

## EDUCATION

---

<b>PhD Computer Science; 4.0</b> <i>University of Washington, advised by Yejin Choi</i>	Dec 2025 Seattle, WA
<b>MS Computer Science; 4.0</b> <i>University of Washington, advised by Yejin Choi</i>	June 2024 Seattle, WA
◦ GRE, 169 / 170 Quantitative (96 <sup>th</sup> percentile), 165 / 170 Verbal (96 <sup>th</sup> percentile)	
<b>BS Applied and Computational Mathematics; 3.9</b> <i>Brigham Young University</i>	Apr 2021 Provo, UT
◦ ACT, 36 / 36 (99.8 <sup>th</sup> percentile)	

## EXPERIENCE

---

<b>humans&amp; - Member of Technical Staff</b>	Nov 2025 - Present
◦ Founding team member	
<b>Yejin Choi's Xlab, University of Washington - Research Assistant</b>	Sep 2022 - Dec 2025
◦ Researching LLMs, commonsense reasoning, morality+AI, model distillation, and NLP for social good	
◦ Spear-headed joint CS+Philosophy work on value pluralism	
<b>Google DeepMind - Student Researcher</b>	July 2024 - Dec 2024
◦ Modeling diverse perspectives and inferring values with language models (preprint , in review)	
◦ Mentored by Verena Rieser, Michiel Bakker, Roma Patel, MH Tessler, Iason Gabriel	
<b>Allen Institute for AI - PhD Research Intern</b>	June 2023 - Sep 2023
◦ Oral presentation at AAAI 2024. (paper)	
◦ Project exploring AI system's ability to model pluralistic human values, mentored by Chandra Bhagavatula	
<b>Perception, Control, and Cognition Laboratory - Research Assistant</b>	Apr 2020 - Sep 2022
◦ PNAS Paper exploring AI's potential to improve democratic discourse (co-first author, PNAS)	
◦ Demonstrated the effectiveness of mutual information as a prompt selection criterion on 8 datasets and 7 models (first author, ACL)	
◦ Engineered psychology-backed automatic rephrasing technique with GPT-3 to aid productivity of online conversations in collaboration with social scientists at Duke and BYU (in human trials)	
◦ Controlled difficult soft robot in real-time by combining first-principles physics and deep learning (Frontiers in Robotics and AI)	
◦ Contributed to open-source NLP data augmentation library (paper)	
<b>Enveda Biosciences - Data Science Intern</b>	Aug 2022 - Sep 2022
◦ Improved mass spectrometry to chemical structure machine translation model's validation performance by 5% using backtranslation (currently being worked into a paper and deployment)	
<b>Double River Investments - Machine Learning Engineer</b>	Jun 2020 - Aug 2021
◦ Informed live-traded quantitative investment model with transformer-based neural network, combining recent work by implementing and validating 5 research papers	
◦ Deployed production pipeline so model could be used in real time by multi-million dollar hedge fund	
<b>Gray Falkon - Deep Learning Consultant</b>	Dec 2019 - Apr 2020
◦ Sold NLP class project to company for \$10k by solving a crucial business need, saving thousands of monthly person-hours	
<b>Math Department, BYU - Competitive Coding Instructor</b>	Jan 2020 - Apr 2020

- Taught 18 students three times a week by developing coursework from scratch
- Helped place several students at top jobs and internships by refining their coding interview skills

## Computer Vision - Research Assistant

Feb 2019 - Dec 2019

- Awarded top prize in student research conference for work in pose correspondence
- Developed human-led, AI-assisted video annotation website to be used by MTurk workers

## SELECT PUBLICATIONS

---

- **Sorensen**, Jiang, Hwang, Levine, Pyatkin, West, Dziri, Lu, Rao, Bhagavatula, Sap, Tasioulas, Choi (2023) Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *AAAI 2024 - Oral presentation (top 3% of submissions)* <https://arxiv.org/abs/2309.00779>
- **Sorensen**, Moore, Fisher, Gordon, Mireshghallah, Rytting, Ye, Jiang, Lu, Dziri, Althoff, Choi (2024) A Roadmap to Pluralistic Alignment *ICML 2024 - Position Paper* <https://arxiv.org/abs/2402.05070>
- **Sorensen**, Mishra, Patel, Tessler, Bakker, Evans, Gabriel, Goodman, Rieser (2025) Value Profiles for Encoding Human Variation *arXiv preprint* <https://arxiv.org/abs/2402.05070>
- **Sorensen\***, Robinson\*, Rytting\*, Shaw, Rogers, Delorey, Khalil, Fulda, Wingate (2022) An Information Theoretic Approach to Prompt Engineering Without Ground Truth Labels. *Association for Computational Linguistics, 2022* <https://aclanthology.org/2022.acl-long.60/> \*Equal Contribution
- Argyle\*, Bail\*, Busby\*, Gubler\*, Howe\*, Rytting\*, **Sorensen\***, Wingate\* (2023) Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Published in PNAS*. <https://www.pnas.org/doi/10.1073/pnas.2311627120>. \*Equal Contribution, Alphabetical Order
- West, Le Bras, **Sorensen**, Lee, Jiang, Lu, Chandu, Hessel, Baheti, Bhagavatula, Choi (2023) NovaCOMET: Open Commonsense Foundation Models with Symbolic Knowledge Distillation *Findings of EMNLP 2023*. <https://aclanthology.org/2023.findings-emnlp.80/>
- Wingate, Shoeybi, **Sorensen** (2022) Prompt Compression and Contrastive Conditioning for Controllability and Toxicity Reduction in Language Models. *Findings of EMNLP 2022*. <https://aclanthology.org/2022.findings-emnlp.412/>
- Jung, West, Jiang, Brahman, Lu, Fisher, **Sorensen**, Choi (2023) Impossible Distillation: from Low-Quality Model to High-Quality Dataset Model for Summarization and Paraphrasing. *In review*. <https://arxiv.org/abs/2305.16635>
- Rytting, **Sorensen**, Argyle, Busby, Fulda, Gubler, Wingate (2023) Towards Coding Social Science Datasets with Language Models. *arXiv Preprint* <https://arxiv.org/abs/2306.02177>
- Dhole, Gangal, ..., **Sorensen** (2021) NL-Augmenter: A Framework for Task-Sensitive Natural Language Augmentation <https://arxiv.org/abs/2112.02721>
- Johnson, Quackenbush, **Sorensen**, Wingate, and Killpack (2021) Using First Principles for Deep Learning and Model-Based Control of Soft Robots. *Front. Robot. AI* 8:654398. doi: 10.3389/frobt.2021.654398 <https://www.frontiersin.org/articles/10.3389/frobt.2021.654398/full>

## INVITED TALKS

---

**National Symposium for Equitable AI**, Morgan State University - Pluralistic Alignment *Apr 2025*

**MilaNLP**, Pluralistic Alignment: A Roadmap, Recent Work, and Open Problems *Jan 2025*

**Dealing with Meaning Variation in NLP**, University of Utrecht - AI and Pluralistic Human Values *Oct 2024*

**Berkeley Simons Institute** - Pluralistic Alignment: A Roadmap, Recent Work, and Open Problems *Oct 2024* <https://www.youtube.com/live/1F0iaivYdvI?si=jKvSSqO1Eq74m5n>

**University College London** - Aligning AI with Pluralistic Human Values *Sep 2024*

**Vienna Alignment Workshop** - Lightning Talk, Pluralistic Alignment Workshop *July 2024*

**IBM Research** - AI and Pluralistic Human Values. *March 2024*

**BuzzRobot** - Aligning AI with Pluralistic Human Values. *May 2024* <https://www.youtube.com/watch?v=lEoBNBfNklI>

**AAAI Oral** - Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties *Feb 2024*

## GRANTS AND AWARDS

**Institute for Humane Studies** Awarded Funding for ICML 2024 Conference Presentation. *June 2024*

## SERVICE

**Reviewer** at ICML '25, FAaCT '25, NeurIPS'24, EMNLP'24, AAAI'24, EMNLP'23, NeurIPS'23, EMNLP'22

**Program Committee Member** - NeurIPS 2023 MP2: AI meets moral philosophy and moral psychology