Liwei Jiang

Paul G. Allen School of Computer Science & Engineering
University of Washington
Box 352350
Seattle, WA 98195

+1 206-852-2310
lwjiang@cs.washington.edu
https://liweijiang.me
@liweijianglw

RESEARCH INTERESTS

My research centers on **humanistic AI safety**, aiming to foster the synergistic, secure, and sustainable coexistence of AI and society—ultimately shaping the **co-evolution of AI and humanity**.

My **current** research encompasses topics such as pluralistic alignment, self-improving algorithms for steerable and secure language models, anticipatory strategies for long-term risks—including overreliance and the erosion of human creativity—and socially impactful applications, such as clinical AI.

My **primary technical focus** is natural language processing (NLP), complemented by interdisciplinary work in philosophy, cognitive science, and social science. I've recently expanded into reinforcement learning (RL) and multi-agent systems, and aim to further broaden into multi-modality and embodiment to better bridge cross-domain knowledge for addressing complex sociotechnical challenges.

EDUCATION

University of Washington — **Seattle, WA**

09/2019–current — Ph.D. in Computer Science & Engineering
Advisor: *Yejin Choi*
Committee: *Yulia Tsvetkov, Maarten Sap, Oren Etzioni, Lucy Lu Wang*

Colby College *(top 0.5%, Dean's List 15–19)* — **Waterville, ME**

09/2015–01/2019 — B.A. in Computer Science, *summa cum laude*, GPA 4.08

B.A. in Mathematics, *summa cum laude*, GPA 4.13
Advisor: *Bruce Maxwell*

PROFESSIONAL EXPERIENCES

Stanford University, Computer Science — **Palo Alto, CA**

02/2025–current — Visiting Researcher, with *Yejin Choi*

NVIDIA — **Santa Clara, CA**

03/2025–current — Research intern at the NeMo Guardrails team, with *Erick Galinkin* and *Christopher Parisien*

Allen Institute for Artificial Intelligence (Ai2) — **Seattle, WA**

12/2024–present — Research collaborator at the AllenNLP team

08/2024–12/2024 Student researcher at the AllenNLP team

06/2020–08/2024 Student researcher at the Mosaic team, with *Yejin Choi*

University of Washington, Computer Science & Engineering

04/2020–current Research Assistant, with *Yejin Choi*
AI safety [P.16, P.17, P.5, P.18, P.6, P.14, P.3, P.2, P.13, P.15], computational morals and norms [P.8, P.28, P.35, P.36, P.11, P.30], alignment of pluralistic human values [P.21, P.23, P.9] and cultures [P.19, P.10], data and algorithm innovations for socially beneficial applications, e.g., dialog systems [P.39, P.31], hallucinations [P.7], privacy-preserving authorship obfuscation [P.25], toxicity reduction [P.33, P.40, P.37, P.34], and the interplay of human and machine capabilities [P.20, P.27, P.12, P.32]

09/2019–08/2020 Research Assistant, with *James Fogarty*
Personalized health self-tracking tools for migraine patients [P.26]

Stanford University, Computer Science **Palo Alto, CA**

06/2017–09/2019 Research Intern, with *James Landay*
Educational interactive conversational systems, including QuizBot [P.43], EnglishBot [P.42], and BookBuddy [W.1]

PUBLICATIONS

*, †denote equal contribution

Manuscripts and Pre-prints

P.1 **Liwei Jiang**, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Maarten Sap, Yulia Tsvetkov, Nouha Dziri, Alon Albalak, and Yejin Choi. Artificial Hiveminds: The Open-Ended Homogeneity of Language Models (and Beyond).
In submission

P.2 **Liwei Jiang***, Mickel Liu*, Yancheng Liang, Yejin Choi, Simon Shaolei Du, Tim Althoff, and Natasha Jaques. Chasing Moving Targets with Online Self-Play Reinforcement Learning for Safer Language Models.
In submission

P.3 **Liwei Jiang***, Salman Rahman*, James Shiffer*, Genglin Liu, Sheriff Issaka, Md Rizwan Parvez, Hamid Palangi, Kai-Wei Chang, Yejin Choi, and Saadia Gabriel. X-Teaming: Multi-Turn Jailbreaks and Defenses with Adaptive Multi-Agents.
In submission to COLM 2025

P.4

P.5 Xuhui Zhou, Hyunwoo Kim*, Faeze Brahman*, **Liwei Jiang**, Hao Zhu, Ximing Lu, Frank Xu, Bill Yuchen Lin, Yejin Choi, Niloofar Mireshghallah, Ronan Le Bras, and Maarten Sap. HAICosystem: An Ecosystem for Sandboxing Safety Risks in Human-AI Interactions.
In submission to COLM 2025

P.6 Priyanshu Kumar*, Devansh Jain*, Akhila Yerukola, **Liwei Jiang**, Himanshu Beniwal, Tom Hartvigsen, and Maarten Sap. PolyGuard: A Multilingual Safety Moderation Tool for 17 Languages.
In submission to COLM 2025

P.7 Wenting Zhao, Tanya Goyal, Yu Ying Chiu, **Liwei Jiang**, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, and Yejin Choi. WildHallucination: Evaluating Long-Form Factuality in LLMs with Real-World Entity Queries. Preprint

Peer-reviewed Conference and Journal Publications

2025

P.8 **Liwei Jiang**, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Sydney Levine, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Taylor Sorensen, Jon Borchardt, Jack Hessel, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. Investigating Machine Moral Judgment through the Delphi Experiment.
Nature Machine Intelligence

P.9 **Liwei Jiang**, Taylor Sorensen, Sydney Levine, and Yejin Choi. Can Language Models Reason about Individualistic Human Values and Preferences?.
ACL 2025

P.10 Yu Ying Chiu, **Liwei Jiang**, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. CulturalBench: A Robust, Diverse and Challenging Benchmark on Measuring the (Lack of) Cultural Knowledge of LLMs.
ACL 2025

P.11 Yu Ying Chiu, **Liwei Jiang**, and Yejin Choi. DailyDilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life.
ICLR 2025 (Spotlight)

P.12 Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, **Liwei Jiang**, Khyathi Chandu, Nouha Dziri, and Yejin Choi. AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text.
ICLR 2025 (Oral)

P.13 Jillian Fisher, Ruth Elisabeth Appel, Chan Young Park, Yujin Potter, **Liwei Jiang**, Taylor Sorensen, Shangbin Feng, Yulia Tsvetkov, Margaret Roberts, Jennifer Pan, Dawn Song, and Yejin Choi. Position Paper: Political Neutrality in AI is Impossible—But Here is How to Approximate it.
ICML 2025 (Oral)

P.14 Jing-Jing Li, Valentina Pyatkin, Max Kleiman-Weiner, **Liwei Jiang**, Nouha Dziri, Anne G. E. Collins, Jana Schaich Borg, Maarten Sap, Yejin Choi, and Sydney Levine. SafetyAnalyst: Interpretable, Transparent, and Steerable Safety Moderation for AI Behavior.
ICML 2025

P.15 Sean McGregor, Allyson Ettinger, Nick Judd, Paul Albee, **Liwei Jiang**, Kavel Rao, Will Smith, Shayne Longpre, Avijit Ghosh, Christopher Fiorelli, Michelle Hoang, Sven Cattell, and Nouha Dziri. To Err is AI: A Case Study Informing LLM Flaw Reporting Practices.
IAAI 2025

2024

P.16 **Liwei Jiang**, Kavel Rao*, Seungju Han*, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Nouha Dziri, and Yejin Choi. WildTeaming at Scale: From In-the-Wild Jailbreak Tactics to (Adversarially) Safer Models.
NeurIPS 2024

P.17 Seungju Han*, Kavel Rao*, **Liwei Jiang**†, Allyson Ettinger†, Bill Yuchen Lin, Nathan Lambert, Nouha Dziri, and Yejin Choi. WildGuard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs.
NeurIPS D&B 2024

P.18 Jimin Mun, **Liwei Jiang**, Jenny Liang, Inyoung Cheong, Nicole DeCario, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. Particip-AI: A Democratic Surveying Framework for Anticipating Future AI Use Cases, Harms and Benefits.
AIES 2024

P.19 Huihan Li, **Liwei Jiang**, Jena D. Huang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. Culture-Gen: Revealing Global Cultural Perception in Language Models through Natural Language Prompting.
COLM 2024

P.20 Linlu Qiu, **Liwei Jiang**, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Models with Hypothesis Refinement.
ICLR 2024 (Oral)

P.21 Taylor Sorensen, **Liwei Jiang**, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties.
AAAI 2024

P.22 Jaehun Jung, Ximing Lu, **Liwei Jiang**, Faeze Brahman, Peter West, Pang Wei Koh, and Yejin Choi. Information-Theoretic Distillation for Reference-less Summarization.
COLM 2024

P.23 Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, **Liwei Jiang**, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position Paper: A Roadmap to Pluralistic Alignment.
ICML 2024

P.24 Jaehun Jung, Peter West, **Liwei Jiang**, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. Impossible Distillation: from Low-Quality Model to High-Quality Dataset & Model for Summarization and Paraphrasing.
NAACL 2024

P.25 Jillian Fisher, Ximing Lu, Jaehun Jung, **Liwei Jiang**, Zaid Harchaoui, and Yejin Choi. JAMDEC: Unsupervised Authorship Obfuscation Using Constrained Decoding Over Small Language Models.
NAACL 2024

P.26 Yasaman S Sefidgar, Carla L Castillo, Shaan Chopra, **Liwei Jiang**, Tae Jones, Anant Mittal, Hyeyoung Ryu, Jessica Schroeder, Allison Cole, Natalia Murinova, Sean A Munson, and James Fogarty. MigraineTracker: Examining Patient Experiences with Goal-Directed Self-Tracking

for a Chronic Health Condition.
CHI 2024 🏆 **Outstanding Paper Award**

P.27 Peter West*, Ximing Lu*, Nouha Dziri*, Faeze Brahman*, Linjie Li*, Jena D. Hwang, **Liwei Jiang**, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. The Generative AI Paradox: "What It Can Create, It *May* Not Understand".
ICLR 2024

2023 P.28 **Liwei Jiang***, Kavel Rao*, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. What Makes it Ok to Set a Fire? Iterative Self-distillation of Contexts and Rationales for Disambiguating Defeasible Social and Moral Situations.
Findings of EMNLP 2023

P.29 Peter West, Ronan Le Bras, Taylor Sorensen, Bill Yuchen Lin, **Liwei Jiang**, Ximing Lu, Khyathi Chandu, Jack Hessel, Ashutosh Baheti, Chandra Bhagavatula, and Yejin Choi. NovaCOMET: Open Commonsense Foundation Models with Symbolic Knowledge Distillation.
Findings of EMNLP 2023

P.30 Seungju Han, Junhyeok Kim, Jack Hessel, **Liwei Jiang**, Jiwan Chung, Yejin Son, Yejin Choi, and Youngjae Yu. Reading Books is Great, But Not if You Are Driving! Visually Grounded Reasoning about Defeasible Commonsense Norms.
EMNLP 2023 (Oral)

P.31 Hyunwoo Kim, Jack Hessel, **Liwei Jiang**, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization.
EMNLP 2023 🏆 **Outstanding Paper Award**

P.32 Nouha Dziri*, Ximing Lu*, Melanie Sclar*, **Liwei Jiang**†, Xiang Lorraine Li†, Bill Yuchen Lin†, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and Fate: Limits of Transformers on Compositionality.
NeurIPS 2023 (Spotlight)

P.33 Ximing Lu, Faeze Brahman, Peter West, Jaehun Jang, Khyathi Chandu, Abhilasha Ravichander, Lianhui Qin, Prithviraj Ammanabrolu, **Liwei Jiang**, Sahana Ramnath, Nouha Dziri, Jillian Fisher, Bill Yuchen Lin, Skyler Hallinan, Xiang Ren, Sean Welleck, and Yejin Choi. Inference-Time Policy Adapters (IPA): Tailoring Extreme-Scale LMs without Fine-tuning.
EMNLP 2023

P.34 Yiming Zhang, Sravani Nanduri, **Liwei Jiang**, Tongshuang Wu, and Maarten Sap. BiasX: "Thinking Slow" in Toxic Content Moderation with Explanations of Implied Social Biases.
EMNLP 2023

P.35 Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, **Liwei Jiang**, Yejin Choi, and Chandra Bhagavatula. ClarifyDelphi: Reinforced Clarification Questions with Defeasibility Rewards for Social and Moral Situations.
ACL 2023 (Oral)

2022 P.36 Prithviraj Ammanabrolu, **Liwei Jiang**, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. Aligning to Social Norms and Values in Interactive Narratives.
NAACL 2022

P.37 Ximing Lu, Sean Welleck*, Peter West*, **Liwei Jiang**†, Jungo Kasai†, Daniel Khashabi†, Ronan Le Bras†, Lianhui Qin†, Youngjae Yu†, Rowan Zellers†, Noah A. Smith, and Yejin Choi. NeuroLogic A*esque Decoding: Constrained Text Generation with Lookahead Heuristics.
NAACL 2022 🏆 **Best Paper Award**

P.38 Peter West, Chandra Bhagavatula*, Jack Hessel*, Jena D. Hwang*, **Liwei Jiang***, Ronan Le Bras*, Ximing Lu*, Sean Welleck*, and Yejin Choi. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models.
NAACL 2022

P.39 Hyunwoo Kim*, Youngjae Yu*, **Liwei Jiang**, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. ProsocialDialog: A Prosocial Backbone for Conversational Agents.
EMNLP 2022

P.40 Ximing Lu, Sean Welleck, **Liwei Jiang**, Jack Hessel, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. Quark: Controllable Text Generation with Reinforced Unlearning.
NeurIPS 2022

2021 P.41 **Liwei Jiang**, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. "I'm Not Mad": Commonsense Implications of Negation and Contradiction.
NAACL 2021

P.42 **Liwei Jiang***, Sherry Ruan*, Qianyao Xu*, Zhiyuan Liu, Glenn M. Davis, Emma Brunskill, and James A. Landay. EnglishBot: An AI-Powered Conversational System for Second Language Learning.
IUI 2021

2019 P.43 Sherry Ruan, **Liwei Jiang**, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. QuizBot: A Dialogue-Based Adaptive Learning System for Factual Knowledge.
CHI 2019

Posters, Extended Abstracts, Workshop Papers and Technical Reports

2019 W.1 Sherry Ruan, Angelica Willis, Qianyao Xu, Glenn M. Davis, **Liwei Jiang**, Emma Brunskill, and James A. Landay. BookBuddy: Turning Digital Materials Into Interactive Foreign Language Lessons Through a Voice Chatbot.
L@S WIP 2019

HONORS AND AWARDS

2025 Abstract Selection for 2025 Qualcomm Innovation Fellowship (North America)
Research proposal has been selected to advance to the Proposal phase.

2024 Outstanding Paper Award
CHI 2024

2023 Outstanding Paper Award
EMNLP 2023

2022 Best Paper Award
NAACL 2022

2019–2020 Anne Dinning - Michael Wolf Endowed Regental Fellowship
University of Washington, Paul G. Allen School First-Year Ph.D. Fellowship

2018 Member of the Phi Beta Kappa Society
Colby College, elected as a member of Phi Beta Kappa with junior standing

2016, 2017, 2018 Julius Seelye Bixler Scholar
Colby College, top-ranking students as determined by the academic record, three-time recipient

2018 Honorable Mention of Interdisciplinary Contest in Modeling (ICM)
20th annual Interdisciplinary Contest in Modeling (ICM)

2017 Phi Beta Kappa Undergraduate Scholastic Achievement Award
Colby College, top two students in the sophomore and junior classes

2016 Phi Beta Kappa Summer Research Scholar
Colby College, summer research stipend

TALKS

2025 **Can Language Models Reason about Individualistic Human Values and Preferences?**
Speaker, Darpa ITM PI Meeting (2025.3)
How to Build Machines with Deep Concerns of Human Traits, Values, and Needs?—Towards Humanistic AI Alignment
Speaker, University of Washington, Foster School of Business, Computational Minds and Machines lab, hosted by Max Kleiman-Weiner (2025.2)

2024 **AI Safety Panel**
Panelist, Annual Research Showcase and Open House Event, UW CSE (2024.10)
WildTeaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer LMs
Co-speaker, All-Ai2 Meeting, Allen Institute for Artificial Intelligence (Ai2) (2024.7)

2023 **On the Outcomes of Scientific Disagreements on Machine Morality**
Co-speaker, The Big Picture Workshop @ EMNLP, Singapore (2023.12)
Toward Interpretable and Interactive Socially & Ethically Informed AI
Speaker, Darpa ITM Kickoff PI Meeting (2023.5)
Toward Interpretable, Interactive, Informative Machine Moral Reasoning
Discussant, Mosaic Morality & AI Series, Allen Institute for Artificial Intelligence (Ai2) (2023.2)

2022 **Toward Socially Aware & Ethically Informed AI**
Speaker, UW NLP Retreat (2022.9)

2021 **Delphi: Toward Machine Ethics and Norms**
Speaker, All-Ai2 Meeting, Allen Institute for Artificial Intelligence (Ai2) (2021.10)

TEACHING EXPERIENCES

Guest Lecturer

05/2025 **Red-Teaming and Safeguarding Language Models: Current Practices, Challenges, and Future Directions**
In COM SCI 162: Natural Language Processing, UCLA (Instructor: Saadia Gabriel)

02/2025 **Red-Teaming and Safeguarding Language Models: Current Practices, Challenges, and Future Directions**
In 11-830: Ethics, Social Biases, and Positive Impact in Language Technologies, Carnegie Mellon University, w/ Nouha Dziri (Instructor: Maarten Sap)

11/2024 **How to Build AI with Deep Concerns for Human Traits, Values, and Needs?**
In IS504: Sociotechnical Information Systems, University of Illinois Urbana-Champaign (Instructor: Yue Guo)

11/2024 **LLM Reasoning (In-Context Learning, Prompting, and Reasoning)**
In CS475: ML for NLP, KAIST, South Korea (Instructor: Alice Oh)

11/2024 **In-Context Learning, Prompting, and Basics of Reasoning**
In CSE 447: Natural Language Processing, University of Washington (Instructor: Yulia Tsvetkov)

10/2024 **How to Build AI with Deep Concerns for Human Traits, Values, and Needs?**
In CS1684/2084: Bias and Ethical Implications in Artificial Intelligence, University of Pittsburgh (Instructor: Xiang Lorraine Li)

08/2024 **How to Build AI with Deep Concerns for Human Traits, Values, and Needs?**
In CSE 163: Intermediate Data Programming, University of Washington (Instructor: Yuxuan Mei)

09/2023 **Can We Teach Machines Human Ethics and Values?**
In Ethics and Citizenship, w/ Valentina Pyatkin and Taylor Sorensen, The Downtown School, Seattle

03/2023 **Toward Interpretable and Interactive Socially & Ethically Informed AI**
In CS496: AI Perspectives: Symbolic Reasoning to Deep Learning, Northwestern University (Instructor: Mohammed Anwarul Alam)

03/2023 **Toward Interpretable and Interactive Socially & Ethically Informed AI**
In LAW E 553: Technology Law And Public Policy Seminar, University of Washington (Instructor: Inyoung Cheong)

09/2022 **Toward Socially Aware & Ethically Informed AI**
In Ethics and Citizenship, w/ Saadia Gabriel, The Downtown School, Seattle

05/2022 **Toward Ethically Informed & Socially Aware AI**
In HONORS 222 B: Artificial Intelligence Meets Society, University of Washington (Instructor: Richard Freeman)

Tutorials

07/2025 **Guardrails and Security for LLMs: Safe, Secure, and Controllable Steering of LLM Applications**, ACL 2025
Co-Instructor w/ Traian Rebedea, Leon Derczynski, Shaona Ghosh, Makesh Narsimhan Sreedhar, Faeze Brahman, Bo Li, Yulia Tsvetkov, Christopher Parisien, and Yejin Choi

Teaching Assistant

01/2024–03/2024 **CSE447/517 Natural Language Processing**, UW
Head TA for the NLP class with 230+ undergraduate and graduate students
Co-design the class module, including teaching materials and homework

01/2023–03/2023 **CSE599 D1 Exploration on Language, Knowledge, and Reasoning**, UW
TA for a graduate seminar with 30+ students

09/2016–01/2019 **CS151 Introduction to Computational Thinking**, Colby College

CS231 Data Structure & Algorithm, Colby College
CS251 Data Analysis & Visualization, Colby College
Graded programming projects and homework, held TA office hours and tutored sessions weekly

09/2018–01/2019 **MA311 Ordinary Differential Equation**, Colby College
Held TA office hours and graded problem sets weekly for 30 students

MENTORING EXPERIENCES

Junior Ph.D. Students

12/2024–present **Mickel Liu** (PhD student at UW CSE)
Self-Play multi-agent RL training for LM safety enhancement [P.2].

10/2024–present **Salman Rahman** (PhD student at USC)
Extensive red-teaming of LMs for multi-turn adversarial attacks [P.3].

05/2024–10/2024 **Jing-Jing Li** (PhD student at Berkeley)
Interpretable harm and benefit analysis of user queries to language models [P.14].

06/2023–02/2024 **Jimin Mun** (PhD student at CMU)
A democratic surveying framework for future AI harms and benefits [P.18].

06/2023–03/2024 **Huihan Li** (PhD student at USC)
Multicultural symbol generation and evaluation [P.19].

05/2023–09/2023 **Linlu Qiu** (PhD student at MIT)
Inductive reasoning capabilities of language models [P.20].

09/2022–08/2023 **Taylor Sorensen** (PhD student at UW CSE)
Engaging machines with pluralistic human values, rights, and duties [P.21].

01/2022–05/2023 **Jillian Fisher** (PhD student at UW Statistics)
Model revision and authorship obfuscation [P.25].

Undergraduate & Master Students

04/2025–present **Neel Bhandari** (Master student at CMU LTI)
Interpretable safety moderation of reasoning-based models.

04/2024–present **Supriti Vijay** (Master student at CMU LTI)
Interpretable safety moderation of reasoning-based models.

01/2025–04/2025 **James Shiffer** (Master student at UCLA CSE)
Extensive red-teaming of LMs for multi-turn adversarial attacks [P.3].

10/2024–04/2025 **Yuanjun Chai** (Master student at UW ECE)
Diversity-enhancing LM alignment.

09/2024–present **Priyanshu Kumar** (Master student at CMU LTI)
Multi-lingual safety moderation tool [P.6].

09/2024–present **Devansh Jain** (Master student at CMU LTI)
Multi-lingual safety moderation tool [P.6].

01/2022–12/2024 **Kavel Rao** (Undergraduate student at UW CSE → SWE at Jane Street)
Explainable defeasible moral reasoning [P.28].
Open AI safety moderation tool & In-the-wild LM redteaming [P.16, P.17].
🏆 Single Awardee of the 2024 Best Senior Thesis Award at UW CSE

03/2023–12/2024 **Kelly Chiu** (Master student at UW Linguistics → Research Assistant at NYU)
A challenging cultural knowledge benchmark for LMs [P.10].
Using daily dilemmas to test language models' value preferences [P.11].

04/2023–06/2024 **Seungju Han** (Undergraduate student at SNU ECE → Ph.D. Student at Stanford)
Multimodal defeasible social and moral norm reasoning [P.30].
Open AI safety moderation tool & In-the-wild LM redteaming [P.16, P.17].

03/2023–07/2023 **Airei Fukuzawa** (Undergraduate student at UW CSE → SWE at Meta)
Enhancing LLMs with multi-cultural understanding and social norms.

09/2021–02/2023 **Sravani Nanduri** (Undergraduate student at UW CSE)
Co-mentored with Maarten Sap & Tongshuang (Sherry) Wu
Online hate speech moderation with explanations [P.34].

12/2021–03/2022 **Nuria Alina Chandra** (Undergraduate student at UW CSE)

PROFESSIONAL SERVICE

Organizing Committees

2025 **Socially Responsible Language Modeling Research** (SoLaR Workshop, COLM 2025)

2024 **Socially Responsible Language Modeling Research** (SoLaR Workshop, NeurIPS 2024)

2023 **AI Meets Moral Philosophy and Moral Psychology: An Interdisciplinary Dialogue about Computational Ethics** (MP2 Workshop, NeurIPS 2023)

Paper Reviewing

Conf. **EMNLP** 2022, **ACL** 2021, **AAAI** 2023, **NeurIPS** 2024, **NeurIPS D&B** 2024, **ICLR** 2025, **ICML** 2025, **ACL** 2025, **COLM** 2025, **NeurIPS** 2025, **NeurIPS D&B** 2025

Journal **Language Resources and Evaluation (Springer Nature)** 2024 , **Applied Artificial Intelligence** 2025

Community Service

2024 **Liaison**, UW Allen School Faculty Recruiting
Keep students informed about faculty recruiting and coordinate with student hosts to carry out responsibilities.

2023 **Area Chair, Reviewer**, UW Allen School PhD Admissions
Coordinate between PhD students/postdocs reading for PhD admissions and advising staff/admissions committee.
Liaison, UW Allen School Faculty Recruiting
Keep students informed about faculty recruiting and coordinate with student hosts to carry out responsibilities.
Co-chairs, UW Allen School Prospective Student Committee
Organize visit days for prospective PhD students at UW Allen School.

2022 **Volunteer**, UW NLP Retreat
Area Chair, Reviewer, UW Allen School PhD Admissions
Coordinate between PhD students/postdocs reading for PhD admissions and advising staff/admissions committee.
Liaison, UW Allen School Faculty Recruiting
Keep students informed about faculty recruiting and coordinate with student hosts to carry out responsibilities.
Student Representative, UW Allen School Diversity and Inclusion Committee
Co-chairs, UW Allen School Prospective Student Committee

Organize visit days for prospective PhD students at UW Allen School.

Volunteer Coordinator, NAACL 2022

2021 **Mentor**, UW Allen School Pre-Application Mentorship Service (PAMS)

A program supporting potential CS PhD applicants, with 80% from underrepresented communities.

Liaison, Reviewer, UW Allen School PhD Admissions

2020 **Co-organizer**, UW Allen School Pre-Application Mentoring Service (PAMS)

SELECTED MEDIA COVERAGE

2025 Delphi Experiment Tries to Equip an AI Agent with Moral Judgment
Tech Xplore (2025.1)

2024 WildTeaming: An Automatic Red-Team Framework to Compose Human-like Adversarial Attacks Using Diverse Jailbreak Tactics Devised by Creative and Self-Motivated Users in-the-Wild
MarkTechPost, (2024.7)

2023 How Moral Can A.I. Really Be?
The New Yorker (2023.11)
How Robots Can Learn to Follow a Moral Code
Nature Outlook (2023.10)

2022 Can Computers Learn Common Sense?
The New Yorker, (2022.4)

2021 Can a Machine Learn Morality?
The New York Times (2021.12)

2021 This Program Can Give AI a Sense of Ethics–Sometimes
Wired, (2021.12)

2021 Machines Learn Good From Commonsense Norm Bank
IEEE Spectrum, (2021.11)

Updated May 2025