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.github.io
@liweijianglw

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 *(GPA 4.08, top 0.5%, Dean's List 15–19)* **Waterville, ME**

09/2015–01/2019 B.A. in Computer Science and B.A. in Mathematics, *summa cum laude*

HONORS AND AWARDS

2025 Best Paper Award
NeurIPS 2025 (Datasets and Benchmarks Track)

Outstanding Paper Award
AI Agents: Capabilities and Safety (AIA) Workshop @ COLM 2025

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

2024 Best 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

PUBLICATIONS

*, † denote equal contribution; Google Scholar; Semantic Scholar

Manuscripts and Pre-Prints

P.1 **Liwei Jiang**, Erick Galinkin, Makesh Narsimhan Sreedhar, Chong Xiang, Yejin Choi, Traian Rebedea, and Christopher Parisien. HieraSuite: A Holistic Toolkit for Building Versatile System-User Instruction Hierarchy.
In submission to ICLR 2026

P.2 Mickel Liu*, **Liwei Jiang***, 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 to ICLR 2026
Multi-Agent Systems in the Era of Foundation Models (MAS) Workshop @ ICML 2025
AI Agents: Capabilities and Safety (AIA) Workshop @ COLM 2025 **Outstanding Paper Award**

P.3 Hongjue Zhao, Haosen Sun, Jiangtao Kong, Xiaochang Li, Qineng Wang, **Liwei Jiang**, Qi Zhu, Tarek F. Abdelzaher, Yejin Choi, Manling Li, and Huajie Shao. Activation Steering for LLM Alignment via a Unified ODE-Based Framework.
In submission to ICLR 2026

P.4 Taylor Sorensen, Benjamin Newman, Jared Moore, Chan Young Park, Jillian Fisher, Niloofar Mireshghallah, **Liwei Jiang**, and Yejin Choi. Spectrum Tuning: Post-Training for Distributional Coverage and In-Context Steerability.
In submission to ICLR 2026

P.5 Prasoon Varshney*, Makesh Narsimhan Sreedhar*, **Liwei Jiang**, Traian Rebedea, and Christopher Parisien. PluralisticBehaviorSuite: Stress-Testing Multi-Turn Adherence to Custom Behavioral Policies. In submission
Multi-Turn Interactions in LLMs (MTI-LLM) Workshop @ NeurIPS 2025

P.6 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 *Featured in TechCrunch*

Journal Publications

2025 J.1 **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 *Featured in The New York Times, The New Yorker, Vox, IEEE Spectrum, The Guardian, Nature Outlook, Wired, TechXplore*
Catalyzed the Darpa In the Moment (ITM) program
The public demo hosted by Ai2 received 4M+ queries

Peer-Reviewed Conference Publications

2025 P.7 **Liwei Jiang**, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial Hiveminds: The Open-Ended Homogeneity of Language Models (and Beyond).
NeurIPS 2025 (Datasets & Benchmarks Track) **Best Paper Award, Oral**

P.8 **Liwei Jiang**, Taylor Sorensen, Sydney Levine, and Yejin Choi. Can Language Models Reason about Individualistic Human Values and Preferences?.
ACL 2025
Pluralistic Alignment Workshop @ NeurIPS 2024

P.9 Salman Rahman*, **Liwei Jiang***, 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.
COLM 2025
Multi-Agent Systems in the Era of Foundation Models (MAS) Workshop @ ICML 2025

P.10 Salman Rahman, Sheriff Issaka, Ashima Suvarna, Genglin Liu, James Shiffer, Jaeyoung Lee, Md Rizwan Parvez, Hamid Palangi, Shi Feng, Nanyun Peng, Yejin Choi, Julian Michael, **Liwei Jiang**, and Saadia Gabriel. AI Debate Aids Assessment of Controversial Claims.
NeurIPS 2025
Multi-Agent Systems in the Era of Foundation Models (MAS) Workshop @ ICML 2025

P.11 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.12 Yu Ying Chiu, **Liwei Jiang**, and Yejin Choi. DailyDilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life.
ICLR 2025 **Spotlight, Top 5.1%**

P.13 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.
COLM 2025
Towards Safe & Trustworthy Agents Workshop @ NeurIPS 2024

P.14 Priyanshu Kumar*, Devansh Jain*, Akhila Yerukola, **Liwei Jiang**, Himanshu Beniwal, Tom Hartvigsen, and Maarten Sap. PolyGuard: A Multilingual Safety Moderation Tool for 17 Languages.
COLM 2025

P.15 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, Top 1.8%**
Featured in Science Magazine News

P.16 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, Top 3.3%**
Featured in Stanford HAI Policy Brief

P.17 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
Pluralistic Alignment Workshop @ NeurIPS 2024

P.18 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.19 **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 Language Models.
NeurIPS 2024
NextGenAISafety Workshop @ ICML 2024 *Data and evaluation suite were used in the safety post-training of Ai2's OLMo 2, OLMo 3, and Tulu 3 models*
Adopted in the training of Skywork-Reward, the top-performing reward model on RewardBench

P.20 Seungju Han*, Kavel Rao*, Allyson Ettinger†, **Liwei Jiang**†, 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 2024 (Datasets & Benchmarks Track) *230K total model downloads on HuggingFace*

P.21 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.22 Huihan Li, **Liwei Jiang**, Jena D. Hwang, 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
TrustNLP Workshop @ NAACL 2024

P.23 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, Top 1.2%**

P.24 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 **Oral, Top 3%**

P.25 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.26 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 *#22 most influential 2024 arXiv AI paper by PaperDigest*
Featured in Jack Clark's Import AI and Interconnects
Featured in the Keynote Talk at ACL 2025 by Verena Rieser
Catalyzed the Pluralistic Alignment Workshop @ NeurIPS 2024

P.27 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.28 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.29 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 **Best Paper Award, Oral**

P.30 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.31 Kavel Rao*, **Liwei Jiang***, 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.32 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.33 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.34 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, Oral**

P.35 Nouha Dziri*, Ximing Lu*, Melanie Sclar*, Xiang Lorraine Li†, **Liwei Jiang**†, 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, Top 3.1%**
Featured in ScienceNews, Quanta Magazine

P.36 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.37 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.38 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.39 Prithviraj Ammanabrolu, **Liwei Jiang**, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. Aligning to Social Norms and Values in Interactive Narratives.
NAACL 2022 **Oral**

P.40 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, Oral**

P.41 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 **Oral**

P.42 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 *Featured in BBC Science Focus*

P.43 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 **Oral**

2021 P.44 **Liwei Jiang**, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. "I'm Not Mad": Commonsense Implications of Negation and Contradiction.
NAACL 2021

P.45 Sherry Ruan*, **Liwei Jiang***, 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.46 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 *Featured in World Economic Forum,* Stanford Report

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

GRANTS CONTRIBUTED

2023-Present The In the Moment (ITM) Program at Darpa
The Delphi project served as a seminal catalyst for the establishment of the ITM grant. I contributed to the grant proposal, led the preparation of progress reports and weekly meeting participation, and delivered 4 presentations at PI meetings, including the program's kickoff meeting.

PROFESSIONAL EXPERIENCES

NVIDIA **Santa Clara, CA**

03/2025–current Research intern at the NeMo Guardrails team, with *Erick Galinkin* and *Christopher Parisien*
Secure and versatile instruction hierarchy of language models [P.1].
Multi-turn pluralistic behavioral policy adherence of language models [P.5].

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

02/2025–09/2025 Visiting graduate student, with *Yejin Choi*

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

12/2024–06/2025 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 and alignment [P.2, P.3, P.9, P.10, P.13, P.14, P.16, P.17, P.18, P.19, P.20, P.21].
Computational morals and norms [J.1, P.12, P.31, P.33, P.38, P.39].
Pluralisticc alignment to values [P.4, P.7, P.8, P.24, P.26] and cultures [P.11, P.22].
Various prosocial AI applications [P.6, P.28, P.34, P.36, P.37, P.40, P.42, P.43].
The interplay between human and machine capabilities [P.15, P.23, P.30, P.35].
Symbolic data distillation and data synthesis [P.25, P.27, P.32, P.41].
Commonsense negation [P.44].

09/2019–08/2020 Research Assistant, with *James Fogarty*
Personalized, longitudinal health self-tracking tools for migraine patients [P.29].

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

06/2017–09/2019 Research Intern, with *James Landay*
Educational interactive conversational systems, including QuizBot for factual knowledge learning [P.46], EnglishBot for second language learning [P.45], and BookBuddy for narrative reading learning of children [W.1].

TALKS

2026 **Humanistic, Pluralistic, and Coevolutionary AI Safety and Alignment**
(Upcoming) Invited Talk, UCLA NLP Seminar (2026.1)
Humanistic, Pluralistic, and Coevolutionary AI Safety and Alignment
(Upcoming) Invited Talk, UIUC ECE, hosted by Huan Zhang (2026.1)

2025 **Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)**
Invited Speaker, NVIDIA (2025.12)
Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)
Oral Talk Presenter, NeurIPS 2025 (2025.12)
Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)
Fireside Discussion Speaker, Ploutos (2025.12)
WildTeaming and WildGuard: Building Robust Model-Level and System-Level Safeguards of Language Models
Speaker, Netskope (2025.5)
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
Speaker, All-Ai2 Meeting, Allen Institute for Artificial Intelligence (Ai2) (2024.7)

2023 **On the Outcomes of Scientific Disagreements on Machine Morality**
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

Conference Tutorials

07/2025 **Guardrails and Security for LLMs: Safe, Secure, and Controllable Steering of LLM Applications**
ACL 2025 Tutorial
Co-Instructor w/ Traian Rebedea, Leon Derczynski, Makesh Narsimhan Sreedhar, Prasoon Varshney, and Yulia Tsvetkov
Consistently ∼400 in-person and 50 ∼online attendees throughout the tutorial.
Featured in the Keynote Talk at ACL 2025 by Verena Rieser.

Guest Lectures

11/2025 **In-Context Learning, Prompting, and Basics of Reasoning**
In CSE 447: Natural Language Processing, University of Washington (Instructor: Yulia Tsvetkov)

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, w/ Nouha Dziri, CMU (Instructor: Maarten Sap)

11/2024 **How to Build AI with Deep Concerns for Human Traits, Values, and Needs?**
In IS504: Sociotechnical Information Systems, UIUC (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)

Teaching Assistant

12/2023–03/2024 **CSE447/517 Natural Language Processing (Grad + Undergrad)**, 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 (Grad)**, UW
TA for a graduate-level seminar with over 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

09/2018–01/2019 **MA311 Ordinary Differential Equation**, Colby College

MENTORING EXPERIENCES

Junior Ph.D. Students

10/2025–present **Carrie Yuan** (PhD student at UW CSE)
Improving the mode collapse of large language and reasoning models.

07/2025–present **Mingqian Zheng** (PhD student at CMU LTI)
Realistic over-refusal simulation and evaluation for multi-turn dialogues.

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

10/2024–09/2025 **Salman Rahman** (PhD student at USC)
Extensive multi-turn red-teaming of LMs [P.9].
Scalable oversight over controversial claims [P.10].

05/2024–10/2024 **Jing-Jing Li** (PhD student at Berkeley)
Co-mentored with Sydney Levine
Interpretable harm and benefit analysis of user queries to language models [P.17].

06/2023–02/2024 **Jimin Mun** (PhD student at CMU)
A democratic surveying framework for future AI harms and benefits [P.21].

06/2023–03/2024 **Huihan Li** (PhD student at USC)
Multicultural symbol generation and evaluation [P.22].

09/2022–01/2024 **Taylor Sorensen** (PhD student at UW CSE)
Engaging machines with pluralistic human values, rights, and duties [P.24].

01/2022–05/2023 **Jillian Fisher** (PhD student at UW Statistics/CSE)
Model revision and authorship obfuscation [P.28].

Undergraduate, Master, Pre-Doctoral Students

08/2025–present **Hangoo Kang** (Master student at Stanford University)
PluralisticDataSmith, an open-source, scalable synthetic data engine.

07/2025–present **Ahnjae Shin** (Master student at Seoul National University)
Co-mentored with Hyunwoo Kim
PluralisticDataSmith, an open-source, scalable synthetic data engine.

07/2025–present **James (Jihao) Liu** (Undergrad student at Stanford)
Multi-agent training of LMs for emergent alter egos.

08/2024–present **Supriti Vijay** (Master student at CMU LTI)
Co-mentored with Jimin Mun
Interpretable safety moderation of reasoning-based models.

08/2025–present **Neel Bhandari** (Master student at CMU LTI)
Co-mentored with Jimin Mun
Interpretable safety moderation of reasoning-based models.

05/2025–09/2025 **Matthias Kleiner** (Master student at ETH Zürich)
Co-mentored with Irene Chen
Pluralistic medical opinions.

01/2025–04/2025 **James Shiffer** (Master student at UCLA CSE)
Extensive red-teaming of LMs for multi-turn adversarial attacks [P.9].

10/2024–04/2025 **Yuanjun Chai** (Master student at UW ECE)
Artificial Hivemind: the lack of open-endedness of LMs and beyond [P.7].

09/2024–06/2025 **Priyanshu Kumar** (Master student at CMU LTI → MLE at Apple)
Multi-lingual safety moderation tool [P.14].

09/2024–06/2025 **Devansh Jain** (Master student at CMU LTI)
Multi-lingual safety moderation tool [P.14].

01/2022–12/2024 **Kavel Rao** (Undergraduate student at UW CSE → SWE at Jane Street)
Explainable defeasible moral reasoning [P.31].
Open AI safety moderation tool & In-the-wild LM redteaming [P.19, P.20].
🏆 *Single Awardee of the 2024 Best Senior Thesis Award at UW CSE for [P.31]*

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

04/2023–06/2024 **Seungju Han** (Undergraduate student at SNU ECE → Ph.D. Student at Stanford)
Multimodal defeasible social and moral norm reasoning [P.33].
Open AI safety moderation tool & In-the-wild LM redteaming [P.19, P.20].

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.37].

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

High School Students

07/2025–present **Abhay Gupta** (High school student at John Jay Senior High School)
PluralisticDataSmith, an open-source, scalable synthetic data engine.

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

Reviewer

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, **ICLR** 2026

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

Area Chair

WS. **MTI-LLM @ NeurIPS** 2025

Community Service

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

2023 **Area Chair, Reviewer**, UW Allen School PhD Admissions
Managing the distribution and review of 200+ PhD applications by PhD students, postdocs, advising staff, and the admissions committee.
Liaison, UW Allen School Faculty Recruiting
Keeping students informed about faculty recruiting and coordinate with student hosts to carry out responsibilities.
Co-Chairs, UW Allen School Prospective Student Committee
Organizing visit days for prospective PhD students at UW Allen School.

2022 **Co-Organizer**, UW NLP Retreat
Organizing the UW NLP offsite retreat for 200+ students, faculty, and external collaborators.
Area Chair, Reviewer, UW Allen School PhD Admissions
Managing the distribution and review of 200+ PhD applications by PhD students, postdocs, advising staff, and the admissions committee.
Liaison, UW Allen School Faculty Recruiting
Keeping students informed about faculty recruiting and coordinate with student hosts to carry out responsibilities.
Student Representative, UW Allen School Diversity and Inclusion Committee
Reviewing diversity statements for the UW Allen School faculty search.
Co-Chairs, UW Allen School Prospective Student Committee
Organizing visit days for prospective PhD students at UW Allen School.
Volunteer Coordinator, NAACL 2022
Volunteering for overseeing onsite check-ins, registration, and attendee support.

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
Managing the distribution and review of 200+ PhD applications by PhD students, postdocs, advising staff, and

the admissions committee.

2020 **Co-organizer**, UW Allen School Pre-Application Mentoring Service (PAMS)
A program supporting potential CS PhD applicants, with 80% from underrepresented communities.

SELECTED RESEARCH MEDIA COVERAGE

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

Chatbot Software Begins to Face Fundamental Limitations
Quanta Magazine (2025.1)

2024 AI writing is improving, but it still can't match human creativity
Science Magazine News, (2024.12)

Study suggests that even the best AI models hallucinate a bunch
TechCrunch, (2024.8)

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)

AI's understanding and reasoning skills can't be assessed by current tests
ScienceNews, (2024.7)

Faith and Fate: Transformers as fuzzy pattern matchers
Answer.AI, (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)

How ChatGPT – and Ted Lasso – could rid the internet of hate speech
BBC Science Focus, (2023.2)

2022 Can Computers Learn Common Sense?
The New Yorker, (2022.4)
AI Is Already Making Moral Choices for Us. Now What?
Nautilus, (2022.1)

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

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

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

'Is it OK to …': the bot that gives you an instant moral judgment
The Guardian, (2021.11)

How well can an AI mimic human ethics?
Vox, (2021.10)

2019 This AI quizbot is helping students to learn
World Economic Forum, (2019.5)

Stanford's 'QuizBot' helps students retain 25 percent more information
EdScoop, (2019.6)

Stanford's 'QuizBot' – a chatbot that teaches – beats flashcards for learning factual information
Stanford Report, (2019.5)

Updated December 2025