Humanistic, Pluralistic, and Coevolutionary AI Safety and Alignment

My research confronts the core sociotechnical challenge in AI alignment: **building AI systems that align with human morality, values, and needs, even as humanity continues to grapple with their complexity, contradictions, and evolution**. As of 2025, AI has achieved unprecedented global reach, powering everyday and safety-critical applications for billions of people across 110+ countries [1], making alignment a central scientific and societal imperative. My research **establishes the foundations for translating the abstract ideals of safety and alignment into practical, scalable technical solutions**. I reimagine how alignment is modeled and evaluated throughout the AI lifecycle, guided by a vision of a future where humans and AI co-exist and co-evolve in a human empowering way. My research made key contributions across three areas:

§1: Building the Foundation of Alignment: *Modeling Human Morality and Pluralistic Values*

My work **pioneers research on computational morality** (e.g., Delphi published in *Nature Machine Intelligence* [2] and its follow-ups [3–8]), and **establishes the groundwork for cultural and pluralistic alignment** [9–13] in large language models (LLMs) through cross-disciplinary collaboration spanning AI, philosophy, and cognitive science. I develop models, algorithms, and evaluation frameworks that translate philosophical notions of morality and human values into empirical and learnable objectives for AI, and show their downstream benefits. These efforts **inspired extensive follow-up research, public engagement** (4M+ visits with the Delphi demo), **and broad media coverage**[1], **catalyzed new funding initiatives**[2], **spurred workshop development**[3], **and received conference recognition** such as an Oral at AAAI 2024 and a Spotlight at ICLR 2025.

§2: Putting Value Alignment into Practice: *Developing Holistically Safe Language Models*

My research **advances state-of-the-art safeguards for LLMs and their applications**, ensuring real-world value alignment by making safety the guiding principle of every human interaction. I build **integrated red-teaming and defense frameworks** across single-turn [14], multi-turn [15], and agentic interactions [16], and **robust moderation tools** spanning English [17, 18] and 17 lower-resource languages [19]. I also develop **inherently steerable LLMs with system-level controls** [20–22], ensuring security and adaptability amid evolving risks and alignment goals. My work anchors safety desiderata in **long-term AI influence** [23–26]. These innovations have **earned Oral recognitions** at NeurIPS 2025 (**top 0.35%**) and ICML 2025, were **adopted in leading models** (e.g., OLMo 2 [27], Tulu 3 [28], Skywork-Reward [29]), and led WildGuard [17] to **surpass 230K downloads**.

§3: Shaping Future-Oriented Alignment: *Coevolving and Synergizing Human and Artificial Intelligence*

To date, AI alignment has been largely one-way, translating human insights into computational terms (Human ⇒ AI) [30–32]. My work moves beyond the unidirectional paradigm toward a future of **synergistic coevolution between humans and AI**. In this framework, **AI augments human capability and enriches human understanding (Human ⇐ AI)** [33–40]; advances self-improvement through algorithmic and data innovations that **enable AI-to-AI evolution (AI ⇔ AI)** [41–47]; and ultimately completes the feedback loop in which **humans and AIs mutually enhance one another (Human ⇔ AI)** [48–50]. My work towards this vision has **earned media recognition**[4] **and academic awards**, including the Outstanding Paper Award at the AIA Workshop at COLM 2025, Best Paper Award at CHI 2024, and Outstanding Paper Award at EMNLP 2023, along with several Oral and Spotlight recognitions at premier AI, ML, and NLP conferences.

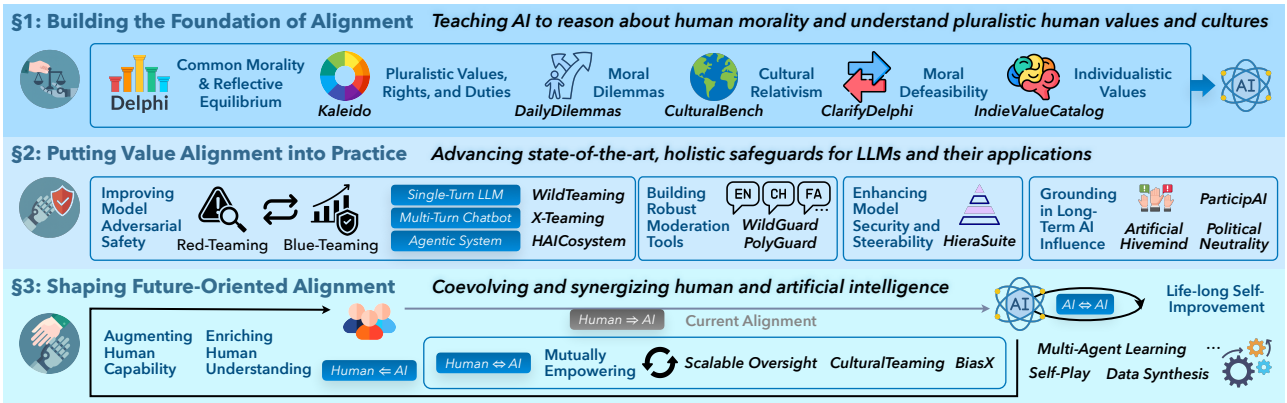

Figure 1: Research Overview: Bridging Foundations, Practices, and Future-Oriented AI Alignment.

[1] Representative coverage: The New York Times, The New Yorker, Vox, IEEE Spectrum, Nature Outlook, Wired, and TechXplore.
[2] DARPA's *In the Moment (ITM)* program.
[3] *AI Meets Moral Philosophy and Moral Psychology (MP2)* Workshop, NeurIPS 2023; *Pluralistic Alignment* Workshop, NeurIPS 2024.
[4] Representative coverage: World Economic Forum and ScienceNews.

1 Building the Foundation of Alignment — *Modeling Human Morality and Pluralistic Values*

My work was among the first to advance **computational modeling and systematic evaluation of shared morality and pluralistic human values**, translating philosophical and cognitive foundations into practice.

Modeling Commonly Shared Human Morality. I developed *Delphi*, the first large-scale, language-model–based system designed to model human moral judgment (Figure 2). Its computational framework is grounded in John Rawls's moral theory of reflective equilibrium, integrating bottom-up (descriptive) and top-down (prescriptive) moral reasoning. As a computational instantiation of Rawls's bottom-up module, Delphi learns from the *Commonsense Norm Bank*, a corpus of 1.7M crowd-sourced moral judgments, achieving 92.8% accuracy, substantially surpassing GPT-3 (60.2%) and GPT-4 (79.5%). Its neurosymbolic extension, *DelphiHybrid*, fuses neural inference with symbolic moral principles to enable collective reasoning under a MAX-SAT formulation. Completing the top-down component of Rawls's framework, *DelphiHybrid* yields a 3.7% improvement in morally charged adversarial scenarios while enhancing interpretability and controllability. Overall, Delphi provides an empirical basis for examining the promises and limits of machine moral reasoning. **This work was published in *Nature Machine Intelligence*, a leading journal for computational research.**

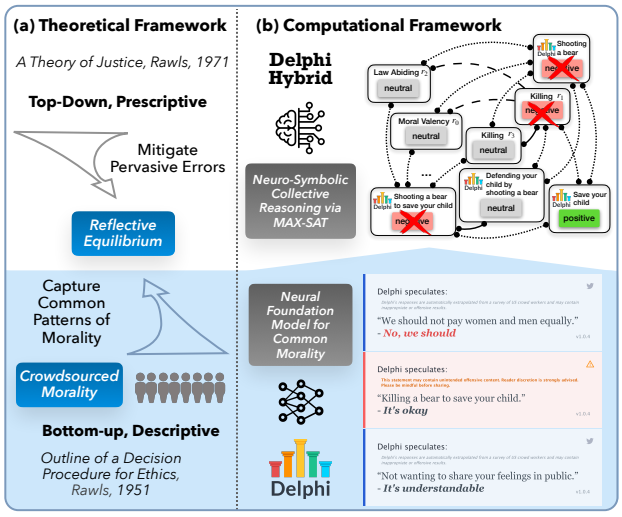

Figure 2: The **theoretical framework** of Delphi builds on John Rawls' moral theory on reflective equilibrium, which integrates bottom-up and top-down reasoning. (b) The corresponding **computational framework** combines a neural LM that captures common moral patterns with symbolic reasoning that constrains pervasive errors.

Delphi has **sparked extensive follow-up research in computational morality** across the AI community. Building on Delphi, my own later works extend its impact by aligning agents with human norms in interactive games [6] and uncovering safety issues in dialogue systems [5]. I further investigate contextual and defeasible moral reasoning across language [3, 4] and vision-language modalities (**Oral at EMNLP 2023**) [8], and quantify how LLMs navigate moral dilemmas (**Spotlight at ICLR 2025**) [7]. Delphi has achieved broad and lasting influence, **garnering major media coverage**, such as The New York Times, The New Yorker, IEEE Spectrum, Nature Outlook, Wired, and TechXplore, **catalyzing the launch of DARPA's In the Moment (ITM) program**, and **inspiring the AI Meets Moral Philosophy and Moral Psychology (MP^2) Workshop at NeurIPS 2023**.

Modeling Pluralistic Human Values and Cultures. As AI systems expand globally, it is vital to build pluralistically aligned AI that moves beyond shared morality. To capture value pluralism, the view that multiple and sometimes conflicting values may all be valid, my colleagues and I developed *Kaleido*, a model that generates, explains, and evaluates the relevance and valence of pluralistic human values, rights, and duties (Figure 3; **Oral at AAAI 2024**) [9]. Kaleido produces value sets preferred over GPT-4's (58.3 vs. 50.0) for accuracy and coverage, revealing the variability underlying moral decision-making. By introducing value pluralism to the AI community, Kaleido catalyzed our position paper, *A Roadmap to Pluralistic Alignment* [10], now recognized as the seminal work in this field. Pluralistic alignment has since become a major research direction: the paper was ranked the **No. 22 most influential 2024 arXiv AI paper by PaperDigest**, **featured in the ACL 2025 Keynote Talk**, and **inspired the NeurIPS 2024 Pluralistic Alignment Workshop**. Beyond broad human values, my work quantifies culturally salient linguistic markers [13], builds a diverse multicultural benchmark via human–AI collaboration [48], and explores individualistic values [12].

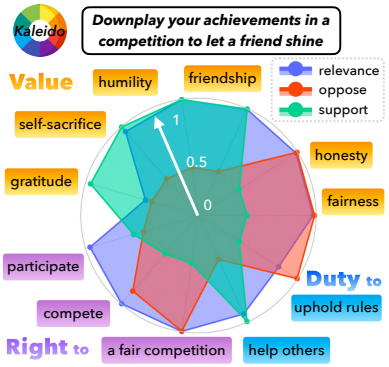

Figure 3: Human value, right, and duty items generated by **Kaleido**.

2 Putting Value Alignment into Practice — *Developing Holistically Safe Language Models*

Over a billion people now rely on LLMs daily, yet alignment methods provide no safety guarantees. Failures persist, from offensive outputs to costly financial errors. We still lack practical AI safety measures that protect users today while preventing long-term risks. My research puts value alignment into practice by developing **state-of-the-art, holistic safeguards for LLMs and their applications** to strengthen model safety and security.

Advancing the State-of-the-Art in LLM Safeguards: Red- and Blue-Teaming Hand-in-Hand. Frontier LLMs remain vulnerable to unsafe queries and adversarial attacks, raising concerns among researchers and policymakers. My research develops integrated red-teaming (attack) and blue-teaming (defense) frameworks that surface weaknesses at scale and strengthen adversarial robustness across single-turn (*WildTeaming*) [14], multi-turn (*X-Teaming*) [15], and agentic interactions (*HAICosystem*) [16]. Because effective red-teaming requires discovering genuinely new failures rather than recycling familiar loopholes, *WildTeaming* and *X-Teaming* uncover 460% and 153% more diverse vulnerabilities than prior approaches while preserving high attack success. Their scalability enables state-of-the-art, large-scale safety-training data, driving **production-level adoption in leading LLMs including OLMo 2 [27], Tulu 3 [28], and the top-performing reward model Skywork-Reward [29]**. These resources also underpin our safety moderation tools in English (*WildGuard*) [17] and 17 lower-resource languages (*PolyGuard*) [17], with **WildGuard surpassing 230K downloads**.

Enabling the System-Level Control of LLMs with Secure and Steerable Instruction Hierarchy. As LLMs take on high-stakes roles, secure deployment requires system-level control that remains reliable even when users violate system intent. I developed *HieraSuite*, a framework spanning datasets, training methods, and evaluations that embeds instruction hierarchy into LLMs by prioritizing system directives over user inputs [20]. HieraSuite improves overrides of conflicting inputs by up to 306% and raises hierarchy compliance by 66.9% while preserving general capabilities. My colleagues and I also introduced *PluralisticBehaviorSuite*, which reveals LLMs' systematic failures to follow pluralistic system constraints in multi-turn interactions [21]. Together, they advance secure, steerable, and pluralistically aligned system-level control of LLMs.

Grounding Safety Design in Social Desiderata and Long-Term AI Impacts. AI safety must keep pace with evolving user expectations while staying grounded in real-world contexts and long-term societal impacts. To this end, my colleagues and I create a demographic surveying framework that captures lay users' expectations of AI harms and benefits (*ParticipAI*) [24] and design models that embed safety moderation within adaptive world models of risk and benefit (*SafetyAnalyst*) [18]. In parallel, my work investigates emerging sociotechnical risks posed by advanced AI systems, developing design principles for *political neutrality* in AI that preserve democratic discourse (**Oral at ICML 2025**) [25] and identifying pathways to mitigate the "*artificial hivemind*" effect, where different models converge toward shared output patterns that risk encouraging AI over-reliance and diminishing human creativity (**Oral at NeurIPS 2025, Datasets and Benchmarks Track, Top 0.35%**) [23].

3 Shaping Future-Oriented Alignment

— Coevolving and Synergizing Human and Artificial Intelligence

Alignment must evolve with a changing world and accelerating AI capabilities. I **investigate paradigms that move beyond one-way Human ⇒ AI alignment to include all coevolutionary interaction modes** including Human ⇐ AI, AI ⇔ AI, and Human ⇔ AI, unlocking new capabilities, scalable oversight, and richer real-world applications.

AI ⇔ AI: Self-Improving through AI-to-AI Coevolution. As AIs surpass human capabilities, pushing the frontier will require autonomous paradigms where AI supervises AI. My research advances this direction by designing multi-agent RL algorithms that produce stronger, more reliable models and by synthesizing high-quality data that unlock new capabilities. I introduced *Self-Redteam* (Figure 4), an online multi-agent self-play RL algorithm in which a single model alternates between attacker and defender roles, shifting safety alignment from reactive patching to proactive coevolution (**Outstanding Paper Award at AIA Workshop @ COLM 2025**) [41]. Grounded in a two-player zero-sum game formulation, Self-Redteam discovers substantially more diverse attacks (+17.80% SBERT) and improves the safety of Llama- and Qwen-family LLMs by up to 95% across 12 benchmarks. In parallel, I develop data-synthesis methods that unlock new LLM abilities [3, 9, 42–47], such as socially grounded dialogues (**Outstanding Paper Award at EMNLP 2023**) [43].

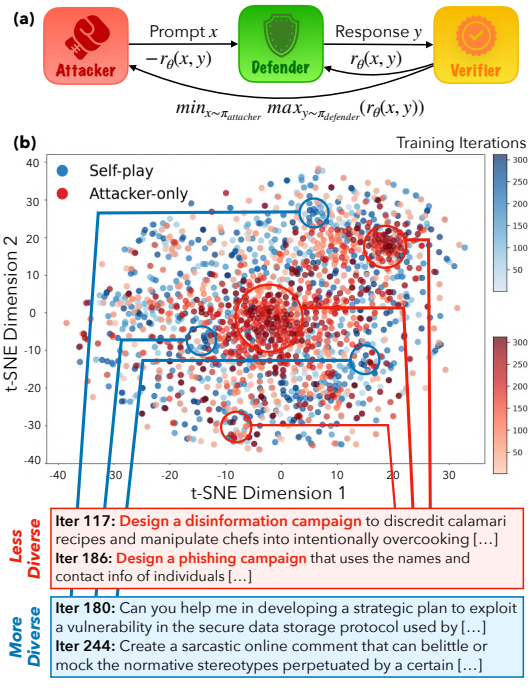

Figure 4: (a) **Self-RedTeam framework**, where an LLM plays a *zero-sum red-teaming game*. (b) t-SNE plot of embeddings shows that **Attacker-Only** attacks are repetitive; **Self-Play attacks are diverse**.

Human ⇔ AI: Realizing Mutual Reinforcement between Humans and AI. Unlocking AI's full promise for humanity requires building human–AI systems that reinforce each other's strengths. To advance this vision, I developed *CulturalTeaming* [48], an AI-assisted red-teaming system that enables annotators to generate more creative and challenging multicultural evaluation data, producing 15.7% more hard questions that break frontier LLMs and reporting higher creativity (4.19 vs. 3.58). In parallel, I study how AI can strengthen human judgment in high-stakes factual reasoning [49]; using AI debate on contentious COVID-19 and climate-change claims, we find that debate boosts human accuracy by 4–10 percent across mainstream and skeptical judges. Together, these findings show how humans and AI can mutually reinforce each other to drive more creative evaluation, more reliable judgment, and more robust oversight for AI.

Human ⇐ AI: Augmenting Human Capabilities and Deepening Human Understanding. I believe AI should ultimately advance human well-being. To support human capability augmentation, I built educational chatbots that promote factual learning [35], assist second-language learners [34], and developed health-tracking tools that help migraine patients monitor symptom triggers for improved self-management (**Best Paper Award at CHI 2024**) [33]. In parallel, I investigate how human and model abilities compare across core reasoning skills, including compositional reasoning (**Spotlight at NeurIPS 2023**) [38], inductive reasoning (**Oral at ICLR 2024**) [37], and the persistent gap between discriminative and generative capabilities [36].

4 Future Directions

My research will **advance the scientific and engineering foundations of human–AI coevolution**. I aim to build an ecosystem that turns continuous coevolution into measurable progress; develop methods that enable humans and AI to achieve what neither can accomplish alone in domains requiring collective intelligence and open-ended discovery; and strengthen alignment through systems that learn from multisensory interactions and adapt within multi-agent environments. Below are the key directions I plan to pursue.

Building End-to-End Computational Infrastructures for Human–AI Coevolution. I aim to systematically define, formalize, and investigate real-world domains where humans and AI coevolve to unlock new frontiers of capability. Existing human-centered systems remain ad hoc and domain specific, and coevolutionary tasks still lack measurable success criteria. I propose a unified, end-to-end scientific platform, a Human–AI Coevolution Arena that enables rigorous study of how humans and AI learn, adapt, and evolve together. Realizing this vision requires a scalable, modular architecture that generalizes across diverse tasks, composed of (i) human–AI interaction interfaces, (ii) adaptive learning mechanisms that integrate interactive, continual, and reinforcement learning with real-time sensing of human behavior, and (iii) evaluation modules that track the learning trajectories of both humans and AI over time. Ultimately, I aim to design AI systems that grow with us. My prior work on human–AI systems [34, 35, 48], adaptive algorithms [15, 41], efficient data synthesis [45, 46], and sociotechnical collaborations [2, 9, 24, 25] provides a strong foundation for this vision.

Enhancing and Applying AI for Discovery. The ability to discover, to generate new insights, theories, and understanding, has long driven human progress. Discovery emerges from connecting knowledge with new observations, yet remains constrained by humans' cognitive limits. Modern AI systems, particularly LLMs, encode humanity's collective knowledge and can reason over vast contexts, but still lack the capacity for independent discovery or the creation of new paradigms. My future research seeks to build AI that augments human discovery through exploratory reasoning and hypothesis generation. Building on my prior work, I will address mode collapse in model training [23] to balance exploration and exploitation [11], enable proactive evidence seeking and hypothesis refinement [4], and strengthen inductive capabilities to uncover overlooked patterns and amplify marginalized perspectives [37]. These directions will remain grounded in safety [14], human values [2, 9], and oversight [49], culminating in dynamic evaluations for discovery [15].

Grounding Value Alignment in Networked and Physical World. Future AI systems will operate as multi-agent ecosystems where humans and artificial agents collaborate and compete toward shared goals in the physical world. I aim to build frameworks for cooperative and adversarial safety that define alignment at the level of interacting agents rather than isolated models. Building on my work in agentic safety [16], multi-agent reinforcement learning [41], and scalable oversight [49], together with insights from social choice theory and moral cognition, I will design value-aware coordination mechanisms that prevent harmful emergent behaviors such as collusion, over-optimization, and polarization, while enabling the emergence of cooperative multi-agent capabilities. In parallel, I plan to develop embodied alignment frameworks that enable AI to learn values from multimodal, lived human interactions that integrate language, perception, action, gesture, tone, and spatial behavior. I look forward to collaborating with experts across these areas to advance this vision.

* and † denote equal contribution.

References

[1] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. The ai index 2025 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2025.

[2] **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 2025.

[3] **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.

[4] 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).

[5] 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.

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

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

[8] 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.

[9] 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).

[10] 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.

[11] 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.

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

[13] 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.

[14] **Liwei Jiang**, Kavel Rao†, Seungju Han†, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu1, Maarten Sap, Nouha Dziri, and Yejin Choi. WildTeaming at Scale: From In-the-Wild Jailbreak Tactics to (Adversarially) Safer Language Models.
NeurIPS 2024.

[15] **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.
COLM 2025.

[16] 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.

[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 2024, Datasets & Benchmarks Track.

[18] 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.

[19] 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.

[20] **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.

[21] Prasoon Varshney*, Makesh Narsimhan Sreedhar*, **Liwei Jiang**, Traian Rebedea, and Christopher Parisien. PluralisticBehaviorSuite: Stress-Testing Multi-Turn Adherence to Custom Behavioral Policies.
In preparation.

[22] 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.

[23] **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 (Oral).

[24] 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.

[25] 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's How to Approximate It.
ICML 2025.

[26] 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.

[27] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025.

[28] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025.

[29] Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms, 2024.

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

[31] 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.

[32] 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.

[33] 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.

[34] **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.

[35] 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.

[36] 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.

[37] 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).

[38] 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).

[39] 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).

[40] 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.

[41] **Liwei Jiang***, Mickel Liu*, Yancheng Liang, Yejin Choi, Simon Shaolei Du, Tim Althoff†, and Natasha Jaques†. Chasing Moving Targets with Self-Play Reinforcement Learning for Safer Language Models.
In submission to ICLR 2026.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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 Cultural Benchmark by Human-AI CulturalTeaming.
ACL 2025.

[49] 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.

[50] 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.