: Chris Manning and Danqi Chen

Research & Industry Experience

Summary: A machine learning scientist and engineer at several leading research labs: Anthropic, Google, Cohere, Apple, BigScience, & Stanford. I founded the Data Provenance Initiative, a volunteer research organization with 50+ members.

Anthropic, *Member of Technical Staff* — 2026 – Present

The Data Provenance Initiative, *Founder & Lead* — 2023 - Present
A collective of AI researchers passionate about auditing AI data, and AI ecosystems. Now spans 50+ contributors, from 20+ countries.
Research Advisors: Sara Hooker, Stella Biderman, Sandy Pentland

Google Deepmind / Cloud, *Google Student Researcher* — 2024 - 2025
Research Advisors: Sayna Ebrahimi

Cohere 4 AI, Aya Open Science Initiative Co-Lead — 2023 - 2024
Research Advisors: Sara Hooker

Google Brain, *Google Student Researcher* — 2022
Research Advisors: Jason Wei, Barret Zoph, Denny Zhou, & Adam Roberts

BigScience, *Volunteer Contributor* — 2022
BLOOM [35] & ROOTS [34] Teams

Apple, *Senior Applied Machine Learning Scientist* — 2018 - 2021
Siri & Information Intelligence Team
Research Advisor: Chris Dubois

Stanford NLP Group, *Research Assistant* — 2016 - 2017
Research Advisors: Chris Manning & Danqi Chen

Salesforce AI Research, *Deep Learning Research Intern* — 2016
Research Advisors: Richard Socher & Caiming Xiong

Select AI Publications

Summary: Since 2021, I've published 43 AI conference/journal papers, 17 of which were first-authored. Publication venue range includes Science, Nature, AI, NLP, law, and economics journals, as well as op-eds, policy briefs, open letters, and comments to the US Copyright Office. My research has received 6 Best or Outstanding Paper Awards: (ICML RLxF Workshop 2026, NAACL [2024, 2025], ACL 2024, TMLR [2024, 2025]), 4 conference Oral presentations, and 4 Spotlights. This has culminated in 18k+ citations and an h-index of 39.

[1] The AI Observatory: A Public Measure of Real-World AI Use
Shayne Longpre, Anka Reuel, Dayeon Ki, Cathy Mengying Fang, Megan Richards, Zhiping Zhang, Chuanyang Jin, Jenn Mickel, Cedric Deslandes Whitney, Ahmet Üstün, Niloofar Mireshghallah, Victor Ojewale, Ariel Lee, Alex Pentland, Tianshi Li, Yuntian Deng, Mykel Kochenderfer, Sara Hooker, Sanmi Koyejo
Preprint 2026, ***[Featured in WaPo (+2)]***

[2] FLARE-AI: Flaw Reporting for AI
Shayne Longpre, Elaine Zhu, Carson Ezell, Avijit Ghosh, Sean McGregor, Kevin Paeth, Kevin Klyman, Sayash Kapoor, Rishi Bommasani, Ruth Appel, Gregory Strom, Lauren McIlvenny, Mark M. Jaycox, Peter Slattery, Nathan Butters, Arvind Narayanan, Percy Liang, Alex Pentland
ICML 2026, ***[Featured in Wired]***

[3] ThoughtTrace: Understanding User Thoughts in Real-World LLM Interactions
Chuanyang Jin, Binze Li, Haopeng Xie, Cathy Mengying Fang, Tianjian Li, **Shayne Longpre**, Hongxiang Gu, Maximillian Chen, Tianmin Shu
ArXiv 2026, Best Paper Award at ICML 2026 RLxF Workshop

[4] ATLAS: Adaptive Transfer Scaling Laws for Multilingual Pretraining, Finetuning, and Decoding the Curse of Multilinguality
Shayne Longpre, Sneha Kudugunta, Niklas Meunnighoff, I-Hung Hsu, Sandy Pentland, Chen-Yu Lee, Sercan Arik, Sayna Ebrahimi
ICLR 2026, ***[Featured in Google Research]***

[5] FlexOlmo: Open Language Models for Flexible Data Use
Weijia Shi, Akshita Bhagia, Kevin Farhat, Niklas Muennighoff, Pete Walsh, Jacob Morrison, Dustin Schwenk, **Shayne Longpre**, Jake Poznanski, Allyson Ettinger, Daogao Liu, Margaret Li, Dirk Groeneveld, Mike Lewis, Wen-tau Yih, Luca Soldaini, Kyle Lo, Noah A. Smith, Luke Zettlemoyer, Pang Wei Koh, Hannaneh Hajishirzi, Ali Farhadi, Sewon Min
NeurIPS 2025, Spotlight, ***[Featured in Wired]***

[6] Establishing Best Practices for Building Rigorous Agentic Benchmarks
Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, **Shayne Longpre**, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Jasjeet Sekhon, Jacob Steinhardt, Antony Kellermann, Sarah Schwettmann, Matei Zaharia, Ion Stoica, Percy Liang, Daniel Kang
NeurIPS 2025, 1st Place at Berkeley AgentX Summit

[7] The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text
Nikhil Kandpal, Brian Lester, Colin Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A. Feder Cooper, Aviya Skowron, John Kirchenbauer, **Shayne Longpre**, Lintang Sutawika, Alon Albalak, Zhenlin Xu, Guilherme Penedo, Loubna Ben Allal, Elie Bakouch, John David Pressman, Honglu Fan, Dashiell Stander, Guangyu Song, Aaron Gokaslan, Tom Goldstein, Brian R. Bartoldson, Bhavya Kailkhura, Tyler Murray
NeurIPS 2025,

[8] The Leaderboard Illusion
Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, **Shayne Longpre**, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, Sara Hooker
NeurIPS 2025, *[Featured in TechCrunch (+7)]*

[9] To Err Is AI: A Case Study Informing LLM Flaw Reporting Practices
Sean McGregor, Allyson Ettinger, Nick Judd, Paul Albee, Liwei Jiang, Kavel Rao, William H. Smith, **Shayne Longpre**, Avijit Ghosh, Christopher Fiorelli, Michelle Hoang, Sven Cattell, Nouha Dziri
AAAI 2025

[10] In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI
Shayne Longpre, Kevin Klyman, Ruth E. Appel, Sayash Kapoor, Rishi Bommasani, Michelle Sahar, Sean McGregor, Avijit Ghosh, Borhane Blili-Hamelin, Nathan Butters, Alondra Nelson, Amit Elazari, Andrew Sellars, Casey John Ellis, Dane Sherrets, Dawn Song, Harley Geiger, Ilona Cohen, Lauren McIlvenny, Madhulika Srikumar, Mark M. Jaycox, Markus Anderljung, Nadine Farid Johnson, Nicholas Carlini, Nicolas Miailhe, Nik Marda, Peter Henderson, Rebecca S. Portnoff, Rebecca Weiss, Victoria Westerhoff, Yacine Jernite, Rumman Chowdhury, Percy Liang, Arvind Narayanan
ICML 2025, **Spotlight**, *[Featured in Wired (+3)]*

[11] Global MMLU: Understanding and Addressing Cultural & Linguistic Biases in Multilingual Evaluation
Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, **Shayne Longpre**, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre FT Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, Sara Hooker
ACL 2025

[12] The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models
Seungone Kim, Juyoung Suk, Ji Yong Cho, **Shayne Longpre**, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, Minjoon Seo
NAACL 2025, **Best Paper Award**

[13] The foundation model transparency index v1. 1: May 2024
Rishi Bommasani, Kevin Klyman, Sayash Kapoor, **Shayne Longpre**, Betty Xiong, Nestor Maslej, Percy Liang
TMLR 2025

[14] Bridging the Data Provenance Gap Across Text, Speech and Video
Shayne Longpre, ... (50 authors), Caiming Xiong, Luis Villa, Stella Biderman, Alex Pentland, Sara Hooker, Jad Kabbara
ICLR 2025, *[Featured in MIT Tech Review (+2)]*

[15] The Responsible Foundation Model Development Cheatsheet: A Review of Tools & Resources
Shayne Longpre, Stella Biderman, Alon Albalak, Hailey Schoelkopf, Daniel McDuff, Sayash Kapoor, Kevin Klyman, Kyle Lo, Gabriel Ilharco, Nay San, Maribeth Rauh, Aviya Skowron, Bertie Vidgen, Laura Weidinger, Arvind Narayanan, Victor Sanh, David

Adelani, Percy Liang, Rishi Bommasani, Peter Henderson, Sasha Luccioni, Yacine Jernite, Luca Soldaini
TMLR 2025, **Outstanding Survey**

[16] Consent in Crisis: The Rapid Decline of the AI Data Commons
Shayne Longpre, ... (50 authors), Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, Jad Kabbara, Sandy Pentland
NeurIPS 2024, ***[Featured in The New York Times (+17)]***

[17] Foundation Model Transparency Reports
Rishi Bommasani, Kevin Klyman, **Shayne Longpre**, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, Percy Liang
AIES 2024, **Oral**

[18] A Systematic Review of NeurIPS Dataset Management Practices
Yiwei Wu, Leah Ajmani, **Shayne Longpre**, Hanlin Li
NeurIPS 2024

[19] A Safe Harbor for AI Evaluation and Red Teaming
Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Sandy Pentland, Arvind Narayanan, Percy Liang, Peter Henderson
ICML 2024, **Oral (1.5%)**, ***[Featured in The Washington Post (+5)]***

[20] On the Societal Impact of Open Foundation Models
Sayash Kapoor, Rishi Bommasani, Kevin Klyman, **Shayne Longpre**, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E Ho, Percy Liang, Arvind Narayanan
ICML 2024, **Oral (1.5%)**

[21] AI-Powered Autonomous Weapons Risk Geopolitical Instability and Threaten AI Research
Riley Simmons-Edler, Ryan Badman, **Shayne Longpre**, Kanaka Rajan
ICML 2024, **Oral (1.5%)**

[22] Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them?
Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Katy Gero, Sandy Pentland, Jad Kabbara
ICML 2024, **Spotlight**

[23] The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI
Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, Sara Hooker
Nature Machine Intelligence 2024, 🤗 **10k+ downloads**, ***[Featured in The Washington***

[24] A Survey on Data Selection for Language Models
Alon Albalak, Yanai Elazar, Sang Michael Xie, **Shayne Longpre**, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, William Yang Wang
TMLR 2024, **Featured Paper Award**

[25] Aya model: An instruction finetuned open-access multilingual language model
Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, **Shayne Longpre**, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, Sara Hooker
ACL 2024, **Best Paper Award**, 🤗 **100k+ downloads**

[26] The Foundation Model Transparency Index
Rishi Bommasani, Kevin Klyman, **Shayne Longpre**, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, Percy Liang
TMLR 2024, **Featured Paper Award**, ***[Featured in The New York Times (+8)]***

[27] Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models
Seungone Kim, Juyoung Suk, **Shayne Longpre**, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, Minjoon Seo
EMNLP 2024

[28] Prometheus: Inducing Fine-grained Evaluation Capability in Language Models
Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, **Shayne Longpre**, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, Minjoon Seo
ICLR 2024

[29] OctoPack: Instruction Tuning Code Large Language Models
Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, **Shayne Longpre**
ICLR 2024, **Spotlight**

[30] Mixture-of-Experts Meets Instruction Tuning: A Winning Combination for Large Language Models
Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, **Shayne Longpre**, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, Denny Zhou
ICLR 2024

[31] A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity
Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, Daphne Ippolito
NAACL 2024, **Outstanding Paper Award**

[32] Scaling instruction-finetuned language models
{Hyung Won Chung, Le Hou, **Shayne Longpre**}, ... Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, Jason Wei (35 authors)
JMLR 2024, 🤗 **4M+ downloads**

[33] The Flan Collection: Designing data and methods for effective instruction tuning
Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, Adam Roberts
ICML 2023, **1.5k+ stars.** 🤗 **100k+ downloads**

[34] The Bigscience Roots Corpus: A 1.6 tb composite multilingual dataset
Hugo Laurençon... **Shayne Longpre**... Margaret Mitchell, Sasha Luccioni, Yacine Jernite (52 authors)
NeurIPS 2022

[35] BLOOM: A 176B-Parameter Open-Access Multilingual Language Model
Teven Le Scao, ... **Shayne Longpre**, ... Matteo Manica (128 authors)

ArXiv, 2022.

[36] Combining Compressions for Multiplicative Size Scaling on Natural Language Tasks
Rajiv Movva, Jinhao Lei, **Shayne Longpre**, Ajay Gupta, Chris DuBois
COLING 2022

[37] You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings
Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, **Shayne Longpre**, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, Oskar Van Der Wal
ACL 2022 BigScience Workshop

[38] MIA 2022 Shared Task: Evaluating Cross-lingual Open-Retrieval Question Answering for 16 Diverse Languages
Akari Asai, **Shayne Longpre**, Jungo Kasai, Chia-Hsuan Lee, Rui Zhang, Junjie Hu, Ikuya Yamada, Jonathan H. Clark, Eunsol Choi
NAACL 2022 Multilingual Information Access Workshop

[39] Active Learning Over Multiple Domains in Natural Language Tasks
Shayne Longpre, Julia Reisler, Edward Greg Huang, Yi Lu, Andrew Frank, Nikhil Ramesh, Chris DuBois
NeurIPS 2022 Workshop on Distribution Shifts

[40] Entity-Based Knowledge Conflicts in Question Answering
{**Shayne Longpre**, Kartik Perisetla, Anthony Chen}, Nikhil Ramesh, Chris DuBois, Sameer Singh
EMNLP 2021

[41] MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering
Shayne Longpre, Yi Lu, Joachim Daiber
TACL 2021

[42] Evaluating Entity Disambiguation and the Role of Popularity in Retrieval-Based NLP
Anthony Chen, Pallavi Gudipati, **Shayne Longpre**, Xiao Ling, Sameer Singh
ACL 2021

[43] Evaluating Question Rewriting for Conversational Question Answering
Svitlana Vakulenko, **Shayne Longpre**, Zhucheng Tu, Raviteja Anantha
WSDM 2021

[44] Open-Domain Question Answering Goes Conversational via Question Rewriting
Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, **Shayne Longpre**
NAACL 2021

[45] Pivot Through English: Reliably Answering Multilingual Questions without Document Retrieval
Ivan Montero, **Shayne Longpre**, Ni Lao, Andrew Frank, Christopher DuBois
NAACL 2021 Multilingual Information Access Workshop

[46] On the Transferability of Minimal Prediction Preserving Inputs in Question Answering
Shayne Longpre, Yi Lu, Chris DuBois
NAACL 2021

[47] A Wrong Answer or a Wrong Question? An Intricate Relationship between Question Reformulation and Answer Selection in Conversational Question Answering
Svitlana Vakulenko, **Shayne Longpre**, Zhucheng Tu, Raviteja Anantha
EMNLP 2020 Search-Oriented Conversational AI Workshop **Best Paper Award**

Writings on Policy & Economics

[48] Op-Ed: Policymakers Overlook How Open Source AI Is Reshaping Global Power
Lucie-Aimee Kaffee, **Shayne Longpre**.
Tech Policy Press 2025

[49] Op-Ed: AI crawler wars threaten to make the web more closed for everyone
Shayne Longpre.
MIT Tech Review 2025

[50] International AI Safety Report
A team of authors, led by Yoshua Bengio. **Shayne Longpre**, as part of the core writing group.

[51] Considerations for governing open foundation models
Rishi Bommasani, Sayash Kapoor, Kevin Klyman, **Shayne Longpre**, Ashwin Ramaswami, Daniel Zhang, Marietje Schaake, Daniel E Ho, Arvind Narayanan, Percy Liang.
Science 2024
Stanford HAI, Foundation Model Issue Brief Series, 2023.

[52] UK Gov: International Scientific Report on the Safety of Advanced AI — Interim Report (2024)
A team of authors, led by Yoshua Bengio. **Shayne Longpre**, as part of the core writing group.

[53] An Open Letter: A Safe Harbor for Independent AI Evaluation
An **Open Letter signed by 400+** researchers, journalists, and civil society members. Effort led by **Shayne Longpre**, as part of [19], 2024. ***[Featured in The Washington Post (+***

[54] A Safe Harbor for AI Evaluation and Red Teaming
Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Arvind Narayanan, Percy Liang, Peter Henderson
Knight First Amendment Institute Blog at Columbia University 2024.

[55] Long Comment to the US Copyright Office, Ninth Triennial Proceeding, Class 4
Kevin Klyman, **Shayne Longpre**, Sayash Kapoor, Arvind Narayanan, Aleksandra Korolova, Peter Henderson
Comment to the US Copyright Office, 2024.

[56] Discit Ergo Est: Training Data Provenance and Fair Use
Robert Mahari, **Shayne Longpre**
Network Law Review, 2024

[57] Request for Comment on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights
Researchers from Stanford HAI, CRFM, RegLab, and other institutions
Stanford HAI, Comment to National Telecommunications and Information Administration, 2024.

[58] Lethal autonomous weapons systems & artificial intelligence: Trends, challenges, and policies
Shayne Longpre, Marcus Storm, Rishi Shah
MIT Science Policy Review, Volume III, 2022

[59] Invigorating Competition in Social Networking: An Interoperability Remedy
Cristian Santesteban, **Shayne Longpre**
Competition Policy International, 2021

[60] How Big Data Confers Market Power to Big Tech: Leveraging the Perspective of Data Science
Cristian Santesteban, **Shayne Longpre**
The Antitrust Bulletin, 2020

Awards & Fundraising

Academic Awards

MIT Technology Review **Innovators Under 35** 2026
ICML 2026 RLxF Workshop **Best Paper Award** [3] 2026
NAACL 2025 **Best Paper Award** [12] 2025
TMLR 2025 **Featured Survey Award** [15] 2025
ACL 2024 **Best Paper Award** [25] 2024
NAACL 2024 **Outstanding Paper Award** [31] 2024
TMLR 2024 **Featured Paper Award** [26] 2024
Federation of American Scientists AI Legislative Proposal Award 2024
Awarded to top "New Legislative Proposals To Deploy Artificial Intelligence Strategically"
ICLR Highlighted Reviewer (Top 3%) 2022
EMNLP 2020, *Search-Oriented Conversational AI Workshop* **Best Paper Award** 2020
EMNLP 2019, *MRQA Workshop Shared Task* **2nd place** 2019
TREC 2019, *Conversational Assistance Track (CAsT) Shared Task* **1st place ("A Team")** 2019

Fundraising

AI Safety Tactical Opportunities Fund, AI Coordinated Disclosure Grant Award, 2025 2025
Awarded $25,000 in funding to build an AI flaw/incident reporting platform.
Mozilla Data Futures Lab, Infrastructure Grant Award, 2023 2024
Awarded for the Data Provenance Initiative, accompanied by $25,000 research grant
MIT Generative AI Impact Award, 2023 2023
Awarded for the Data Provenance Initiative, accompanied by $70,000 research grant.

Teaching Experience

Summary: I've been an instructor for multiple MIT courses, including the research seminar MAS.S68 I co-lead and designed from scratch. Prior to this, I lead an AI4ALL computer vision summer course for women in STEM, and was a teaching assistant for two graduate deep learning course at Stanford.

Instructor, MIT xPro, MIT 2024-2025
Instructed seminars on foundations of language modeling, and AI evaluations, safety and security.
Instructor, Lincoln Laboratory: Large Language Models, MIT 2024
Instructed seminars on foundations of language modeling, AI auditing, and large data curation.
Instructor, MAS.S68 Generative AI: Evaluation and New Research Methods, MIT 2023
Instructed research seminar on large language models, and the landscape of socio-political concerns with their adoption.
Instructor, AI4ALL, Stanford University 2017
Instructed course on Computer Vision fundamentals to young women in STEM.
Teaching Assistant, Natural Language Processing w/ Deep Learning (CS224N), Stanford 2017
Teaching Assistant, Computer Vision with Deep Learning (CS231N), Stanford2017

Service, Leadership, & Organization

Summary: Lead organizer of a 50+ member research initiative. Co-organized 1 tutorial (AAAI), 3 AI conference workshops (NeurIPS, DEFCON, NAACL), and 3 independently hosted workshops. Area chair and reviewer across several AI conferences. Former ACL professional conduct committee trained volunteer.

Data Provenance Initiative 2023 - Present

Founded and lead volunteer research organization, mentoring ∼ 20 undergraduate researchers.

Tutorial Organizer, AAAI 2025 2025

AI Data Transparency: The Past, the Present, & Beyond, AAAI 2025

Lead Workshop Organizer Stanford / MIT / Princeton 2024

The Future of Third Party AI Evaluation, Remote through Stanford University

Workshop Co-Organizer & Bug Bounty Adjudicator DEFCON 2024, AI Village 2024

AI Village Generative Red Teaming (GRT) 2 Event, 500+ red teamers

Workshop Organizer, NeurIPS 2023 2023

Instruction Following & Finetuning Workshop (ITIF) NeurIPS 2023

Workshop Organizer Stanford / Princeton 2023

Workshop on Responsible and Open Foundation Models, Remote through Princeton University

Aya Initiative Model Training Co-Lead 2023

Cohere For AI Aya Initiative

MozFest Facilitator 2023

Bringing Light to Shadow Data

Workshop Co-Lead Organizer, Brown University 2022

Defining Transparency Workshop Brown University

(*hosted w/ the Algorithmic Transparency Institute and Brown's Information Futures Lab*)

Workshop Organizer, NAACL 2022 2022

Multilingual Information Access Workshop (MIA), NAACL 2022

Shared-task Organizer, NAACL 2022 2022

MIA 2022 Shared task, NAACL 2022

Academic Service

Area Chair (2026): NeurIPS
Reviewer (2026): ICML
Area Chair (2025): ARR
Reviewer (2025): ARR, ICML, COLM, NeurIPS, AAAI, FAAcT
Reviewer (2024): ARR, ICML, NeurIPS, AAAI, ICLR, COLM
Area Chair (2023): EMNLP, *Large Language Models and the Future of NLP Track*
ACL Professional Conduct Committee 2021 - 2023

Trained volunteer, responding to complaints from ACL's Anti-Harassment Policy.

Reviewer (2023): ARR, NeurIPS, FAccT, ICML, ICLR, EMNLP
Reviewer (2022): ARR, NeurIPS, ICLR, EMNLP, various workshops

Advising & Mentorship

Volunteer at MIT **Students Offering Support Program** 2022-2024

Advising underrepresented students in MIT graduate applications.

Hamidah Oderinwale, McGill University Undergrad. 2024 - 2025
Naana Obeng-Marnu, MIT Masters → Now UPenn PhD. Published [23] [14]. 2024 - 2025
Emily Chen, UNC Chapel Hill Undergrad. 2025
Elaine Zhu, Northeasern University Undergrad 2025
Carson Ezell, Harvard University Undergrad → RAND. 2025
Niklas Muennighoff, Peking University → Stanford CS PhD. Published [16][25][29]. 2023 - 2025
Campbell Lund, Wellesley CS → Edinburgh University AI Ethics MS. Published [16]. 2023

- 2025

Seungone Kim, KAIST AI MS → CMU CS PhD student. Published [28] [27] [14]. 2022 - 2024

An My Dinh, MIT Undergrad. Published [16]. 2023 - 2024

Minnie Liang, MIT Undergrad. Published [16]. 2023 - 2024

Rajiv Movva, MIT CS → Now Cornell Tech CS PhD. Published [36]. 2021 - 2022

Xuhui Zhou, UW MS → Now LTI-CMU CS PhD. 2021 - 2022

Erik Jones, Stanford MS → UC Berkeley CS PhD → Anthropic. 2021

Ivan Montero, UW CS MS student → Now Apple Machine Learning. Published [45]. 2020 - 2021

Invited Lectures & Panels

Summary: I've been fortunate to give 60+ guest lectures on my research at MIT, Harvard, Stanford, UC Berkeley, UW, UT Austin, USC, Brown; invited talks at ICLR 2025, Google Deepmind, AI2, Cohere, Databricks, Amazon, Mozilla, ML Commons, and various Alignment Workshops. I've also been hosted on BBC Radio 4, the StackOverflow podcast, and Yahoo! Finance.

General Talks & Panels

TAIGR @ ICML 2026, Panel Moderator, "AI Security, Governance, and Cooperation" 2026

Seoul Forum on AI Safety and Security Standards (SFASS), Building Safe AI for a Global Userbase 2026

Cohere Labs Connect 2025, Keynote Speaker, Research Collaborations 2025

OECD GPAI Expert Workshop, Expert Presentation on AI Incidents and Hazard Monitor (OECD.AI AIM) 2025

Stanford UN AI for Good Law Track, Panel on Benchmarking & Auditing AI Systems in High-Risk Domains 2026

Knight First Amendment Institute, AI and Democratic Freedoms, Panel Moderator 2025

US Patent and Trademark Office Talk, Listening Session on AI 2024

US Copyright Office DMCA Section 1201 Hearing, Invited Participant and Testimony 2024

Talks & Lectures for ATLAS [4]

Y Combinator Paper Club, ATLAS presentation (August 12) 2026

UC Berkeley CS 294/288, Guest Lecture (Host: Sewon Min) 2026

UC Berkeley Simons Institute, Agency in Collaborative Learning, Invited Talk 2026

USC ISI Natural Language Seminar (Host: Atharva Kulkarni) 2026

NYU CILVR Seminar Series (Host: Megan Richards) 2026

Amazon AGI Team (Host: Gamaleldin Elsayed) 2026

University of Washington ATLAS Reading Group (Host: Luke Zettlemoyer) 2026

International Workshop on Pretraining and Posttraining of Sovereign Foundation Models 2025

Invited Talks at ICLR 2025 (Singapore)

Building Trust in LLMs and LLM Applications on AI Coordinated Disclosure [10] 2025

Open Science for Foundation Models (SCI-FM) 2025 on Open Data [14][16] 2025

Future of Machine Learning Data Practices (MLDPR) 2025 on Data Provenance [14][16] 2025

Talks & Lectures for the Third-Party AI Disclosure [10].

University of Washington GenAI Ethics, Guest Lecture on Safe Harbor and Third-Party Evaluation (Host: Belen Saldias) 2025

Singapore Alignment Workshop, Invited Talk 2025
US DoD Technical Exchange on Frontier AI, Invited Talk (Host: Rumman Chowdhury) 2025
AAAI Session on AI Safety, Reliability, and Incident Management (Host: Sean McGregor) 2025
Cybersecurity at MIT Sloan Invited Talk (Host: Stuart Madnick) 2025
MIT EECS Lingo Lab Invited Talk (Host: Jacob Andreas) 2025
Microsoft New England Research & Development Center, Invited Talk 2025

Talks & Lectures for the Multimodal Data Provenance [14]
Brown University — Tech & Society Reading Group (Host: Suresh Venkatasubramanian) 2025
MIT Media Lab Rising Stars in AI — Keynote Talks (Host: MIT Media Lab) 2025
Twelve Labs — Multimodal Weekly (Host: James Le) 2024

Talks & Lectures for the Foundation Model Development Cheatsheet [15]
Linux Foundation AI & Data Seminar Series (Host: Anni Lai) 2024

Talks & Lectures for the Consent in Crisis [16]
BBC Radio 4 — Will AI Eat Itself? — Podcast Guest 2024
Risks of AI in the Military — Panel Moderator 2024
Women in AI & Robotics Reading Group (Host: Cleo Norris) 2024
UC Berkeley Responsible AI Workshop (Host: Genevieve Smith) 2024
MIT Sloan Initiative on the Digital Economy Guest Speaker (Host: Sinan Aral) 2024
AI Tinkerer Paper Club (Host: Human Feedback Foundation) 2024
Mozilla AI Salon (Host: Abeba Birhane) 2024
PLAMADISO – Platforms, Markets, and the Digital Society (Host: Volker Stocker) 2024
Creative Commons – Workshop on Preference Signals (Host: Anna Tumadóttir) 2024
MozFest Data Futures Lab Showcase (Host: Mozilla) 2024
Stanford HAI (Host: Digital Economy Lab) 2024
Sony AI (Host: Wiebke Hutiri) 2024
MIT CIO Symposium on AI (Host: Irving Wladawsky-Berger) 2024

Talks & Lectures for A Safe harbor for AI Evaluation & Red Teaming [19]
Harvard Kempner Center — Reading Group (Host: Ryan Badman) 2024
Alignment Workshop Lightning Talk Santa Cruz 2024

Talks & Lectures for the Data Provenance Initiative [23]
MIT Imagination in Action (Host: MIT Media Lab) 2024
USC NLG Seminar Series (Host: Justin Cho) 2024
UT Austin Data Ethics course, Guest Lecture (Host: Hanlin Li) 2024
MLCommons Croissant Group 2024
Mozilla Data Futures Lab Speaker Series 2024
Ethical Commerce Alliance (Host: Nina Müller) 2023
Harvard Library Innovation Lab (Host: Greg Leppert) 2023
MIT Algorithmic Alignment Group (Host: Dylan Hadfield-Menell) 2023

Talks & Lectures for the A Pretrainer's Guide [31]
Microsoft Research India (Host: Sanchit Ahuja) 2024
Salesforce AI (Host: Caiming Xiong) 2024
Cohere For AI (Host: Sara Hooker) 2024
Google Deepmind (Host: Daphne Ippolito) 2023

Mosaic ML (Host: Jonathan Frankle) 2023
Harvard & MIT: Policymaking for AI Series, Invited Talk & Panel (Host: Getting Plurality Research Network) 2023
University of Washington (Hosts: Akari Asai, Sewon Min) 2023
Allen Institute of AI (AI2) (Host: Maria Antoniak) 2023

Talks & Lectures for Effective Instruction Tuning [33][32]
Instituto Superior Técnico Seminar Series (Host: Nuno Guerreiro) 2023
Amazon Data-centric AI Seminar Series (Host: Li Lihong) 2023
Databricks Seminar Series (Host: Mike Conover) 2023
Apple Applied ML Reading Group (Host: Michael Tu) 2023
Kailua Labs AI Seminar Series (Host: Pablo Mendes) 2023
Oracle ML Seminar Series (Host: Ari Kobren) 2023
Google Research (Host: Denny Zhou) 2022

Other Talks & Lectures
Truth & Trust Online 2022 *Evaluating Transparency in Online Social Platforms* 2022
Panel moderator at NAACL 2022, *MIA Workshop* 2022
UC Irvine Reading Group *Knowledge Conflicts in QA* [40] (Host: Sameer Singh) 2021
Question Answering Evaluation Panel moderator at EMNLP 2021, *SCAI Workshop* 2021

Press & Media

Select Press for The AI Observatory [1] 2026
CBS News. *More people are using AI chatbots for health and relationship advice, study shows*
The Washington Post. *New data shows people are getting personal with AI*
MIT Technology Review. *We still don't know how people are really using AI*

Select Coverage on Research & Quoted Pieces 2024 - 2026
The New York Times. *We Are Just Beginning to Understand the Risks of Using A.I. at Work*
The Wall Street Journal. *Workers Sue $10 Billion AI Startup for Collecting and Exposing Personal Data*
The Washington Post. *Elon Musk's "truth-seeking" chatbot often disagrees with him*
MIT Technology Review. *Cloudflare will now, by default, block AI bots from crawling its clients' websites*
Bloomberg Law. *Websites Turn to Charging AI Scrapers They've Struggled to Block*
Nature (Feature). *The AI revolution is running out of data. What can researchers do?*
MIT Sloan. *Bringing transparency to the data used to train artificial intelligence*
IT Brew. *How an IT director deals with AI crawlers*

Select Press for Open Model Ecosystems and Economies of Open Intelligence 2025 & 2026
Financial Times. *China leapfrogs US in global market for 'open' AI models*
Financial Times. *Mistral unveils new models in race to gain edge in 'open' AI*
Financial Times. *Three questions AI needs to answer*
MIT Technology Review. *What's next for Chinese open-source AI*
AI World. *Chinese developers account for over 45% of top open-model public downloads*
Stanford HAI Policy Brief. *Beyond DeepSeek: China's Diverse Open-Weight AI Ecosystem and Its Policy Implications*
Stanford AI Index Report. *Featured Foundation Model Transparency Index; featured re-*

search on data provenance and consent; featured open-model ecosystem data

Select Press for The Leaderboard Illusion [8] 2025

TechCrunch. *Study accuses LM Arena of helping top AI labs game its benchmark*

Ars Technica. *Researchers claim LM Arena's AI leaderboard is biased against open models*

404 Media. *Researchers Say the Most Popular Tool for Grading AIs Unfairly Favors Meta, Google, OpenAI*

New Scientist. *Meta, Amazon and Google accused of 'distorting' key AI rankings*

BetaKit. *Cohere Labs head calls "unreliable" AI leaderboard rankings a "crisis" in the field*

Computerworld. *Leaderboard illusion: How big tech skewed AI rankings on Chatbot Arena*

The Rundown AI. *AI benchmarking under fire*

The Logic. *Tech firms are gaming the most popular ranking of AI models, researchers claim*

Select Press for FlexOlmo [5] 2025

Wired. *A New Kind of AI Model Lets Data Owners Take Control*

Select Press for The Common Pile [7] 2025

The Washington Post. *AI firms say they can't respect copyright. These researchers tried.*

Select Press for AI Flaw Reporting [10] [2] 2024 - 2026

WIRED. *You Can Now Sound the Alarm on AI Behaving Badly*

WIRED. *Researchers Propose a Better Way to Report Dangerous AI Flaws*

CNBC. *Researchers say AI chatbots need stronger standards and tests*

ZDNET. *OpenAI used to test its AI models for months—now it's days. Why that matters*

Stanford HAI. *A Framework to Report AI's Flaws*

Select Press for Multimodal Data Provenance [14] 2024 & 2025

The Washington Post. *OpenAI won't say whose content trained its video tool. We found some clues.*

MIT Technology Review. *This is where the data to build AI comes from*

The Globe and Mail. *AI-generated video has come a long way. Can you spot the difference between real and fake?*

Select Press for Consent in Crisis [16] 2024 & 2025

The New York Times. *The Data That Powers A.I. Is Disappearing Fast*

The Wall Street Journal. *The AI Scraping Fight That Could Change the Future of the Web*

Vox. *It's practically impossible to run a big AI company ethically*

Yahoo! Finance. *Data to train AI models is becoming restricted. Here's why.*

404 Media. *The Backlash Against AI Scraping Is Real and Measurable*

404 Media. *Anthropic AI Scraper Hits iFixit's Website a Million Times in a Day*

The StackOverflow Podcast. *The Stack Overflow Podcast: The Data Provenance Initiative*

MIT Technology Review. *AI that Feeds on a Diet of AI Garbage Ends up Spitting out Nonsense*

IEEE Spectrum. *AI Has Created a Battle Over Web Crawling*

Wired. *A New Group Is Trying to Make AI Data Licensing Ethical*

Le Monde. *A cause des intelligences artificielles, le Web se ferme de plus en plus*

Variety. *Generative AI & Licensing: A Special Report*

Nature Press. *The AI revolution is running out of data. What can researchers do?*

Riff Reporter. *Immer weniger aktuelle Daten für das Training: Künstliche Intelligenz in*

der Kris

The Observer. *AI Companies Are Running Out of Training Data: Study*

Futurism. *Crisis Looms as AI Companies Rapidly Losing Access to Training Data*

Start Magazine. *L'intelligenza artificiale è a corto di dati?*

Mozilla. *AI Training Can Undermine the Open Web. This Team Is Thinking Through Solutions*

Select Press for the Geopolitical risks of Autonomous Weaponry [21] 2024

Harvard Medical School News. *The Risks of Artificial Intelligence in Weapons Design*

Select Press for A Safe Harbor for AI Evaluation & Red Teaming [19] [53] 2024 & 2025

The Verge. *California is trying to regulate its AI giants—again*

The Washington Post. *Top AI researchers say OpenAI, Meta hinder independent evaluations*

VentureBeat. *Experts call for 'safe harbor' so researchers, journalists & artists can evaluate AI*

404 Media. *It May Soon Be Legal to Jailbreak AI to Expose How it Works*

Vox. *How would we even know if AI went rogue?*

Decipher. *The Emerging Ecosystem Dedicated to AI Accountability*

Select Press for the Aya Model [25] 2023

Cohere For AI. *The Journey of Aya - Accelerating Multilingual AI Through Open Science*

Select Press for Data Provenance Initiative [23] 2023

The Washington Post. *AI researchers uncover ethical, legal risks to using popular data sets*

VentureBeat. *MIT, Cohere for AI, others launch platform to track and filter audited AI datasets*

IEEE Spectrum. *Public AI Training Datasets Are Rife With Licensing Errors*

MIT News. *Study: Transparency is often lacking in datasets used to train large language models*

Reuters. *Legal transparency in AI finance: facing the accountability dilemma in digital decision-making*

TechCircle. *MIT, Cohere for AI, others launch platform to enhance transparency in AI data*

Cohere Blog. *Data Provenance Explorer Launches to Tackle Data Transparency Crisis*

Select Press for Foundation Model Transparency Index [26] 2023

Stanford HAI Blog. *Introducing The Foundation Model Transparency Index*

The New York Times. *Maybe we will finally learn more about how AI works*

The Atlantic. *We Don't Actually Know If AI Is Taking Over Everything*

Bloomberg. *Klobuchar Says AI Regulation Still Possible Before End of Year*

The Information. *How Transparent is your model?*

VentureBeat. *How transparent are AI models? Stanford researchers found out.*

The Verge. *The world's biggest AI models aren't very transparent, Stanford study says*

Reuters. *Stanford researchers issue AI transparency report, urge tech companies to reveal more*

Fast Company. *Why everyone seems to disagree on how to define Artificial General Intelligence*

Hacker News Our course on Generative AI trending 2023

Google AI Blog The Flan Collection: Advancing open source methods for instruction tuning 2023
PaLM 2 Technical Report Flan Collection & Methods cited several times as key components. 2023
DAIR.AI Top ML Papers of the Week 2023

References

Alex "Sandy" Pentland
Toshiba Professor of Media Arts & Science
Human-Centered AI (HAI) Fellow, Stanford University
Director, Human Dynamics, Media Lab, Massachusetts Institute of Technology
pentland@mit.edu

Percy Liang
Associate Professor of Computer Science (and courtesy in Statistics), Stanford University
Director of Center for Research on Foundation Models (CRFM)
pliang@cs.stanford.edu

Sara Hooker
(Former) Lead & VP of Research at Cohere
(Now) Founder, Adaptable Intelligence
sara@adaptionlabs.ai

Daphne Ippolito
Assistant Professor, School of Computer Science (SCS), Carnegie Mellon University (CMU)
Affiliate Cylab Security and Privacy Institute, SCS, CMU
Senior Research Scientist, Google Deepmind
daphnei@cmu.edu

Last Updated *September 2026.*