Sharon Levy

Assistant Professor, Rutgers University
Department of Computer Science

Email: s.levy@rutgers.edu
Web URL: https://sharonlevy.github.io/

Education

- **Ph.D., Computer Science - Natural Language Processing**
  University of California, Santa Barbara, 2018 - 2023
  Advisor: William Wang

- **M.S., Computer Science**
  University of California, Santa Barbara, 2017 - 2018
  Advisor: William Wang

- **B.S., Computer Science**
  University of California, Santa Barbara, 2013 - 2017
  College of Creative Studies

Research Interests

Natural language processing; Responsible AI; Text generation; Computational social science; Question-answering; Human-AI collaboration

Experience

- **Rutgers University**, New Brunswick, NJ 9/2024 – Present.
  Assistant Professor of Computer Science
- **Johns Hopkins University**, Baltimore, MD 8/2023 – 8/2024.
  Postdoctoral Fellow, Advisors: Mark Dredze and Michelle Kaufman
- **University of California, Santa Barbara**, CA 12/2017 – 6/2023.
  Graduate Student Researcher, Advisor: William Wang
- **Amazon Web Services (AWS) AI**, New York City, NY 06/2022 – 09/2022.
  Applied Scientist Intern (AWS Comprehend Team), Mentors: Neha Anna John and Ling Liu
- **Facebook**, 06/2021 – 10/2021.
  Facebook AI Applied Research Intern (AI Integrity Team), Mentor: Yi-Chia Wang
- **Pinterest**, 06/2020 – 08/2020.
  Pinterest Labs Research Intern (Ph.D., Machine Learning), Mentor: Jacob Gao

- **Akamai Technologies**, Santa Clara, CA 06/2017 – 09/2017.
  Security Engineer Intern, Mentor: Richard Lin
- **University of California, Santa Barbara**, CA 06/2014 – 09/2014.
  Web Developer Intern
- **KLA-Tencor**, Milpitas, CA 06/2012 – 12/2012.
  Software Developer Intern

Publications and Preprints

1. Iain Xie Weissburg, Sathvika Anand, **Sharon Levy**, Haewon Jeong. "LLMs are Biased Teachers: Evaluating LLM Bias in Personalized Education". In Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025)
2. **Sharon Levy**, Tahilin Sanchez Karver, William D. Adler, Michelle R. Kaufman, Mark Dredze. "Evaluating Biases in Context-Dependent Sexual and Reproductive Health Questions". In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)
3. **Sharon Levy**, William D. Adler, Tahilin Sanchez Karver, Mark Dredze, Michelle R. Kaufman. "Gender Bias in Decision-Making with Large Language Models: A Study of Relationship Conflicts". In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)
4. Michael Saxon*, Yiran Luo*, **Sharon Levy**, Chitta Baral, Yezhou Yang, William Yang Wang. "Lost in Translation? Translation Errors and Challenges for Fair Assessment of Text-to-Image Models on Multilingual Concepts". In Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)
5. **Sharon Levy**, Tahilin Sanchez Karver, William D. Adler, Michelle R. Kaufman, Mark Dredze. "Evaluating Biases in Context-Dependent Health Questions". (2024)
6. **Sharon Levy**, Neha Anna John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, Dan Roth. "Comparing Biases and the Impact of Multilingual Training across Multiple Languages". In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)
7. Alex Mei*, **Sharon Levy***, William Yang Wang. "ASSERT: Automated Safety Scenario Red Teaming for Evaluating the Robustness of Large Language Models". In Findings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)
8. Alex Mei*, **Sharon Levy***, William Yang Wang. "Foveate, Attribute, and Rationalize: Towards Safe and Trustworthy AI". Findings of the Association for Computational Linguistics (ACL 2023)
9. Matthew Ho*, Aditya Sharma*, Justin Chang*, Michael Saxon, **Sharon Levy**, Yujie Lu and William Yang Wang. "WikiWhy: Answering and Explaining Cause-and-Effect Questions", to appear in Proceedings of the International Conference on Learning Representations (ICLR 2023), Oral Paper: Top 5% out of all 4019 submissions.
10. Alon Albalak, **Sharon Levy**, William Yang Wang. "Addressing Issues of Cross-Linguality in Open-Retrieval Question Answering Systems For Emergent Domains". In Proceedings of the 2023 Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2023)

11. **Sharon Levy**, Emily Allaway, Melanie Subbiah, Lydia Chilton, Desmond Patton, Kathleen McKeown and William Yang Wang. "SafeText: A Benchmark for Exploring Physical Safety in Language Models", to appear in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), Long Paper, ACL.
12. Alex Mei*, Anisha Kabir*, **Sharon Levy**, Melanie Subbiah, Emily Allaway, John N. Judge, Desmond Patton, Bruce Bimber, Kathleen McKeown and William Yang Wang. "Mitigating Covertly Unsafe Text within Natural Language Systems", to appear in Findings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022).
13. Samhita Honnavalli*, Aesha Parekh*, Lily Ou*, Sophie Groenwold*, **Sharon Levy**, Vicente Ordonez and William Yang Wang. "Towards Understanding Gender-Seniority Compound Bias in Natural Language Generation", to appear in Proceedings of The 13th Language Resources and Evaluation Conference (LREC 2022).
14. Kai Nakamura, **Sharon Levy**, Yi-Lin Tuan, Wenhu Chen, William Yang Wang, "HybriDialogue: An Information-Seeking Dialogue Dataset Grounded on Tabular and Textual Data", to appear in Findings of 60th Annual Meeting of the Association for Computational Linguistics (Findings of ACL 2022), long paper, Dublin, Ireland.
15. **Sharon Levy**, Robert E. Kraut, Jane A. Yu, Kristen M. Altenburger, Yi-Chia Wang, "Understanding Conflicts in Online Conversations", to appear in Proceedings of the ACM Web Conference 2022 (WWW 2022), online, ACM.
16. Michael Saxon, **Sharon Levy**, Xinyi Wang, Alon Albalak and William Yang Wang, "Modeling Disclosive Transparency in NLP Application Descriptions", to appear in Proceedings of The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), long paper, online, ACL.
17. **Sharon Levy**, Kevin Mo, Wenhan Xiong and William Yang Wang, "Open-Domain Question-Answering for COVID-19 and Other Emergent Domains", to appear Proceedings of The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), demos track, online, ACL.
18. **Sharon Levy**, Michael Saxon and William Yang Wang, "Investigating Memorization of Conspiracy Theories in Text Generation", to appear in Findings of The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Findings of ACL-IJCNLP 2021), long paper, online, ACL.
19. Sophie Groenwold*, Lily Ou*, Aesha Parekh*, Samhita Honnavalli*, **Sharon Levy**, Diba Mirza and William Yang Wang, "Investigating African-American Vernacular English in Transformer-Based Text Generation", to appear in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), short paper, ACL.
20. **Sharon Levy**, Wenhan Xiong, Elizabeth Belding, and William Yang Wang, "SafeRoute: Learning to Navigate Streets Safely in an Urban Environment", to appear in ACM Transactions on Intelligent Systems and Technology (ACM TIST), journal paper, ACM, 2020.
21. Kai Nakamura*, **Sharon Levy***, and William Yang Wang, "Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection", to appear in Proceedings of 12th International Conference on Language Resources and Evaluation (LREC 2020), full paper, Marseille, France, May 11-16, 2020, ELRA.
22. **Sharon Levy** and William Yang Wang, "Cross-lingual Transfer Learning for COVID-19 Outbreak Alignment", Presented at the 1st Workshop on NLP for COVID-19 at ACL 2020.
23. *Sophie Groenwold, *Samhita Honnavalli, *Lily Ou, *Aesha Parekh, **Sharon Levy**, Diba Mirza, William Yang Wang, "Evaluating Transformer-Based Multilingual Text Classification", 2020.

Invited Talks

1. Rutgers University, An Ethical World: Knowledge, Power, and Justice, 2024
2. Montclair State University, Discovering Implicit Social Biases in Large Language Models, 2024
3. Johns Hopkins University, Panel on Artificial Intelligence - Gender, Data and Equity: Expert Conversations, 2024
4. Johns Hopkins University, *Foundational Concepts in Gender and Health Data and their Application*, 2024
5. Megagon Labs, 2024
6. Laguna Blanca School, 2023
7. UCSB CS 190i, *Introduction to Natural Language Processing*, 2023
8. Johns Hopkins University CLSP, 2023
9. Stanford University, 2023
10. UT Austin NLP Group, 2022
11. UCSB CS 165B, *Introduction to Machine Learning*, 2022
12. Fakespeak Workshop, University of Oslo, 2021
13. UCSB INT 200, *Seminar in Information Technology & Society*, 2021
14. SBCC Computer Science Club, 2021.

Grants, Awards and Honors

- Google Academic Research Award, 2024
- UCSB Computer Science Dissertation Award, 2023
- Fiona and Michael Goodchild Graduate Mentoring Award, 2023
- UCSB Computer Science Outstanding Mentoring Award, 2023
- EECS Rising Star, 2022
- Amazon Alexa AI Fellowship, 2020-2022
- CS Outstanding Teaching Assistant, 2020
- CRA-WP Grad Cohort for Women, 2019
- Regents Fellowship, 2018-2019
- Holbrook Fellowship, 2019
- Grace Hopper Poster Presentation, 2018
- Highest Honors (Top 2.5%), 2017

PhD Students

- Fatima Jahara (Rutgers University, 2024-Present)
- Adel Khorramrouz (Rutgers University, 2024-Present)

High School & Undergraduate Student Mentoring

- Kuleen Sasse (JHU BS, 2023-2024)
- Alex Mei (UCSB BS/MS, 2022-2023)
- Anisha Kabir (UCSB BS, 2022, CRA Outstanding Undergraduate Researcher Award Honorable Mention)
- John Judge (UCSB BS, 2022)
- Matthew Ho (UCSB BS, 2021-2022, Chancellor's Award for Undergraduate Research, CRA Outstanding Undergraduate Researcher Award Honorable Mention)
- Justin Chang (UCSB BS, 2021-2022)
- Aditya Sharma (UCSB BS, 2021-2022)
- Nga Ngo (UCSB BS, 2021-2022)
- Kevin Mo (Princeton BS, 2021)
- Aesha Parekh (UCSB BS, 2019-2021, Chancellor's Award in Undergraduate Research, CRA Outstanding Undergraduate Researcher Award Finalist)
- Lily Ou (UCSB BS, 2019-2021)
- Samhita Honnavalli (UCSB BS, 2019-2021, CRA Outstanding Undergraduate Researcher Award Honorable Mention)
- Sophie Groenwold (UCSB BS, 2019-2021)
- Kai Nakamura (High school/Caltech BS, 2019-2022)
- Ksenia Zhizhimontova (Cornell BS, 2019)

Teaching Experience

- Rutgers University, Spring 2025
  Instructor, *Natural Language Processing* 16:198:533
- Johns Hopkins University, Spring 2024
  Co-instructor, *Trustworthy and Responsible NLP*
- UC Santa Barbara, Fall 2019
  Teaching Assistant, *Introduction to Machine Learning*

Service

- Organizer: MASC-SLL (2024)
- Reviewer: ACL Rolling Review (2022-Present), FAccT (2024), MASC-SLL (2024), ACL (2023), SoCalNLP (2022)