Rebecca Dorn
(301) 832-2668 | rdorn@usc.edu | rebedorn.github.io

EDUCATION

Pursuing Ph.D., Computer Science Aug. 2021-Present
University of Southern California GPA 3.54
Advisors: Kristina Lerman and Fred Morstatter
Research Interests: Dialect Bias in Language Models, Machine Learning Fairness, Algorithm Auditing

B.S., Computer Science Sep. 2016-Jun. 2020
University of California, Santa Cruz GPA 3.47
Advisor: Lise Getoor
Thesis: Bias Exploration in Face Verification Systems

PUBLICATIONS

WORKING PAPERS

Chu, D., He, Z., **Dorn, R.** & Lerman, K. Improving and Assessing the Fidelity of Large Language Model Alignment to Online Communities. *Under Submission.*

Sanchez, C., Chu, D., He, Z., **Dorn, R.**, Murray, S. & Lerman, K. Feelings about Bodies: Emotions on Diet and Fitness Forums Reveal Gendered Stereotypes and Body Image Concerns. *Under Submission.*

PEER REVIEWED PAPERS

Ranjit, J., Joshi, B., **Dorn, R**., Petry, L., Koumoundouros, O., Bottarini, J., Liu, P., Rice, E. & Swayamdipta, S. OATH-Frames: Characterizing Online Attitudes Towards Homelessness via LLM. *ACM EMNLP 2024*. (20.8% Acceptance Rate)

He, Z., **Dorn, R.**, Guo, S., Chu, D. & Lerman, K. COMMUNITY-CROSS-INSTRUCT: Unsupervised Instruction Generation for Aligning Large Language Models to Online Communities. *ACM EMNLP 2024*. (20.8% Acceptance Rate)

Dorn, R., Kezar, L., Morstatter, F. & Lerman, K. Harmful Speech Detection by Language Models Exhibit Gender-Queer Dialect Bias. *ACM EAAMO 2024.* **(11.5% Acceptance Rate)**

Dorn, R., Mokhberian, N., Jiang, J. Abramson, J., Morstatter, F. & Lerman, K. Non-Binary Gender Expression in Online Interactions*. IEEE ASONAM 2024*.

INVITED TALKS

Dorn, R. (2024, November). Gender-Queer Dialect Bias in Harmful Speech Detection by Large Language Models. University of California Los Angeles, NLP Fairness Group. Los Angeles, CA.

Dorn, R. (2023, November). Gender-Queer Dialect Bias in Harmful Speech Detection by Large Language Models. University of Southern Information Science Institute Fairness and Bias Group. Marina Del Rey, CA.

Dorn, R. (2023, March). Non-Binary Gender Expression in Online Interactions. University of Southern California Center for AI in Society. Los Angeles, CA.

Dorn, R. (2022, October). Studying Gender Equity with Data Science. University of Southern California Women in Science and Engineering (WiSE) STEM Bytes. Los Angeles, CA.

POSTER PRESENTATIONS

Lee, E, Baird, A., Young, L. & **Dorn, R.** The Role of Public Facing Organizations in the Transgender Rights Debate: An Issue Management Framework. *Organizational Communication Research Escalator at ICA 2024*.

Ranjit, J., Joshi, B., **Dorn, R**., Petry, L., Koumoundouros, O., Bottarini, J., Liu, P., Rice, E. & Swayamdipta, S. OATH-Frames: Characterizing Online Attitudes Towards Homelessness via LLM. *USC ShowCAIS 2024*.
***Best Poster Award**

Dorn, R., Kezar, L., Morstatter, F. & Lerman, K. Harmful Speech Detection by Language Models Exhibit Gender-Queer Dialect Bias. *SoCalNLP 2023*.

Ranjit, J., Joshi, B., **Dorn, R**., Petry, L., Koumoundouros, O., Bottarini, J., Liu, P., Rice, E. & Swayamdipta, S. OATH-Frames: Characterizing Online Attitudes Towards Homelessness via LLM. *SoCalNLP 2023*.

Dorn, R., Mokhberian, N., Jiang, J. Abramson, J., Morstatter, F. & Lerman, K. Non-Binary Gender Expression in Online Interactions. *CMU SBP-BRiMS 2023*.

RESEARCH EXPERIENCE

Graduate Research Assistant, AAVE Dialect Bias in Text-Based Emotion Classification Sep. 2024-Present
University of Southern California, Information Science Institute

- Conducting a literature review surrounding the sociolinguistics of African American Vernacular English and emotion.
- Devising robust prompting schemas and running language and classification models to measure potential model bias.

Graduate Research Assistant, Unsupervised Alignment of LMs in Online Communities Jan. 2024-May 2024
University of Southern California, Information Science Institute

- Designed method for **self-supervised instruction tuning** using **LoRA** and **BERTopic** to align LLMs with online communities.
- Demonstrated alignment improvement through a 40% accuracy increase in comparison to base models for fitness groups.

Graduate Research Assistant, Characterizing Online Attitudes Towards Homelessness via LLM Jan. 2023-Feb. 2024
University of Southern California, Computer Science Department

- Scraped 5 million relevant tweets and tested 50 versions of **few-shot prompting** to probe GPT 4's use as an annotator.
- Collaborated with USC Social Work Department to formulate framework of online attitudes towards homelessness.

Graduate Research Assistant, Dialect Bias in LLM Classification Tasks Aug. 2023-Apr. 2024
University of Southern California, Information Science Institute

- Curated 108 natural templates to measure dialect bias by **GPT 3.5**, **Llama 2** and **Mistral** toxicity scores on reclaimed slurs.
- Directed team of 6 gender-queer annotators towards community-centric **chain-of-thought prompting** for bias mitigation.

Data Science Intern, Families USA Dec. 2020–Aug. 2021

- Integrated multiple Census datasets to examine demographics of uninsured populations receiving unemployment benefits.
- ***Utilized by Nancy Pelosi's office** to add provisions to the American Relief Act (2021).

TEACHING EXPERIENCE

Graduate Teaching Assistant, USC Computer Science Department Aug. 2023-Present

- Supported student learning by holding regular office hours, running lab sections and regularly meeting with project teams.
- Provided logistical support to four courses by grading assignments, uploading grades and monitoring online course forums.

Computer Science Education Coordinator, USC Stimulating STEM Summer Program May–August 2024

- Ran coding and empowerment lessons for high school students with identities historically underrepresented in STEM.
- Curated 10 Jupyter Notebooks to teach Python and guide students to create a project they are passionate about.

Educational Director, Habonim Dror Camp Moshava Mar.–Aug. 2020

- Led over 100 staff members in building social justice educational curriculum for campers entering grades 3 through 11.
- Adapted programming and staff training to meet evolving needs during the Coronavirus pandemic

Undergraduate Teaching Assistant, UC Santa Cruz Computer Science Department Jan.-Jun. 2020

- Designed assignments and rubrics for new Computer Science course, serving as sole undergraduate teaching assistant.
- Surveyed research in fair machine learning spanning differential privacy, bias mitigation and algorithm accountability.

Undergraduate Peer Tutor, UC Santa Cruz Learning Support Services Jan.-Dec. 2019

- Planned and led tutoring sessions for Artificial Intelligence, Intro to Algorithm Analysis, and Algorithms/Abstract Data Types
- Coordinated with Multicultural Education Program (MEP) to better support students historically underrepresented in STEM.

HONORS AND AWARDS

EAAMO Conference Travel Award, 2024
WiSE Travel Award, 2024
Best Poster Award for *OATH-Frames: Characterizing Online Attitudes Towards Homelessness…,* USC ShowCAIS 2024
SBP-BRiMS Scholarship, 2023
Inclusive STEM Educator Award, 2019

MEDIA COVERAGE

Cohen, Julia. “Flagged for Being Queer”. Article in *USC Viterbi School of Engineering Newsletter*. October 2024.

Soetirto, Rania. “AI Solutions for Social Good: Ph.D. Student Enlists LLM Assistants on Project Addressing Homelessness”. Article in *USC Viterbi School of Engineering Newsletter*. September 2024.

Russell, Adam (Host). “There is no end goal for AI ethics, there will always be something new to mitigate”. In *AI/nsiders Podcast*, June 2024.

INVOLVEMENT

Memberships: EAAMO (2024-Present)**,** QueerinAI (2023-Present), Women in Science and Engineering (2021-Present)
External Reviewer: CLARe6 (2024), LREC (2024), IJHCI (2024), CSCW (2023), WebSci (2022), JMIR (2022)
Organizer: Fairness and Bias ISI Seminar Series (2024-Present)

TECHNICAL SKILLS

Programming	Python (Expert), C++ (Intermediate), Bash (Basic), HTML (Basic)
Machine Learning	TensorFlow, PyTorch, Scikit-Learn, NumPy, Pandas, Matplotlib, Seaborn
Large Language Models	Fine Tuning (RAG, Instruction Tuning), Value Alignment, Prompt Engineering, Dataset Curation