

# Saadia Gabriel

Email: [skgabrie@cs.ucla.edu](mailto: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

-  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
-  Best Paper Nomination, *Early Fusion for Goal Directed Robotic Vision*, IROS 2019

## Awards & Honors (continued)

- 🏆 Best Paper Nomination, *The Risk of Racial Bias in Hate Speech Detection*, ACL 2019

## Research Publications

### Preprints and Working Papers









- 1 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.
- 2 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.*, “An Empirical Investigation of Machines’ Capabilities for Moral Judgment with the Delphi Experiment,” *Nature Machine Intelligence (Conditionally Accepted)*, 2024. 🔗 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 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.
- 2 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>.
- 3 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>.
- 4 K. Deng, A. Ray, R. Tan, S. Gabriel, B. A. Plummer, and K. Saenko, “Socratis: Are large multimodal models emotionally aware?” In *ICCV WECEIA*, 2023. 🔗 URL: <https://api.semanticscholar.org/CorpusID:261395214>.
- 5 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.
- 6 S. Gabriel, H. Palangi, and Y. Choi, “NaturalAdversaries: Can naturalistic adversaries be as effective as artificial adversaries?” In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5635–5645. 🔗 DOI: 10.18653/v1/2022.findings-emnlp.413.
- 7 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.

- 8 S. Gabriel, A. Bosselut, J. Da, *et al.*, “Discourse understanding and factual consistency in abstractive summarization,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds., Online: Association for Computational Linguistics, Apr. 2021, pp. 435–447.  DOI: 10.18653/v1/2021.eacl-main.34.
- 9 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.
- 10 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.
- 11 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>.
- 12 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.
- 13 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.
- 14 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.
- 15 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*.

## PhD Committees

- |          |   |                                      |
|----------|---|--------------------------------------|
| UCLA CS  |  | Di Wu, Fall 2024.                    |
| UCLA ECE |  | Natarajan Balaji Shankar, Fall 2024. |

## Teaching

- |            |   |   |
|------------|---|---|
| Instructor |  | Data Science Capstone (DS-GA 1006). New York University, Fall 2023. |
|------------|---|---|

## Teaching (continued)

Guest Lecturer	<ul style="list-style-type: none"><li>■ Ethical Machine Learning In Human Deployments. Massachusetts Institute of Technology, Spring 2024.</li><li>Computational Ethics. Carnegie Mellon University, Spring 2023 and 2024.</li><li>Natural Language Processing. Massachusetts Institute of Technology, Fall 2023.</li><li>AI Ethics. Oakton College, Fall 2023.</li><li>Natural Language Processing (Undergrad). University of Washington, Spring 2023.</li><li>Intro to Machine Learning. University of Washington, Fall 2022.</li><li>Natural Language Processing. Mount Holyoke College, Fall 2019.</li><li>NLP State-of-the-Art Methods. Carlson School of Management, Fall 2019.</li></ul>
Teaching Assistant	<ul style="list-style-type: none"><li>■ Natural Language Processing (Undergrad/Grad). University of Washington.</li><li>Real Analysis. Mount Holyoke College.</li></ul>

## Selected Talks & Panels



- Invited talk at UCLA Statistics and Data Science Seminar, March 2024.
- NeurIPS panelist on "Becoming a Successful AI Researcher/Engineer," December 2023.
- Invited talk at NYU Center for Data Science Lunch Seminar, November 2023.
- Invited talk at Northeastern, November 2023.
- Invited talk at NYU-KAIST Inclusive AI Workshop, November 2023.
- Invited talk at Mount Holyoke College Computer Science and Mathematics Lunch Seminar, October 2023.
- NYU Academic Careers panelist, Fall 2023.
- University of Washington Academic Careers panelist, Fall 2023.
- Panelist for CHIL 2023 session on *LLMs and Healthcare*, June 2023.
- Invited talk at Spotify NYC, June 2023.
- Invited talk at Cornell JEDI Dialogues Seminar, April 2022.
- Darpa SemaFor Workshop keynote, February 2022.
- Invited talk at Stanford NLP Seminar, December 2021.
- Presentation at MIT EECS Rising Stars Workshop, October 2021.
- Invited talk at UMass Amherst Rising Stars Colloquium, January 2021.
- Invited talk at NeurIPS 2020 Resistance AI Workshop, December 2020.
- Panelist for Voice Tech Global session on *Implicit Bias in Conversational AI*, July 2020.

## Service

Area Chairing	<ul style="list-style-type: none"><li>■ ARR (EMNLP 2024), ICLR 2025</li></ul>
Reviewing	<ul style="list-style-type: none"><li>■ ACL*, NAACL*, EMNLP, NeurIPS, AAAI, ICLR, ICML, Computational Linguistics, Journal of Artificial Intelligence, Journal of the American Medical Informatics Association, npj Digital Medicine (*Outstanding Reviewer)</li></ul>






## Service (continued)

---

- Organizing Committee     2024-2025 Faculty Advisory Committee for the Ralph J. Bunche Center for African American Studies  
2024-2025 UCLA CS MS Admissions Committee  
NeurIPS 2023, Tutorial Co-Chair  
NAACL 2022, Socio-Cultural Inclusion Co-Chair and Generation Session Chair  
SIGDIAL 2021, Safety for E2E Conversational AI Special Session
- Peer Mentoring     GEM Computer Science Mentor at Mount Holyoke College (Spring 2016)

## Misc

---

-  Author of "The Nightghosts' Child" novel.
-  Inventor of a solar charging jacket called "The Turtle."
-  University of Washington NLP Retreat Organizer (2018, 2019).
-  University of Washington CSE Visit Days Committee (2018).
-  Mount Holyoke College CS Department Chair Student Search Committee (2016-2017).