

Hanjie Chen

📍 Duncan Hall 2081, 6100 Main St MS 364, Houston, TX 77005

✉ hanjie@rice.edu 🏠 <https://hanjiechen.github.io/> 🌐 HanjieChen 🐦 @hanjie_chen

📄 <https://scholar.google.com/citations?user=DyY0gLwAAAAJ>

RESEARCH INTEREST

Natural Language Processing (NLP), Interpretable Machine Learning, Trustworthy AI

CURRENT POSITION

Assistant Professor (tenure-track) at Rice University *Jul. 2024 - Present*
Department of Computer Science
George R. Brown School of Engineering and Computing
Member of The Ken Kennedy Institute

EDUCATION

The University of Virginia (UVA) *Aug. 2018 - May 2023*
Ph.D. in Computer Science
Thesis: Neural Model Interpretability for Natural Language Processing
Advisor: Yangfeng Ji

University of Science and Technology of China (USTC) *Sept. 2015 - Jun. 2018*
M.S. in Information and Communication Engineering

Nanjing University of Aeronautics and Astronautics (NUAA) *Sept. 2011 - Jun. 2015*
B.S. in Information Engineering

ACADEMIC EXPERIENCE

Postdoctoral Fellow at Johns Hopkins University *Jun. 2023 - Jun. 2024*
Center for Language and Speech Processing
Advisor: Mark Dredze

Research Assistant at The University of Virginia *Aug. 2018 - May 2023*
Department of Computer Science
Advisor: Yangfeng Ji

Research Intern at Allen Institute for AI (AI2), Seattle, WA USA *May 2022 - Oct. 2022*
Mosaic Group
Manager: Yejin Choi
Mentors: Swabha Swayamdipta, Faeze Brahma

Research Intern at Microsoft Research, Redmond, WA USA *May 2021 - Aug. 2021*
Language and Information Technologies Group
Manager: Ahmed H. Awadallah
Mentors: Guoqing Zheng, Srinagesh Sharma

Research Intern at IBM Research, New York, NY USA *Jun. 2020 - Aug. 2020*
Thomas J. Watson Research Center
Manager: Luis Lastras
Mentors: Chulaka Gunasekara, Song Feng, Hui Wan, Jatin Ganhotra, Sachindra Joshi

PUBLICATIONS

*equal contribution

1. **When Embedding-Based Defenses Fail: Rethinking Safety in LLM-Based Multi-Agent Systems**
Lingxi Zhang, Guangtao Zheng, **Hanjie Chen**
The Forty-Third International Conference on Machine Learning (ICML), Seoul, South Korea, Jul. 2026
2. **Spherical Steering: Geometry-Aware Activation Rotation for Language Models**
Zejia You, Chunyuan Deng, **Hanjie Chen**
The Forty-Third International Conference on Machine Learning (ICML), Seoul, South Korea, Jul. 2026
3. **SportR: A Benchmark for Multimodal Large Language Model Reasoning in Sports**
Haotian Xia*, Haonan Ge*, Junbo Zou*, Hyun Woo Choi, Xuebin Zhang, Danny Suradja, Botao Rui, Ethan Tran, Wendy Jin, Zhen Ye, Xiyang Lin, Christopher Lai, Shengjie Zhang, Junwen Miao, Shichao Chen, Rhys Tracy, Vicente Ordonez, Weining Shen, **Hanjie Chen**
The Fourteenth International Conference on Learning Representations (ICLR), Rio de Janeiro, Brazil, Apr. 2026
4. **Personality Structured Interview for Large Language Model Simulation in Personality Research**
Pengda Wang, Huiqi Zou, Han Jiang, **Hanjie Chen**, Tianjun Sun, Xiaoyuan Yi, Ziang Xiao, Frederick L. Oswald
The 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Rabat, Morocco, Mar. 2026
5. **Language Models are Symbolic Learners in Arithmetic**
Chunyuan Deng, Zhiqi Li, Roy Xie, Ruidi Chang, **Hanjie Chen**
Transactions on Machine Learning Research (TMLR), Jan. 2026
6. **Steering Information Utility in Key-Value Memory for Language Model Post-Training**
Chunyuan Deng, Ruidi Chang, **Hanjie Chen**
The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS), San Diego, US, Dec. 2025
7. **Findings of the BlackboxNLP 2025 Shared Task: Localizing Circuits and Causal Variables in Language Models**
Dana Arad, Yonatan Belinkov, **Hanjie Chen**, Najoung Kim, Hosein Mohebbi, Aaron Mueller, Gabriele Sarti, Martin Tutek
EMNLP BlackboxNLP Workshop, Suzhou, China, Nov. 2025
8. **Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models**
Yang Sui, Yu-Neng Chuang, Guanclu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, **Hanjie Chen**, Xia Hu
Transactions on Machine Learning Research (TMLR), Jul. 2025
9. **Learning Distribution-Wise Control in Representation Space for Language Models**
Chunyuan Deng, Ruidi Chang, **Hanjie Chen**
The Forty-Second International Conference on Machine Learning (ICML), Vancouver, Canada, Jul. 2025
10. **Rethinking Diverse Human Preference Learning through Principal Component Analysis**
Feng Luo, Rui Yang, Hao Sun, Chunyuan Deng, Jiarui Yao, Jingyan Shen, Huan Zhang, **Hanjie Chen**
Findings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vienna, Austria, Jul. 2025
11. **SAFR: Neuron Redistribution for Interpretability**
Ruidi Chang, Chunyuan Deng, **Hanjie Chen**
Findings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL), Albuquerque, New Mexico, United States, May 2025
12. **MiCEval: Unveiling Multimodal Chain of Thought’s Quality via Image Description and Reasoning Steps**
Xiongtao Zhou*, Jie He*, Lanyu Chen, Jingyu Li, Haojing Chen, Victor Gutierrez Basulto, Jeff Z. Pan, **Hanjie Chen**

2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL), Albuquerque, New Mexico, United States, May 2025

13. **Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions**
Hanjie Chen*, Zhouxiang Fang*, Yash Singla, Mark Dredze
2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL), Albuquerque, New Mexico, United States, May 2025
14. **SPORTU: A Comprehensive Sports Understanding Benchmark for Multimodal Large Language Models**
Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, Yuan-fang Wang, Weining Shen, **Hanjie Chen**
The Thirteenth International Conference on Learning Representations (ICLR), Singapore, Apr. 2025
15. **Will the Real Linda Please Stand up...to Large Language Models? Examining the Representativeness Heuristic in LLMs**
Pengda Wang*, Zilin Xiao*, **Hanjie Chen**, Frederick L. Oswald
The First Conference on Language Modeling (COLM), Philadelphia, Pennsylvania, United States, Oct. 2024
Oral Spotlight, top 2%
16. **RORA: Robust Free-Text Rationale Evaluation**
Zhengping Jiang*, Yining Lu*, **Hanjie Chen**, Daniel Khashabi, Benjamin Van Durme, Anqi Liu
The 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Bangkok, Thailand, Aug. 2024
17. **Explanation in the Era of Large Language Models**
Zining Zhu, **Hanjie Chen**, Xi Ye, Chenhao Tan, Ana Marasović, Sarah Wiergreffe, Veronica Qing Lyu
2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) Tutorial, Mexico City, Mexico, Jun. 2024
18. **Explainability for Large Language Models: A Survey**
Haiyan Zhao, **Hanjie Chen**, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Mengnan Du
ACM Transactions on Intelligent Systems and Technology, Jan. 2024
19. **REV: Information-Theoretic Evaluation of Free-Text Rationales**
Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta
The 61st Annual Meeting of the Association for Computational Linguistics (ACL), Toronto, Canada, Jul. 2023
20. **KNIFE: Knowledge Distillation with Free-Text Rationales**
Aaron Chan*, Zhiyuan Zeng*, Wyatt Lake, Brihi Joshi, **Hanjie Chen**, Xiang Ren
TrustML-(un)Limited Workshop at ICLR 2023, Kigali, Rwanda, May 2023
21. **Improving Interpretability via Explicit Word Interaction Graph Layer**
Arshdeep Sekhon, **Hanjie Chen**, Aman Shrivastava, Zhe Wang, Yangfeng Ji, and Yanjun Qi
The 37th AAAI Conference on Artificial Intelligence (AAAI), Washington, DC, USA, Feb. 2023
22. **Explaining Predictive Uncertainty by Looking Back at Model Explanations**
Hanjie Chen, Wanyu Du and Yangfeng Ji
AAAI Workshop on Uncertainty Reasoning and Quantification in Decision Making, Washington, DC, USA, Feb. 2023
23. **Identifying the Source of Vulnerability in Explanation Discrepancy: A Case Study in Neural Text Classification**
Ruixuan Tang, **Hanjie Chen**, and Yangfeng Ji
EMNLP BlackboxNLP Workshop, Abu Dhabi, Dec. 2022
24. **Self-training with Two-phase Self-augmentation for Few-shot Dialogue Generation**
Wanyu Du, **Hanjie Chen**, and Yangfeng Ji
Findings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, Dec. 2022

25. Pathologies of Pre-trained Language Models in Few-shot Fine-tuning
Hanjie Chen, Guoqing Zheng, Ahmed Hassan Awadallah, and Yangfeng Ji
ACL Workshop on Insights from Negative Results in NLP, Dublin, Ireland, May 2022
26. Adversarial Training for Improving Model Robustness? Look at Both Prediction and Interpretation
Hanjie Chen and Yangfeng Ji
The 36th AAAI Conference on Artificial Intelligence (AAAI), Vancouver, BC, Canada, Feb. 2022
Oral Presentation, top 1.5%
27. Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing
Sanchit Sinha, **Hanjie Chen**, Arshdeep Sekhon, Yangfeng Ji, Yanjun Qi
EMNLP BlackboxNLP Workshop, Punta Cana, Dominican Republic, Nov. 2021
28. Explaining Neural Network Predictions on Sentence Pairs via Learning Word-Group Masks
Hanjie Chen, Song Feng, Jatin Ganhotra, Hui Wan, Chulaka Gunasekara, Sachindra Joshi, Yangfeng Ji
2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Jun. 2021
29. Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers
Hanjie Chen and Yangfeng Ji
The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Nov. 2020
30. Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection
Hanjie Chen, Guangtao Zheng and Yangfeng Ji
The 58th Annual Meeting of the Association for Computational Linguistics (ACL), July 2020
31. Improving the Explainability of Neural Sentiment Classifiers via Data Augmentation
Hanjie Chen and Yangfeng Ji
NeurIPS Workshop on Robust AI in Financial Services, Vancouver, Canada, Dec. 2019

Preprint

32. Agentic Discovery with Active Hypothesis Exploration for Visual Recognition
Jaywon Koo, Jefferson Hernandez, Ruozhen He, **Hanjie Chen**, Chen Wei, Vicente Ordonez
arXiv preprint arXiv:2604.12999
33. PRISM: A Dual View of LLM Reasoning through Semantic Flow and Latent Computation
Ruidi Chang, Jiawei Zhou, **Hanjie Chen**
arXiv preprint arXiv:2603.22754
34. GR-SAP: Generative Replay for Safety Alignment Preservation during Fine-Tuning
Zhouxiang Fang, Jiawei Zhou, **Hanjie Chen**
arXiv preprint arXiv:2603.10243
35. Spherical Beyond Explainable AI (XAI): An Overdue Paradigm Shift and Post-XAI Research Directions
Saleh Afroogh, Seyd Ishtiaque Ahmed, Petra Ahrweiler, David Alvarez-Melis, Mansur Maturidi Arief, Emilia Barakova, Falco J. Bargagli-Stoffi, Erdem Biyik, **Hanjie Chen**, Xiang 'Anthony' Chen, Robert Clements, Keeley Crockett, Amit Dhurandhar, Fethiye Irmak Dogan, Mollie Dollinger, Motahhare Eslami, Aldo A Faisal, Arya Farahi, Melanie Fernandez Pradie, Saadia Gabriele, Diego Garcia-Olano, Marzyeh Ghassemi, Shaona Ghosh, Hatice Gunes, Ehsan Hajiramezanali, Stefan Haufe, Biwei Huang, Angel Hwang, Md Tauhidul Islam, Junfeng Jiao, Amir-Hossein Karimi, Saber Kazeminasab, Anastasia Kuzminykh, William La Cava, Brian Y. Lim, Xiaofeng Liu, Mohammad R. K. Mofrad, Alicia Parrish, Maria Perez-Ortiz, Shriti Raj, Swabha Swayamdipta, Salmon Talebi, Kush R. Varshney, Mihaela Vorvoreanu, Lily Weng, Alice Xiang, Yiming Xu, Ding Zhao, Jieyu Zhao
arXiv preprint arXiv:2602.24176
36. DeepSport: A Multimodal Large Language Model for Comprehensive Sports Video Reasoning via Agentic Reinforcement Learning
Junbo Zou*, Haotian Xia*, Zhen Ye, Shengjie Zhang, Christopher Lai, Vicente Ordonez, Weining Shen, **Hanjie Chen**
arXiv preprint arXiv:2511.12908

37. **The Generalization Ridge: Information Flow in Natural Language Generation**
Ruidi Chang, Chunyuan Deng, **Hanjie Chen**
arXiv preprint arXiv:2507.05387
38. **Political-LLM: Large Language Models in Political Science**
Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui...**Hanjie Chen**...Yue Zhao, Yushun Dong
arXiv preprint arXiv:2412.06864
39. **From Babbling to Fluency: Evaluating the Evolution of Language Models in Terms of Human Language Acquisition**
Qiyuan Yang*, Pengda Wang*, Luke D. Plonsky, Frederick L. Oswald, **Hanjie Chen**
arXiv preprint arXiv:2410.13259
40. **Language and Multimodal Models in Sports: A Survey of Datasets and Applications**
Haotian Xia, Zhengbang Yang, Yun Zhao, Yuqing Wang, Jingxi Li, Rhys Tracy, Zhuangdi Zhu, Yuan-fang Wang, **Hanjie Chen**, Weining Shen
arXiv preprint arXiv:2406.12252

Previous Publications

41. **A Two-Dimensional Constellation Design Method for Visible Light Communications with Signal-Dependent Shot Noise**
Hanjie Chen and Zhengyuan Xu
IEEE Communications Letters, vol. 22, no. 9, pp. 1786-1789, Sept. 2018
Impact Factor: 3.457
42. **The Near-Field Radiation Pattern of an OLED Panel and Its Application in Detection**
Rui Xu, **Hanjie Chen**, and Zhengyuan Xu
The 11th International Symposium on Communication Systems, Networks, and Digital Signal Processing (CSNDSP), Budapest, Hungary, July 18-20, 2018
43. **OLED Panel Radiation Pattern and Its Impact on VLC Channel Characteristics**
Hanjie Chen and Zhengyuan Xu
IEEE Photonics Journal, vol. 10, no. 2, pp. 1-10, April 2018
Impact Factor: 3.03
44. **A 51.6 Mbps Experimental VLC System Using a Monochromic Organic LED**
Hanjie Chen, Zhengyuan Xu, Qian Gao, and Shangbin Li
IEEE Photonics Journal, vol. 10, no. 2, pp. 1-12, April 2018
Impact Factor: 3.03
45. **Radiation Pattern Modeling of a Bent OLED Panel for Visible Light Communication**
Hanjie Chen and Zhengyuan Xu
Asia Communications and Photonics Conference (ACP), Guangzhou, China, November 10 – 13, 2017
46. **Volterra-Based Nonlinear Equalization for Nonlinearity Mitigation in Organic VLC**
Xiangyu Li, **Hanjie Chen**, Shangbin Li, Qian Gao, Chen Gong, and Zhengyuan Xu
The 13th International Wireless Communications & Mobile Computing Conference (IWCMC), Valencia, Spain, June 26-30, 2017
47. **A 1.9 Mbps OFDM-Based All-Organic Visible Light Communication System**
Hanjie Chen, Shangbin Li, Boyang Huang, Zhengyuan Xu, Wenhai Li, Guifang Dong, and Jing Xie
The 15th IEEE International Conference on Communication Systems (ICCS), Shenzhen, China, December 14 - 16, 2016
48. **Squarylium and Rubrene Based Filterless Narrowband Photodetectors for an All-Organic Two-Channel Visible Light Communication System**
Wenhai Li, Shangbin Li, Lian Duan, **Hanjie Chen**, Liduo Wang, Guifang Dong, and Zhengyuan Xu
Organic Electronics, vol. 37, pp. 346-351, Oct. 2016
Impact Factor: 3.50

RESEARCH GRANTS AND FUNDING

- **Open Philanthropy Technical AI Safety Research**, “Open Alignment Faking Behavior”
Co-PI. PI: Jiawei Zhou
Total funding: \$549,384 *Jun. 2026 - Jun. 2028*
- **NVIDIA Academic Grant Program Award**, “PathLoop: A Looped MoE Architecture via Internal Path Group Optimization”
PI. Award: 32K A100 GPU-Hours *Apr. 2026 - Sept. 2026*
- **Lambda’s Research Grant Program Award**,
PI. Award: \$5,000 *Mar. 2026*
- **Rice Creative Ventures Fund: Conference and Workshop Development**, “LLM Frontiers: Advancement Meets Responsibility”
PI. Co-PI: Vicente Ordonez
Total funding: \$10,000 *Jan. 2026 - Dec. 2026*
- **NIH OT2**, “EMED: An Ethical Mixture-of-Experts Digital Twin Framework for Medical Device Surveillance”
Other PI. Contact PI: Hongfang Liu; Co-Is: Tianlong Chen, Kaixiong Zhou, Liwei Wang, Assaf Gottlieb, Cheryl Brown
Total funding: \$1,686,929 *Sept. 2025 - Sept. 2026*
- **NSF ReDDDoT Phase 2**, “Responsible Multi-Modal AI Systems for Multi-Hazard Resilience and Situational Awareness”
Co-PI. PI: Jamie E Padgett; Co-PIs: David P Retchless, Avantika Gori.
Total funding: \$1,500,000 *Sept. 2025 - Sept. 2027*
- **Ken Kennedy Institute Research Cluster Initiative Award**, “Human-centered Artificial Physical Intelligence”
Co-I. PI: Vaibhav Unhelkar; Co-Is: Jing Chen, Moshe Vardi
Total funding: \$80,000 *Jul. 2025 - Jun. 2026*
- **Rice CS Postdoctoral Fellowship**
MPI. MPI: Xia Hu
Total funding: \$60,000 *Oct. 2024 - Sept. 2025*
- **Ken Kennedy Institute Research Aspiring Cluster Initiative Award**, “Human-AI Collaboration”
Co-I. PI: Vaibhav Unhelkar; Co-Is: Jing Chen, Moshe Vardi
Total funding: \$2,000 *Jul. 2024 - Jun. 2025*

HONORS AND AWARDS

- NVIDIA Academic Grant Program Award *2026*
- Lambda’s Research Grant Program Award *2026*
- TAMEST (Texas Academy of Medicine, Engineering, Science and Technology) 2026 Protégé *2026*
- Outstanding Doctoral Student Award, UVA
(One of the five awardees in the School of Engineering & Applied Science) *2023*
- John A. Stankovic Graduate Research Award, UVA *2023*
- Carlos and Esther Farrar Fellowship Award, UVA *2022 - 2023*

- WiML Travel Funding 2022
- University-wide Graduate Teaching Awards Nominee
(**Top 5%** of graduate instructors at UVA) 2022
- UVA CS Outstanding Graduate Teaching Award 2022
- UVA Engineering Graduate Teaching Fellow 2022
- Best Poster Award at the ACM Capital Region Celebration of Women
in Computing (CAPWIC) 2021
- 2021 National Center for Women & Information Technology (NCWIT)
Collegiate Award Finalist 2021
- WiML Travel Funding 2021
- CAPWIC scholarship 2021
- CRA-WP Grad Cohort for Women Travel Funding 2020
- IBM Ph.D. Fellowship Award Nomination 2020
- Microsoft Research Ada Lovelace Fellowship Nomination 2019
- UVA Computer Science Fellowship 2018 - 2019
- Outstanding Graduates in Anhui Province, China (**Top 3%**) 2018
- Outstanding Graduates Awards, USTC (**Top 3%**) 2018
- National Scholarship for Graduate Students, China Ministry of Education (**Top 3%**) 2017
- National Graduate Mathematical Contest in Modeling, Third Prize, China 2016
- Outstanding Student Scholarship, First Prize, USTC 2015 - 2018
- Outstanding Graduates Awards, NUAA (**Top 1%**) 2015
- CATIC Special Scholarship, NUAA (**1/290** \approx **0.34** %) 2014
- Annual Special Award, NUAA (**Top 1%**) 2014
- National Undergraduate Electronic Design Contest, Second Prize, Jiangsu, China 2014
- National Scholarship, China Ministry of Education (**Top 1%**) 2013
- College Physics and Experimental Technology Contest, Third Prize, Jiangsu, China 2013
- Outstanding Student Scholarship, First Prize, NUAA 2012 - 2015

TEACHING EXPERIENCE

- **Instructor**
 - COMP 484/584 001: Natural Language Processing *Spring 2026, Rice*
 - COMP 640 003: Graduate Seminar in Machine Learning *Fall 2025, Rice*
 - COMP 652 001: Natural Language Processing *Spring 2025, Rice*
 - COMP 677 Explainable Natural Language Processing *Fall 2024, Rice*
 - EN.601.867 Trustworthy and Responsible NLP *Spring 2024, JHU*

- CS 6501/4501 Interpretable Machine Learning *Spring 2022, UVA*
 - ★ UVA CS Outstanding Graduate Teaching Award
 - ★ University-wide Graduate Teaching Awards Nominee (top 5% of graduate instructors)
- **Guest Lecturer**
 - Ken Kennedy Institute AI and Machine Learning Boot Camp: *Natural Language Processing* *May 2025, Rice*
 - 601.467/667 Introduction to Human Language Technology: *Interpretable and Explainable NLP* *Nov. 2023, JHU*
 - CSCI 699 Ethics in NLP: *Interpretable and Explainable NLP* *Nov. 2023, USC*
 - CS 4501 Machine Learning for NLP: *Interpretability of NLP Models* *Nov. 2021, UVA*
 - CS 6501 Natural Language Processing: *Interpretability of NLP Models* *May 2021, UVA*
 - CS 4501 Machine Learning for NLP: *Model Interpretability for NLP* *Nov. 2020, UVA*
- **Teaching Assistant**
 - CS 4750 Database Systems *Spring 2021, UVA*
 - CS 5010 Programming and Systems for Data Analysis *Fall 2020, UVA*
 - CS 6316 Machine Learning *Spring 2020, UVA*
 - CS 6501 Natural Language Processing *Fall 2019, UVA*
 - INY5101 Matrix Analysis and Its Applications *Spring 2017, USTC*

PROFESSIONAL SERVICES

- **Organizer** for 2026 Rice Workshop on Large Language Models, BlackboxNLP Workshop @ EMNLP 2024 - 2026, The First Workshop on the Application of LLM Explainability to Reasoning and Planning @ COLM 2025
- **Senior Area Chair** for EMNLP 2026, ACL 2025
- **Publication Co-Chair** for COLING 2027
- **Area Chair**
 - NAACL 2025
 - COLING 2025
 - EMNLP 2024
 - ACL ARR 2024 - Present
 - WiML Workshop @ NeurIPS 2022
- **Program Committee**
 - NAACL-HLT/ACL/EMNLP Tutorial 2025
 - ACL 2023
 - AAAI 2023

- EMNLP 2021 - 2023
- NAACL 2021
- EACL 2023
- CoNLL 2021 - 2022
- NLPCC 2022
- ACL DialDoc Workshop 2022
- EMNLP BlackboxNLP Workshop 2021, 2023
- NeurIPS Explainable AI Approaches for Debugging and Diagnosis Workshop 2021
- Document-grounded Dialogue Workshop 2021
- MASC-SLL 2020
- **Reviewer**
 - TACL 2023 - Present
 - ICLR 2024 - 2026
 - ICML 2026
 - CVPR 2026
 - ECCV 2026
 - COLM 2024 - 2026
 - NeurIPS 2023, 2025 - 2026
 - EMNLP 2023
 - ACL ARR 2021 - Present
 - ACL 2020 - 2021
 - EMNLP BlackboxNLP Workshop 2022
 - CoNLL 2019 - 2020
 - NLPCC 2019 - 2021
- **Panel Reviewer** for NSF CISE/IIS III Core Programs, May 2024, April 2025
- **Mentor** for the ACL 2026 Student Research Workshop, the Summer Undergraduate Research Fellowship (SURF) Program at Rice University, 2025 - 2026
- **Committee Member** for PhD Admission Committee (2024 - Present), Faculty Hiring Committee (2024 - Present), Ken Kennedy Institute Fellowship Review Committee (2024, 2025), AI Major Curriculum Committee (2025), AI4CSEd Committee (2025 - 2026), MCS Application Review Committee (2026) at Rice University
- **Diversity Representative** for UVA Computer Science Graduate Student Group (CSGSG) Council, 2022

TALKS AND ACTIVITIES

- Invited Talk on *The Unsolved Problems of LLMs in Healthcare: Explainability and Safety* @ University of Minnesota Apr. 2026
- Presentation on *Toward Trustworthy and Responsible AI: Large Language Model Interpretability and Control* at the TAMEST 2026 Annual Conference (Protégé Program) Feb. 2026
- Invited Talk on *Toward Interpretable and Controllable Language Models* at the CDS Department Colloquium @ Case Western Reserve University Oct. 2025
- Invited Talk on *Interpretability and Control of Large Language Models* at the LLMs and The Brain @ Rice University Sept. 2025
- Invited Talk on *Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions* at the AI in Health Conference 2025 Sept. 2025
- Invited Talk on *Interpretability and Control of Large Language Models* at the Data-Enabled Science Seminar @ University of Houston Sept. 2025
- Invited Talk on *Large Language Model Interpretability and Application in Medicine* at the Health Atmosphere Nexus Group (HANG) @ UTHealth Houston Aug. 2025
- Lecture on *Natural Language Processing* at the Ken Kennedy Institute AI and Machine Learning Boot Camp, Houston, TX May 2025
- Invited Talk on *Explainable AI in the Era of Large Language Models* at the 2025 SIOP Annual Conference, Denver, Colorado Apr. 2025
- Invited Talk on *Explanation Generation and Evaluation for (Multimodal) Large Language Models* at the Ethical and Explainable GeoAI Workshop @ Texas A&M University Feb. 2025
- Panel Discussion on *Large Language Models, DeepSeek, and the Future of Generative AI* at Rice University Feb. 2025
- Invited Talk on *From Passion to Profession: My Journey as a Woman in STEM* at the Girls Who Code Club @ Bellaire High School, Houston, TX Dec. 2024
- Invited Talk on *Incredible Yet Limited Large Language Models in the Wild* at the Department of Computer Science @ Florida State University Nov. 2024
- Tutorial on *Explanation in the Era of Large Language Models* at NAACL 2024 Jun. 2024
- Invited Talk on *Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions* at How Sustainable is Artificial Intelligence? Mar. 2024
- Presentation on *REV: Information-Theoretic Evaluation of Free-Text Rationales* at ACL 2023 Jul. 2023
- Invited Talk on *Bridging the Trustworthy Gap between AI and Humans: Interpretation Techniques for Modern NLP* at the CLSP Seminar @ Johns Hopkins University Mar. 2023
- Invited Talk at Northeastern University Mar. 2023
- Invited Talk at Florida State University Mar. 2023
- Invited Talk at Indiana University Bloomington Mar. 2023
- Invited Talk at The University of Iowa Feb. 2023
- Invited Talk at The University of Arizona Feb. 2023
- Invited Talk at Rensselaer Polytechnic Institute Feb. 2023

- Invited Talk at The College of William & Mary *Feb. 2023*
- Invited Talk at Rice University *Jan. 2023*
- Presentation on *Information-Theoretic Evaluation of Free-Text Rationales with Conditional V-Information* at Trustworthy and Socially Responsible Machine Learning (TSRML) Workshop @ NeurIPS *Dec. 2022*
- Presentation on *Explaining Predictive Uncertainty by Looking Back at Model Explanations* at WiML Workshop 2022 @ NeurIPS *Nov. 2022*
- Talk on *REV: Information-Theoretic Evaluation of Free-Text Rationales* at the Allen Institute for AI (AI2) *Oct. 2022*
- Presentation on *Pathologies of Pre-trained Language Models in Few-shot Fine-tuning* at Insights Workshop @ ACL 2022 *May 2022*
- Presentation on *Adversarial Training for Improving Model Robustness? Look at Both Prediction and Interpretation* at AAAI 2022 *Feb. 2022*
- Completed the c3Design Program hosted and facilitated by UVA's Course Design Institute offered by the Center for Teaching Excellence *Jan. 2022*
- Invited talk on *Improving Model Robustness via Interpretation-based Adversarial Training @ MLNLP* *Dec. 2021*
- Presentation on *Adversarial Training for Improving Model Robustness? Look at Both Prediction and Interpretation* at WiML Workshop 2021 @ NeurIPS *Dec. 2021*
- Presentation on *Adversarial Training for Improving Model Robustness? Look at Both Prediction and Interpretation* at UVA CS Department Research Symposium *Dec. 2021*
- Paper presentation on *Explaining Neural Network Predictions on Sentence Pairs via Learning Word-Group Masks* at NAACL 2021 *Jun. 2021*
- 2021 CRA-WP Grad Cohort for Women Workshop *Apr. 2021*
- Poster presentation at the ACM Capital Region Celebration of Women in Computing (CAPWIC) *Mar. 2021*
- Paper presentation on *Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers* at EMNLP 2020 *Nov. 2020*
- Presentation on *Learning Variational Masks for Explainable Next Utterance Prediction in Dialog Systems* at IBM Research *Aug. 2020*
- Paper presentation on *Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection* at ACL 2020 *Jul. 2020*
- 2020 CRA-WP Grad Cohort for Women Workshop (postponed to 2021 due to the COVID-19) *Apr. 2020*
- Poster presentation on *Improving the Explainability of Neural Sentiment Classifiers via Data Augmentation* at NeurIPS 2019 Workshop on Robust AI in Financial Services *Dec. 2019*
- Poster presentation on *Building Hierarchical Interpretations in Natural Language via Feature Interaction Detection* at UVA CS Department Research Symposium *Oct. 2019*
- Invited talk on *How to Train a More Interpretable Neural Text Classifier?* at UVA AIML Seminar *Apr. 2019*

- Poster presentation on *An Empirical Comparison on Convolutional and Recurrent Neural Networks for NLP* at the JUMP Undergraduate Research Initiative, UVA *Nov. 2018*

MEMBERSHIPS

- Association for Computing Machinery (ACM), Member *2024 - Present*
- Association for Computational Linguistics (ACL), Member *2020 - Present*