Weiyan Shi (she/her/hers)

Contact

weiyans@stanford.edu
wyshi.github.io
Google Scholar

Research Interests

Natural Language Processing (dialogue systems, privacy-preserving NLP models, etc), Machine Learning, Artificial Intelligence

Employment History

Assistant Professor, Northeastern University 2024.08 -

Full-time Data Scientist, [24]7.ai 2016 - 2018

Education

Stanford University 2023.08 - Present
Postdoc
Advisor: Diyi Yang

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

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

Awards & Honors

Rising Star in AI, KAUST	2024
Rising Star in EECS, Georgia Tech	2022
Rising Star in Machine Learning, UMD	2022
Best Paper Nomination, ACL	2019
Dean's Distinguished Ph.D. Fellowship	2018-2023
Department Citation (top 1), UC Berkeley	2016
Speaker at Department Commencement (top 1), UC Berkeley	2016
National Scholarship, RUC	2014
Presidential Fellowship for Studying Abroad, RUC	2013

Papers

The listed conferences are top-tier in NLP, HCI and AI with acceptance rates between 20%-25%.

22. **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**
arXiv, 2024

21. **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
arXiv, 2024

20. **AutoReply: Detecting Nonsense in Dialogue with Discriminative Replies**
Weiyan Shi, Emily Dinan, Adi Renduchintala, Daniel Fried, Athul Jacob, Zhou Yu, Mike Lewis
Findings of the Association for Computational Linguistics: EMNLP 2023

19. **Controllable mixed-initiative dialogue generation through prompting**
Maximillian Chen, Xiao Yu, **Weiyan Shi**, Urvi Awasthi, Zhou Yu
Annual Meeting of the Association for Computational Linguistics (ACL), 2023

18. **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
(*A core team member. Authors listed alphabetically.)
Science, 2022. [Meta AI blog post] [Science News]
Selected media coverage: [The New York Times front page (01/22/2023)]
[The Washington Post] [The Economist] [MIT Technology Review] [Forbes]

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
European Chapter of the Association for Computational Linguistics (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
Empirical Methods in Natural Language Processing (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
Asia-Pacific Chapter of the Association for Computational Linguistics (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
Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), 2022

13. **Selective Differential Privacy for Language Modeling**
Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, Zhou Yu
North American Chapter of the Association for Computational Linguistics (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
Findings of Empirical Methods in Natural Language Processing (EMNLP), 2021

11. **LEGOEval: An Open-Source Toolkit for Dialogue System Evaluation via Crowdsourcing**

Yu Li, Josh Arnold, Feifan Yan, **Weiyan Shi**, Zhou Yu
Demo of Annual Meeting of the Association for Computational Linguistics (ACL), 2021

10. **PRAL: A tailored pre-training model for task-oriented dialogue generation**
Jing Gu, Qingyang Wu, Chongruo Wu, **Weiyan Shi**, Zhou Yu
Annual Meeting of the Association for Computational Linguistics (ACL), 2021

9. **INSPIRED: Toward Sociable Recommendation Dialogue Systems**
Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, **Weiyan Shi**, Zhou Yu
Empirical Methods in Natural Language Processing (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
Empirical Methods in Natural Language Processing (EMNLP), 2020

7. **Understanding User Resistance Strategies in Persuasive Conversations**
Youzhi Tian, **Weiyan Shi**, Chen Li, Zhou Yu
Findings of Empirical Methods in Natural Language Processing (EMNLP), 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
Conference on Human Factors in Computing Systems (CHI), 2020

5. **End-to-End Trainable Non-Collaborative Dialogue System**
Yu Li, Kun Qian, **Weiyan Shi**, Zhou Yu
AAAI Conference on Artificial Intelligence (AAAI), 2020

4. **How to Build User Simulators to Train RL-based Dialogue Systems**
Weiyan Shi*, Kun Qian* (equal contribution), Xuewei Wang, Zhou Yu
Empirical Methods in Natural Language Processing (EMNLP), 2019

3. **Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good (Best Paper Nomination)**
Xuewei Wang*, **Weiyan Shi*** (equal contribution), Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, Zhou Yu
Annual Meeting of the Association for Computational Linguistics (ACL), 2019

2. **Unsupervised Dialogue Structure Learning**
Weiyan Shi, Tiancheng Zhao, Zhou Yu
North American Chapter of the Association for Computational Linguistics (NAACL), 2019

1. **Sentiment Adaptive End-to-End Dialogue Systems**
Weiyan Shi, Zhou Yu
Annual Meeting of the Association for Computational Linguistics (ACL), 2018

INVITED TALKS

Interactive AI Systems Specialized in Social Influence
University of Hawaii (2023.01)
Rice University (2023.01)

Northwestern University, Statistics and Data Science (2023.01)
Arizona State University (2023.01)
Purdue University (2023.02)
CMU, LTI (2023.02)
Northeastern University (2023.03)
University of Wisconsin-Madison (2023.03)
Penn State University (2023.03)
Cornell Tech (2023.04)
New York University (2023.05)
NC State University (2023.05)
Stanford University NLP Group (2023.06)
Beijing Institute for General Artificial Intelligence (2023.08)
Korea Advanced Institute of Science & Technology (2023.09)
CMU, Robotics Institute (2024.02)

Social Influence Dialogue Systems
University of Maryland, 2022.12

Selective Differential Privacy in Language Modeling
Columbia NLP Seminar, (2021.11)

Reducing Repetition and Contradiction in Dialogue Systems
Columbia NLP Seminar, (2021.02)

Sentiment-adaptive Dialogue Systems
Cresta, (2019.09)

TEACHING

***Co-instructor*: COMS 6998, Conversational AI**, Columbia Fall 2022
· *Topics*: various topics on dialogue systems (task-oriented and open-domain dialogue systems, knowledge-enriched dialogue systems, dialogue systems under low-resource settings, dialogue evaluations, and multimodal dialogues, etc).
· Co-designed (with Prof. Zhou Yu) the curriculum; co-led in-class discussions; co-supervised course projects; co-hosted guest speakers; held office hours and graded assignments.

***Guest Lecturer*: COMS 4705, Natural Language Processing**, Columbia Spring 2022
· *Topic*: dialogue systems (spoken and text-based)
· Prepared and delivered lectures to a 100-student upper-division undergraduate introduction course to NLP.

***Teaching Assistant*: COMS 4156, Advanced Software Engineering**, Columbia Fall 2021
· *Topics*: modern software engineering skills and practices
· Supervised course projects; held office hours and graded assignments

***Teaching Assistant*: COMS 6998, Conversational AI**, Columbia Spring 2021
· *Topics*: various topics on dialogue systems
· Co-designed (with Prof. Zhou Yu) the curriculum; co-supervised course projects; held office hours and graded assignments.

SERVICE

Area Chair: EMNLP 2023

Organization Committee: Social Influence in Conversations, ACL 2023

Area Chair: 17th Women in Machine Learning Workshop (WiML 2022)

Publicity Chair: 4th Workshop for Conversational AI, ACL 2022

Program Commitee/Conference Reviewer: ACL 2019, ACL 2020, *SEM 2020, ICLR 2021, AAAI 2021, EACL 2021, NAACL 2021, ACL 2021, *SEM 2021, NeurIPS 2021, EMNLP 2021, NLPCC 2021, ICLR 2022, AAAI 2022, ICML 2022, ACL 2022, CHI 2022, SIGDIAL 2022, HCI+NLP workshop at NAACL 2022, ACL Rolling Review 2021-2022

Journal Reviewer: ACM Transactions on Human-Robot Interaction, Neurocomputing, ACM Transactions on Information Systems, Nature Human Behaviour

CS Graduate Admission Committee: Columbia University, 2022