Weiyan Shi (she/her/hers)

CONTACT we.shi@northeastern.edu
wyshi.github.io

RESEARCH INTERESTS Human-AI Interaction, AI Safety, AI Agents

POSITIONS

Assistant Professor, Northeastern University — 2024.08 - Present
ECE & Khoury Computer Science

Postdoc, Stanford University — 2023.08 - 2024.06
Advisor: Diyi Yang

EDUCATION

Columbia University — 2021.01 - 2023.05
Ph.D. in Computer Science;
Advisor: Zhou Yu

University of California, Davis — 2018.09 - 2020.12
Ph.D. in Computer Science (transferred with advisor)

University of California, Berkeley — 2015.08 - 2016.05
M.A. in Statistics

Renmin University of China — 2011.09 - 2015.07
B.S. in Mathematics and Applied Mathematics

SELECTED AWARDS

Best Paper, ICML DL4C Workshop	2026
Outstanding Paper, NeurIPS SEA Workshop	2025
AI2050 Early Career Fellowship	2025
MIT Technology Review 35 Innovators under 35 (Global list)	2024
Best Social Impact Paper, ACL	2024
Outstanding Paper, ACL	2024
Best Paper Nomination, ACL	2019

PREPRINTS

4. **SPADE: Self-Play in Adaptive Synthetic Executable Environments**. Bo Liu*, Simon Yu*, Yiding Jiang, Ao Qu, Andrew Zhao, Zichen Liu, Junsu Kim, Zijian Zhou, Seungone Kim, Tongzheng Ren, Mickel Liu, Hanfei Yu, Zhaorun Chen, **Weiyan Shi**, Paul Pu Liang, Luke Zettlemoyer, Yejin Choi, Natasha Jaques (*equal contribution). *arXiv Preprint, 2026*.
3. **Shepherd: A Runtime Substrate Empowering Meta-Agents with a Formalized Execution Trace**. Simon Yu, Derek Chong, Ananjan Nandi, Dilara Soylu, Jiuding Sun, Christopher D Manning, **Weiyan Shi**. *arXiv Preprint, 2026*.
2. **Coding with "Enemy": Can Human Developers Detect AI Agent Sabotage?**. Jingheng Ye, Huiqi Zou, Simon Yu, **Weiyan Shi**. *arXiv Preprint, 2026*. Best Paper@ICML DL4C Workshop, 2026

1. **The Piggyback Hypothesis of Generalization: Explaining and Mitigating Emergent Misalignment**. Jiachen Zhao, Zhengxuan Wu, Aryaman Arora, Yiyou Sun, David Bau, **Weiyan Shi**. *arXiv Preprint, 2026*.

Journal Publications

3. **SocialFusion: Addressing Social Degradation in Pre-trained Vision-Language Models**. Hamza Tahboub, **Weiyan Shi**, Gang Hua, Huaizu Jiang. *TMLR, 2026*.
2. **Labeling Messages as AI-Generated Does Not Reduce Their Persuasive Effects**. Isabel Gallegos, Chen Shani, **Weiyan Shi**, Federico Bianchi, Izzy Gainsburg, Dan Jurafsky, Robb Willer. *PNAS Nexus, 2026*.
1. **Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning**. FAIR, Anton Bakhtin*, Noam Brown*, Emily Dinan*, Gabriele Farina, Colin Flaherty*, Daniel Fried, Andrew Goff, Jonathan Gray*, Hengyuan Hu*, Athul Paul Jacob*, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer*, Mike Lewis*, Alexander H. Miller*, Sasha Mitts, Adithya Renduchintala*, Stephen Roller, Dirk Rowe, **Weiyan Shi***, Joe Spisak, Alexander Wei, David Wu*, Hugh Zhang*, Markus Zijlstra (*core member; authors alphabetical). ***Science***, 2022. [Meta AI blog] [Science News]. **Coverage**: [NYT front page] [Washington Post] [Economist] [MIT Tech Review] [Forbes]

Conference Publications

42. **One Frozen Simulator Is Not Enough: Simulator Collapse in Multi-Agent RL**. Simon Yu, Nicholas Tomlin, Marwa Abdulhai, Ximing Lu, Derek Chong, Abe Hou, Dilara Soylu, Sergey Levine, Christopher D. Manning, **Weiyan Shi**. *EMNLP, 2026*.
41. **Which Metrics Save the Most Human Annotation? Prediction-Powered Evaluation and Meta-Evaluation**. Mingqi Gao, Anthony Sicilia, **Weiyan Shi**. *EMNLP, 2026*.
40. **Can Aha Moments be Fake? Identifying True and Decorative Thinking Steps in Chain-of-Thought**. Jiachen Zhao, Yiyou Sun, **Weiyan Shi***, Dawn Song*. *EMNLP, 2026*. [Quanta Magazine]
39. **Train Yourself as an LLM: Exploring Effects of AI Literacy on Persuasion via Role-playing LLM Training**. Qihui Fan, Min Ge, Chenyan Jia, **Weiyan Shi**. *COLM, 2026*.
38. **Unsafer in Many Turns: Benchmarking and Defending Multi-Turn Safety Risks in Tool-Using Agents**. Xu Li, Simon Yu, Minzhou Pan, Yiyou Sun, Bo Li, Dawn Song, Xue Lin, **Weiyan Shi**. *ICML, 2026*.
37. **Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity**. Jiayi Zhang*, Simon Yu*, Derek Chong*, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, **Weiyan Shi**. *ICML, 2026*. **Coverage**: [Forbes] [Yahoo] [VentureBeat] [TLDR AI] [Neuron Daily]
36. **SPIRAL: Self-Play on Zero-Sum Games Incentivizes Reasoning via Multi-Agent Multi-Turn Reinforcement Learning**. Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, **Weiyan Shi**, Min Lin, Wee Sun Lee, Natasha Jaques. *ICML, 2026*.
35. **Polyskill: Learning generalizable skills through polymorphic abstraction**. Simon Yu, Gang Li, **Weiyan Shi**, Peng Qi. *ICLR, 2026*.
34. **Gem: A Gym for Agentic LLMs**. Zichen Liu, Anya Sims, Keyu Duan, Changyu Chen, Simon Yu, Xiangxin Zhou, Haotian Xu, Shaopan Xiong, Bo Liu, Chenmien Tan, Chuen Yang Beh, Weixun Wang, Hao Zhu, **Weiyan Shi**, Diyi Yang, Michael Shieh, Yee Whye Teh, Wee Sun Lee, Min Lin. *ICLR, 2026*. Outstanding Paper@NeurIPS SEA Workshop 2025
33. **LLMs Encode Harmfulness and Refusal Separately**. Jiachen Zhao, Jing Huang, Zhengxuan Wu, **Weiyan Shi**. *NeurIPS, 2025*.

32. **Distilling an End-to-End Voice Assistant Without Instruction Training Data**. William Held, Yanzhe Zhang, Minzhi Li, **Weiyan Shi**, Michael J Ryan, Diyi Yang. *ACL, 2025*.
31. **NewsInterview: a Dataset and a Playground to Evaluate LLMs' Ground Gap via Informational Interviews**. Michael Lu, Hyundong Justin Cho, **Weiyan Shi**, Jonathan May, Alexander Spangher. *ACL, 2025*.
30. **Proactive Conversational Agents with Inner Thoughts**. Xingyu Bruce Liu, Shitao Fang, **Weiyan Shi**, Chien-Sheng Wu, Takeo Igarashi, Xiang Anthony Chen. *CHI, 2025*.
29. **PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action**. Yijia Shao, Tianshi Li, **Weiyan Shi**, Yanchen Liu, Diyi Yang. *NeurIPS Datasets and Benchmarks, 2024*.
28. **Zero-shot Persuasive Chatbots with LLM-Generated Strategies and Information Retrieval**. Kazuaki Furumai, Roberto Legaspi, Julio Cesar Vizcarra Romero, Yudai Yamazaki, Yasutaka Nishimura, Sina Semnani, Kazushi Ikeda, **Weiyan Shi**, Monica Lam. *EMNLP Findings, 2024*.
27. **CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies**. **Weiyan Shi**, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rlogério Abreu de Paula, Diyi Yang. *EMNLP Findings, 2024*.
26. **How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs**. Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, **Weiyan Shi**. *ACL, 2024*. (Best Social Impact Paper)
25. **The Earth is Flat because...: Investigating LLMs' Belief towards Misinformation via Persuasive Conversation**. Rongwu Xu, Brian S. Lin, Shujian Yang, Tianqi Zhang, **Weiyan Shi**, Tianwei Zhang, Zhixuan Fang, Wei Xu, Han Qiu. *ACL, 2024*. (Outstanding Paper)
24. **Social Intelligence Data Infrastructure: Structuring the Present and Navigating the Future**. Minzhi Li, **Weiyan Shi**, Caleb Ziems, Diyi Yang. *ACL Findings, 2024*.
23. **A Safe Harbor for AI Evaluation and Red Teaming**. Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, **Weiyan Shi**, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Sandy Pentland, Arvind Narayanan, Percy Liang, Peter Henderson. *ICML, 2024*.
22. **The Mirrored Influence Hypothesis: Efficient Data Influence Estimation by Harnessing Forward Passes**. Myeongseob Ko, Feiyang Kang, **Weiyan Shi**, Ming Jin, Zhou Yu, Ruoxi Jia. *CVPR, 2024*.
21. **When Life Gives You Lemons, Make Cherryade: Converting Feedback from Bad Responses into Good Labels**. **Weiyan Shi**, Emily Dinan, Kurt Shuster, Jason Weston, Jing Xu. *NAACL, 2024*.
20. **Dialoging Resonance in Human-Chatbot Conversation: How Users Perceive and Reciprocate Recommendation Chatbot's Self-Disclosure Strategy**. Kai-Hui Liang, **Weiyan Shi**, Yoo Jung Oh, Hao-Chuan Wang, Jingwen Zhang, Zhou Yu. *CSCW, 2024*.
19. **AutoReply: Detecting Nonsense in Dialogue with Discriminative Replies**. **Weiyan Shi**, Emily Dinan, Adi Renduchintala, Daniel Fried, Athul Jacob, Zhou Yu, Mike Lewis. *EMNLP Findings, 2023*.
18. **Controllable mixed-initiative dialogue generation through prompting**. Maximillian Chen, Xiao Yu, **Weiyan Shi**, Urvi Awasthi, Zhou Yu. *ACL, 2023*.
17. **Social Influence Dialogue Systems: A Scoping Survey of the Efforts Towards Influence Capabilities of Dialogue Systems**. Kushal Chawla*, **Weiyan Shi*** (equal contribution), Jingwen Zhang, Gale Lucas, Zhou Yu, Jonathan Gratch. *EACL, 2023*.
16. **Just Fine-tune Twice: Selective Differential Privacy for Large Language Models**. **Weiyan Shi**, Ryan Shea, Si Chen, Chiyuan Zhang, Ruoxi Jia, Zhou Yu. *EMNLP, 2022*.

15. **Seamlessly Integrating Factual Information and Social Content with Persuasive Dialogue**. Maximillian Chen, **Weiyan Shi**, Feifan Yan, Ryan Hou, Jingwen Zhang, Saurav Sahay, Zhou Yu. *AACL-IJCNLP, 2022*.
14. **Towards Socially Intelligent Agents with Mental State Transition and Human Utility**. Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, **Weiyan Shi**, Zhou Yu, Song-Chun Zhu. *SIGDIAL, 2022*.
13. **Selective Differential Privacy for Language Modeling**. **Weiyan Shi**, Aiqi Cui, Evan Li, Ruoxi Jia, Zhou Yu. *NAACL, 2022*.
12. **Refine and Imitate: Reducing Repetition and Inconsistency in Persuasion Dialogues via Reinforcement Learning and Human Demonstration**. **Weiyan Shi**, Yu Li, Saurav Sahay, Zhou Yu. *EMNLP Findings, 2021*.
11. **LEGOEval: An Open-Source Toolkit for Dialogue System Evaluation via Crowdsourcing**. Yu Li, Josh Arnold, Feifan Yan, **Weiyan Shi**, Zhou Yu. *ACL Demo, 2021*.
10. **PRAL: A tailored pre-training model for task-oriented dialogue generation**. Jing Gu, Qingyang Wu, Chongruo Wu, **Weiyan Shi**, Zhou Yu. *ACL, 2021*.
9. **INSPIRED: Toward Sociable Recommendation Dialogue Systems**. Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, **Weiyan Shi**, Zhou Yu. *EMNLP, 2020*.
8. **Structured Attention for Unsupervised Dialogue Structure Induction**. Liang Qiu, Yizhou Zhao, **Weiyan Shi**, Yuan Liang, Feng Shi, Tao Yuan, Zhou Yu, Song-Chun Zhu. *EMNLP, 2020*.
7. **Understanding User Resistance Strategies in Persuasive Conversations**. Youzhi Tian, **Weiyan Shi**, Chen Li, Zhou Yu. *EMNLP Findings, 2020*.
6. **Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies**. **Weiyan Shi**, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, Zhou Yu. *CHI, 2020*.
5. **End-to-End Trainable Non-Collaborative Dialogue System**. Yu Li, Kun Qian, **Weiyan Shi**, Zhou Yu. *AAAI, 2020*.
4. **How to Build User Simulators to Train RL-based Dialogue Systems**. **Weiyan Shi***, Kun Qian* (equal contribution), Xuewei Wang, Zhou Yu. *EMNLP, 2019*.
3. **Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good**. Xuewei Wang*, **Weiyan Shi*** (equal contribution), Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, Zhou Yu. *ACL, 2019*. (Best Paper Nomination)
2. **Unsupervised Dialogue Structure Learning**. **Weiyan Shi**, Tiancheng Zhao, Zhou Yu. *NAACL, 2019*.
1. **Sentiment Adaptive End-to-End Dialogue Systems**. **Weiyan Shi**, Zhou Yu. *ACL, 2018*.

GRANTS

Total funding: **$1.9M** *as single-PI,* **$2.3M** *overall*

Epistemic Sampling for Meta-Calibration 2026/09 - 2027/09
Coefficient Giving, Single-PI, $210K

Self-Exploring Safety Layer for Physical Environments 2026/06 - 2027/06
Samsung, Single-PI, $125K

Decoding Agentic Persuasion: Understanding and Monitoring Persuasion by AI Agents to Mitigate Harm 2025/11 - 2028/11
Schmidt Sciences, Single-PI, $500K

Mitigating Emergent Misalignment 2025/10 - 2027/03
Coefficient Giving, Single-PI, $330K

Agentic Sabotage Capability Evaluation 2025/08 - 2027/08
Coefficient Giving, Single-PI, $707K

Leveraging Generative AI for Updating Coastal Risk Assessments 2025/10 - 2027/10
Massachusetts AI Hub, Co-PI, $200K out of $500k

PATCH of Asclepius: Technology for Cybersecure Healthcare 2025/09 - 2026/09
ARPA-H, Co-I, $200K out of $14M

Students

Ph.D.

Simon Yu *(Google PhD Fellowship Department nominee; Best Paper, ICML DL4C 2026; Outstanding Paper, NeurIPS SEA 2025)* 2024 - Present
Jiachen Zhao *(MATS scholar)* 2024 - Present
Jingheng Ye *(Best Paper, ICML DL4C 2026)* 2025 - Present
Huiqi Zou *(Best Paper, ICML DL4C 2026)* 2025 - Present
Mingqi Gao 2025 - Present
Xu Li 2026 - Present
Barrett Lattimer 2026 - Present

Master

Qihui Fan *(first paper at COLM; → PhD, Stony Brook)* 2025 - 2026
Karmen Cai *(→ Amazon)* 2025 - 2026

Undergraduate

Xu Li *(first paper at ICML; → PhD, Northeastern)* 2025 - 2026

Mentees

Rongwu Xu *(ACL Outstanding Paper, now PhD at UW)* 2023 - 2024

Invited Talks

Git for Agents: Rewind, Fork, and Audit Your Agents with Meta-Agents: Human-Centered Machine Learning Workshop, Apple (2026.08).

Beyond the Surface: How Post-Training Artifacts Shape LLM Diversity and Safety: UPenn NLP Seminar (2026.03); Future of Life (2026.02); Stanford NLP Seminar (2025.10).

Persuasion for Good: How to Build and Break Chatbots: Alignment Workshop (2025.04); AAAI New Faculty Highlights (2025.02); Google Cloud AI (2024.11); Capital One AI (2024.11); Tufts University (2024.10); UMass Lowell Colloquium (2024.10); Meta Llama Team (2024.10); Amazon Research (2024.10); Adobe Research (2024.09).

Interactive AI Systems Specialized in Social Influence: CMU (2024.02); University of Southern California (2023.10); KAIST (2023.09); Stanford University NLP Seminar (2023.06); New York University (2023.05); NC State University (2023.05); Cornell Tech (2023.04).

Teaching

ECE 5668, Large Language Models Fall 2026

ECE 7398, Large-Language-Model-based Agents Spring 2026

(Effectiveness Mean=4.6, Median=5.0)

ECE 7398, Large-Language-Model-based Dialogue Agents Spring 2025

(Effectiveness Mean=4.7, Median=5.0)

CS 7180, Large-Language-Model-based Dialogue Agents Fall 2024

(designed this new course; Mean=3.6, Median=4.0)

Service

Workshop Organizer: 1st Workshop on Responsible Use of Meta-Agents (NeurIPS 2026); 1st Workshop on Multi-Turn Interactions in LLMs (NeurIPS 2025); 1st Workshop on Advancing LLM-Based Multi-Agent Collaboration (AAAI 2025); 2nd Workshop on Social Influence in Conversations (EMNLP 2024); 1st Workshop on Social Influence in Conversations (ACL 2023); 4th Workshop for Conversational AI (ACL 2022).

Area Chair: WiML 2022, EMNLP 2023-2026, NAACL 2023-2025, ACL 2023-2026, ICML 2026, NeurIPS 2026.

Conference Reviewer: ACL 2019-2022, *SEM 2020-2021, ICLR 2021-2022, AAAI 2021-2022, EACL 2021, NAACL 2021, NeurIPS 2021, EMNLP 2021, NLPCC 2021, ICML 2022, CHI 2022, SIGDIAL 2022, ACL Rolling Review 2021-2022, COLM 2025.

Journal Reviewer: ACM Transactions on Human-Robot Interaction; Neurocomputing; ACM Transactions on Information Systems; Nature Human Behaviour; PNAS Nexus.