Saadia Gabriel

Email: skgabrie@cs.ucla.edu
Website: `https://saadiagabriel.com/`
PI: UCLA Misinformation, AI and Responsible Society Lab

Employment History

- 2024 – · · · · **University of California, Los Angeles**
  Samueli School of Engineering, Computer Science
  Assistant Professor
- 2023-2024 **New York University**
  Center for Data Science
  Faculty Fellow/Assistant Professor
- 2023 **Massachusetts Institute of Technology**
  Computer Science & Artificial Intelligence Laboratory
  Postdoctoral Fellow
- 2020, 2021 **Microsoft Research**
  AI/NLP Research Intern
- 2019-2020 **Allen Institute for Artificial Intelligence**
  Mosaic Team Research Intern
- 2019 **SRI International**
  Computer Vision & Learning Research Intern
- 2015, 2016 **University of Massachusetts, Amherst**
  Data Science Research Assistant

Education

- 2017 – 2023 **University of Washington**, PhD in Computer Science & Engineering.
  Advisors: Yejin Choi and Franziska Roesner
- 2013 – 2017 **Mount Holyoke College**, BA (*summa cum laude*) in Computer Science & Mathematics.
  Thesis Advisor: Daniel Sheldon

Awards & Honors

PI Fellowships & Awards

- Google Research Scholar, 2025
- Forbes 30 Under 30, Science 2024 List
- Outstanding Reviewer, ACL 2020 and NAACL 2022
- Google-Leap Fellowship, 2021
- Phi Beta Kappa, 2017
- Weaver Award for Computer Science and Math, Mount Holyoke College, 2017
- David Notkin Endowed Graduate Fellowship, University of Washington, 2017
- ARCS Foundation Fellowship, 2017

Paper Awards

- MIT Generative AI Impact Award, *Generative AI in the Era of "Alternative Facts"*, 2023
- Best Paper, *Social Bias Frames*, West Coast NLP Summit 2020

Awards & Honors (continued)

- Best Paper Nomination, *Early Fusion for Goal Directed Robotic Vision*, IROS 2019
- Best Paper Nomination, *The Risk of Racial Bias in Hate Speech Detection*, ACL 2019

Selected Recent Research Publications

Preprints and Working Papers

1. E. K. Guha, R. Marten, S. S. Keh, *et al.*, "OpenThoughts: Data recipes for reasoning models," 2025. URL: `https://api.semanticscholar.org/CorpusID:279154475`.
2. S. Gabriel, J. X. Han, E. Liu, *et al.*, "Advancing Equality: Harnessing Generative AI to Combat Systemic Racism," https://mit-genai.pubpub.org/pub/1ake7rfu, Mar. 2024.
3. S. Rahman, L. Y. Jiang, S. Gabriel, Y. Aphinyanagphongs, E. K. Oermann, and R. Chunara, "Generalization in Healthcare AI: Evaluation of a Clinical Large Language Model," 2024. URL: `https://api.semanticscholar.org/CorpusID:267750868`.

Journal Articles

1. L. Jiang, J. D. Hwang, C. Bhagavatula, *et al.*, "Investigating machine moral judgement through the Delphi experiment," *Nature Machine Intelligence*, 2025. URL: `https://arxiv.org/abs/2110.07574`.
2. X. Xu, B. Yao, Y. Dong, *et al.*, "Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 1, Mar. 2024. DOI: `10.1145/3643540`.

Conference and Workshop Proceedings

1. G. Liu, V. Le, S. Rahman, E. Kreiss, M. Ghassemi, and S. Gabriel, "MOSAIC: Modeling social AI for content dissemination and regulation in multi-agent simulations," in *EMNLP*, 2025. URL: `https://arxiv.org/abs/2504.07830`.
2. S. Rahman, S. M. Issaka, A. Suvarna, *et al.*, "AI Debate Aids Assessment of Controversial Claims," in *NeurIPS*, 2025. URL: `https://api.semanticscholar.org/CorpusID:279119478`.
3. S. Rahman, L. Jiang, J. Shiffer, *et al.*, "X-teaming: Multi-turn jailbreaks and defenses with adaptive multi-agents," in *COLM*, 2025. URL: `https://arxiv.org/abs/2504.13203`.
4. A. Suvarna, C. A. Chance, K. Naranjo, *et al.*, "ModelCitizens: Representing community voices in online safety," in *EMNLP*, 2025. DOI: `10.18653/v1/2025.emnlp-main.1571`.
5. A. Yerukola, S. Gabriel, N. Peng, and M. Sap, "Mind the gesture: Evaluating AI sensitivity to culturally offensive non-verbal gestures," in *ACL*, 2025. URL: `https://arxiv.org/abs/2502.17710`.
6. S. Gabriel, L. Lyu, J. Siderius, M. Ghassemi, J. Andreas, and A. Ozdaglar, "Generative AI in the Era of 'Alternative Facts'," in *Empirical Methods in Natural Language Processing (EMNLP)*, https://arxiv.org/abs/2410.09949, Nov. 2024.
7. S. Gabriel, I. Puri, X. Xu, M. Malgaroli, and M. Ghassemi, "Can AI Relate: Testing Large Language Model Response for Mental Health Support," in *Empirical Methods in Natural Language Processing (EMNLP) Findings*, Nov. 2024. URL: `https://api.semanticscholar.org/CorpusID:269921604`.

Patents

1. H. Palangi, S. Gabriel, T. Hartvigsen, D. Ray, M. Sap, and E. Kamar, *Adversarial language imitation with constrained exemplars*.

PhD Committees

UCLA Advisees

- Ashima Suvarna
  Sheriff Issaka
  Genglin Liu
  Salman Rahman
  Mehrab Beikzadeh
  Elizabeth Eyeson

UCLA CS

- Haoyi Qiu (Fall 2025)
  Wenhan Yang (Spring 2025)
  Christina Chance (Spring 2025)
  Yufei Tian (Spring 2025)
  Di Wu (Fall 2024)

UCLA ECE

- Natarajan Balaji Shankar (Fall 2024)

Teaching

Instructor

- Inside the Black Box: Artificial Intelligence Safety and Mechanistic Interpretability, 34 students. University of California Los Angeles, Fall 2025.
  Undergraduate Natural Language Processing (CS 162), 80 students. University of California Los Angeles, Spring 2025.
  Computational Ethics, Large Language Models and the Future of NLP (CS 269), 47 students. University of California Los Angeles, Winter 2025.
  Data Science Capstone (DS-GA 1006), 160 students. New York University, Fall 2023.

Guest Lecturer

- ML for Healthcare. UC Berkeley, Spring 2025.
  Speech Processing. University of California Los Angeles, Fall 2024.
  Ethical Machine Learning In Human Deployments. Massachusetts Institute of Technology, Spring 2024.
  Computational Ethics. Carnegie Mellon University, Spring 2023, 2024 and 2025.
  Natural Language Processing. Massachusetts Institute of Technology, Fall 2023 and 2024.
  AI Ethics. Oakton College, Fall 2023.
  Natural Language Processing (Undergrad). University of Washington, Spring 2023.
  Intro to Machine Learning. University of Washington, Fall 2022.
  Natural Language Processing. Mount Holyoke College, Fall 2019.
  NLP State-of-the-Art Methods. Carlson School of Management, Fall 2019.

Teaching Assistant

- Natural Language Processing (Undergrad/Grad). University of Washington.
  Real Analysis. Mount Holyoke College.

Selected Recent Talks & Panels

- Panelist at LA Tech Week 2025 (Tech St Santa Monica).
- Invited talks at COLM 2025 workshops on Social Simulations with LLMs (keynote) and Visions of Language Modeling.
- Invited talk at the Inaugural Conference of the International Association for Safe and Ethical AI, February 2025
- Invited talks at Stanford NLP Seminar, December 2021 and February 2025.
- Invited talk at UCLA Statistics and Data Science Seminar, March 2024.
- Invited talk at NYU Center for Data Science Lunch Seminar, November 2023.

Selected Recent Talks & Panels (continued)

- Invited talk at NYU-KAIST Inclusive AI Workshop, November 2023.
- Panelist for CHIL 2023 session on *LLMs and Healthcare*, June 2023.

Service

Area Chairing	ARR, COLM 2025, ICLR 2025, ICML 2025
Reviewing	ACL*, NAACL*, EMNLP, NeurIPS, AAAI, ICLR, ICML, Computational Linguistics, Journal of Artificial Intelligence, Journal of the American Medical Informatics Association, npj Digital Medicine, Nature Machine Intelligence (*Outstanding Reviewer)
Organizing Committee	Faculty Advisory Committee for the Bedari Kindness Institute (2024-2026) Faculty Advisory Committee for the Ralph J. Bunche Center for African American Studies (2024-2026) UCLA CS MS Admissions Committee (2024-2026) Session Chair for Computational Social Science and Cultural Analytics (EMNLP 2024) NYU Data Science MS Admissions Committee (2023-2024) Tutorial Co-Chair (NeurIPS 2023) Socio-Cultural Inclusion Co-Chair and Generation Session Chair (NAACL 2022) Safety for E2E Conversational AI Special Session (SIGDIAL 2021)
Peer Mentoring	GEM Computer Science Mentor at Mount Holyoke College (Spring 2016)