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 with Marzyeh Ghassemi

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

- MIT Technology Review Innovators Under 35 Semi-Finalist, 2026
- 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

- SAC Highlights Award, *African Languages Lab*, ACL 2026

Awards & Honors (continued)

- Editor's Pick, *Mental-LLM*, ACM IMWUT 2024
- Best Paper, *Social Bias Frames*, West Coast NLP Summit 2020
- 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

Research

Preprints and Working Papers

1. S. Afroogh *et al.*, "Beyond explainable AI (XAI): An overdue paradigm shift and post-XAI research directions," 2026. URL: https://arxiv.org/abs/2602.24176.
2. M. Beikzadeh, Y. Salmanpour, A. Suvarna, *et al.*, "Multi-objective alignment of language models for personalized psychotherapy," 2026. URL: https://api.semanticscholar.org/CorpusID:285726375.
3. C. Chance, R. Pattichis, A. Subramonian, *et al.*, "IYKYK (but AI doesn't): Automated content moderation does not capture communities' heterogeneous attitudes towards reclaimed language," 2026. arXiv: 2604.16654 [cs.CL]. URL: https://arxiv.org/abs/2604.16654.
4. S. M. Issaka, E. R. Gonzalez, L. Liu, *et al.*, "Translation as a scalable proxy for multilingual evaluation," 2026. URL: https://api.semanticscholar.org/CorpusID:284909942.
5. R. Lafuente-Mercado, S. Rahman, J.-Y. Shen, J. Shiffer, S. Gabriel, and P. Izmailov, "LeanPRM: Lean-grounded process supervision for natural language proof verification," 2026. URL: https://api.semanticscholar.org/CorpusID:292897802.
6. H.-Y. Qiu, G. Liu, P. N. Venkit, *et al.*, "Belief cascades drive persuasion in LLM agent networks," 2026. URL: https://api.semanticscholar.org/CorpusID:291367555.
7. S. Rahman, Y. H. Kim, M. Parmar, *et al.*, "Locating hidden failures makes long-horizon agents more reliable," 2026. URL: https://api.semanticscholar.org/CorpusID:292043289.
8. S. Rahman, J. Shen, A. Mordvina, H. Palangi, S. Gabriel, and P. Izmailov, "When can LLMs learn to reason with weak supervision?" 2026. URL: https://arxiv.org/abs/2604.18574.
9. N. Raoof, R. Zhuang, M. Nezhurina, *et al.*, "Openthoughts-Agent: Data recipes for agentic models," 2026. URL: https://api.semanticscholar.org/CorpusID:289625748.
10. 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. E. K. Guha, R. Marten, S. S. Keh, *et al.*, "OpenThoughts: Data recipes for reasoning models," in *ICLR*, Apr. 2026. URL: https://api.semanticscholar.org/CorpusID:279154475.

2 S. Issaka *et al.*, "The African Languages Lab: A collaborative approach to advancing low-resource African NLP," in *ACL*, Jul. 2026. URL: https://arxiv.org/abs/2410.09949.

3 G. Liu and S. Gabriel, "Pm-Bench: Evaluating prospective memory in LLM agents," in *COLM*, 2026. URL: https://api.semanticscholar.org/CorpusID:290148518.

4 R. Rajesh, C. A. Chance, A. Suvarna, K.-W. Chang, and S. Gabriel, "No convo! the illusion of reliability in online toxicity detection," in *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026*, Y. Zhang, R. Xu, A. Purwarianti, T. Rong, and J. Suzuki, Eds., Hengqin, China: Association for Computational Linguistics, Nov. 2026, pp. 4050–4072, ISBN: 979-8-89176-444-6. DOI: 10.18653/2026.findings-aacl.204.

5 A. Suvarna, K. Phan, M. Beikzadeh, H. Bansal, and S. Gabriel, "Supernova: Eliciting general reasoning in llms with reinforcement learning on natural instructions," in *EMNLP Findings*, 2026. URL: https://arxiv.org/abs/2604.08477.

6 A. Yerukola, F. Y. Harel-Canada, S. Khanuja, *et al.*, "NormViz: A benchmark and framework for grounding multimodal reasoning in global cultures," in *COLM*, 2026. URL: https://api.semanticscholar.org/CorpusID:291870700.

7 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.

8 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.

9 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.

10 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.

11 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.

12 R. Zhou, G. Wan, S. Gabriel, *et al.*, "Disparities in LLM reasoning accuracy and explanations: A case study on African American English," in *COLM MELT Workshop*, 2025. URL: https://arxiv.org/abs/2503.04099.

13 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.

14 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.

15 J. Lee, X. Lu, J. Hessel, *et al.*, "How to Train Your Fact Verifier: Knowledge Transfer with Multimodal Open Models," in *Empirical Methods in Natural Language Processing (EMNLP) Findings*, Nov. 2024. URL: https://api.semanticscholar.org/CorpusID:270870580.

16 S. Gabriel, S. Hallinan, M. Sap, *et al.*, "Misinfo Reaction Frames: Reasoning about readers' reactions to news headlines," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3108–3127. DOI: 10.18653/v1/2022.acl-long.222.

17 T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan,

P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3309–3326. DOI: 10.18653/v1/2022.acl-long.234.

18 S. Gabriel, A. Celikyilmaz, R. Jha, Y. Choi, and J. Gao, "GO FIGURE: A meta evaluation of factuality in summarization," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 478–487. DOI: 10.18653/v1/2021.findings-acl.42.

19 Z. Cheng, S. Gabriel, P. Bhambhani, *et al.*, "Detecting and tracking communal bird roosts in weather radar data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, Apr. 2020, pp. 378–385. DOI: 10.1609/aaai.v34i01.5373.

20 S. Gabriel, C. Bhagavatula, V. Shwartz, R. L. Bras, M. Forbes, and Y. Choi, "Paragraph-level commonsense transformers with recurrent memory," in *AAAI Conference on Artificial Intelligence*, 2020. URL: https://api.semanticscholar.org/CorpusID:222134165.

21 M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi, "Social Bias Frames: Reasoning about social and power implications of language," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 5477–5490. DOI: 10.18653/v1/2020.acl-main.486.

22 A. Amini, S. Gabriel, S. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi, "MathQA: Towards interpretable math word problem solving with operation-based formalisms," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2357–2367. DOI: 10.18653/v1/N19-1245.

23 M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1668–1678. DOI: 10.18653/v1/P19-1163.

24 A. Walsman, Y. Bisk, S. Gabriel, *et al.*, "Early fusion for goal directed robotic vision," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1025–1031. DOI: 10.1109/IROS40897.2019.8968165.

Patents

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

Grants

- UCLA Teaching and Learning Center Innovation Grant with Hong Zhou ($32,000), 2026-2027.
- Bedari Kindness Institute Research Innovation Fund with Elisa Borowski ($25,000), 2026-2027.
- Bedari Kindness Institute Enhanced Collaboration Grant Award with Cesi Cruz ($20,000), 2026-2027.
- Amazon Hub Research Awards ($100,000 cash, $50,000 credits), 2026-2027.
- Amazon Fall 2025 Research Awards ($250,000 in credits), 2026-2027.
- Hungarian Ministry of Culture and Innovation Grant with Brigitta Iványi-Bitter ($1,100,000), 2026-2029.

Grants (continued)

- Laude Institute Moonshots Seed Funding with Tom Griffiths, Rachit Dubey, Chris Summerfield, and Yulia Tsvetkov ($100,000), 2026.
- UCLA DataX Pilot Projects Grant with Cindy Nguyen ($20,724), 2026-2027.
- UCLA Initiative to Study Hate Seed Funding ($20,000), 2025-2026.
- Google Research Scholar Award ($60,000), 2025-2026.
- Google ML & Systems Junior Faculty Award ($100,000), 2025-2026.
- UCLA Initative to Study Hate Grant with Tom Hartvigsen and Sophie Hao ($20,000), 2024-2026.
- Samueli Foundation Gift Funding with Elisa Borowski ($25,000), 2024-2026.
- UCLA Racial & Social Justice Fund ($30,000), 2024-2026.
- UCLA ORCA Opportunity Fund ($60,000), 2024-2025.
- Google Gemma GCP Credit Award ($5,000 credits), 2024-2025.
- MIT Generative AI Impact Award with Marzyeh Ghassemi, Jacob Andreas, and Asu Ozdaglar ($60,000), 2023-2024.

PhD Committees

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

UCLA CS
- Elaine Yixin Wan (Winter 2026)
- Haoyi Qiu (Fall 2025)
- Wenhan Yang (Spring 2025)
- Christina Chance (Spring 2025)
- Yufei Tian (Spring 2025)
- Di Wu (Fall 2024)

UCLA Statistics
- Jonathan Vidal Contreras (Fall 2026)

UCLA ECE
- Natarajan Balaji Shankar (Fall 2024)

Teaching

Instructor
- Computational Ethics and Large Language Models (CS 260E), 69 students. UCLA, Fall 2026.
- Graduate Natural Language Processing (CS 263), 102 students. UCLA, Winter 2026.
- Inside the Black Box: Artificial Intelligence Safety and Mechanistic Interpretability, 34 students. UCLA, Fall 2025.
- Undergraduate Natural Language Processing (CS 162), 110 students. UCLA, Spring 2025-26.
- Computational Ethics, Large Language Models and the Future of NLP (CS 269), 47 students. UCLA, Winter 2025.
- Data Science Capstone (DS-GA 1006), 160 students. NYU, Fall 2023.

Teaching (continued)

Guest Lecturer	Black Studies in STEM. Caltech, Spring 2026. ML for Healthcare. UC Berkeley, Spring 2025. Speech Processing. UCLA, Fall 2024. Ethical Machine Learning In Human Deployments. MIT, Spring 2024. Computational Ethics. CMU, Spring 2023, 2024 and 2025. Natural Language Processing. MIT, 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 SF Tech Week 2026, October 2026.
- Invited talk and panelist at CHIL 2026 on *LLMs for Mental Health*, June 2026.
- Invited talk at UW NLP seminar, June 2026.
- Invited talk at UToronto SRI seminar, January 2026.
- Panelist at LA Tech Week 2025 (Tech St Santa Monica), October 2025.
- Invited talks at COLM 2025 workshops on Social Simulations with LLMs (keynote) and Visions of Language Modeling, October 2025.
- 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.
- 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, ICLR, ICML, NeurIPS
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, ACM Computing Surveys (*Outstanding Reviewer)

Service (continued)

Organizing Committee

- General Chair for the Southern California NLP Symposium (SoCalNLP 2026)
- UCLA CS Prospective PhD Student Visit Day Coordinator (2027)
- Faculty Advisory Committee for the Bedari Kindness Institute (2024-2026)
- Faculty Advisory Committee for the Ralph J. Bunche Center for African American Studies (2024-2027)
- UCLA NLP Seminar Co-organizer (2025-2027)
- UCLA CS MS Admissions Committee (2024-2027)
- 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)