
RUIXIANG (RYAN) TANG

CONTACT INFORMATION

Assistant Professor, Department of Computer Science, Rutgers University

HOME PAGE: <https://www.ruixiangtang.net/>

PHONE: (+1)(682)266-9544 EMAIL: ruixiang.tang@rutgers.edu

RESEARCH INTERESTS

My research focuses on developing algorithms and systems that enhance trustworthiness throughout the AI lifecycle. I pursue this goal by formalizing key trustworthiness objectives, including robustness, privacy, interpretability, and safety, and by developing principled algorithms with measurable guarantees and transparent failure modes. I also collaborate closely with biologists and health informaticians to translate trustworthy learning methods to high-impact problems in biology and medicine.

ACADEMIC EMPLOYMENTS

Rutgers University–New Brunswick, New Jersey, USA

September 2024 - Present

Assistant Professor

Department of Computer Science

EDUCATION

Rice University, Houston, USA

August 2021 - August 2024

Ph.D. in Computer Science

Department of Computer Science

Advisor: Dr. Xia (Ben) Hu

Texas A&M University, College Station, USA

August 2019 - August 2021

Ph.D. Student in Computer Science (Transfer Out)

Department of Computer Science and Engineering

Advisor: Dr. Xia (Ben) Hu

Tsinghua University, Beijing, China

September 2014 - June 2019

Bachelor of Science in Automation

Department of Automation

Advisor: Dr. Jiwen Lu

SELECTED HONORS, AWARDS, & FELLOWSHIPS

- 2026 IEEE International Conference on Multimedia and Expo Best Paper Candidate
- AMD University Program: AI & HPC Cluster Allocation Award 2026
- OpenAI Researcher Access Program 2025
- Hershel M. Rich Invention Award 2024
- AMIA 2023 Best Student Paper Award
- CIKM 2023 Best Demo Paper Honorable Mention Award
- AMIA 2022 Best Student Paper (Shortlist) Award
- Microsoft Research Asia Rising Star Intern Award
- Outstanding Freshman Scholarship, Tsinghua University (Top 10%)
- The Silver Medal of IGEM (International Genetically Engineered Machine)
- China High School Biology Olympiad (Ranked Seventh Nationally)

PUBLICATIONS

70+ papers have been published on Top AI Conferences/Journals

AMIA Best Student Paper

CIKM Best Demo Paper Honorable Mention

Communication of ACM Cover Paper

5,000+ citations, h-index 27 (Google Scholar).

* denotes an equal contribution.

[EMNLP' 26] Jiaqian Li, Yanshu Li, Boxuan Zhang, **Ruixiang Tang**, Kuan-Hao Huang. "TRACES: Proactive Safety Auditing for Multi-Turn LLM Agents via Trajectory-State Modeling." In *Proceedings of Empirical Methods in Natural Language Processing, 2026*.

[ECCV' 26] Yanshu Li, Jiaqian Li, Kuai Yu, Xi Xiao, Dongfang Liu, Tianyang Wang, **Ruixiang Tang**. "Personalize Your Large Vision-language Models With In-context Prompt Tuning." In *European Conference on Computer Vision*

[ECCV' 26] Feng He, Zhenting Wang, Qifan Wang, Qiang Guan, Dongfang Liu, **Ruixiang Tang**, Qiankun Li. "HIVE: Understanding Post Hallucination Reasoning in Vision Language Models" In *European Conference on Computer Vision*

[ECCV' 26] Liwei Che, Zhiyu Xue, Yihao Quan, Benlin Liu, Zeru Shi, **Ruixiang Tang**, Ranjay Krishna, Vladimir Pavlovic "Counting Circuits: Mechanistic Interpretability of Visual Reasoning in Large Vision-Language Models" In *European Conference on Computer Vision*

[ICML' 26] Zeru Shi, Zhenting Wang, Fan Yang, Qifan Wang, **Ruixiang Tang**. "A Single Layer to Explain Them All: Understanding Massive Activations in Large Language Models." In *International Conference on Machine Learning*.

[ICML' 26] Jiaqian Li, Yanshu Li, Ligong Han, **Ruixiang Tang**, Wenya Wang. "Towards Generalizable Implicit In-Context Learning with Attention Routing." In *International Conference on Machine Learning*.

[ACL' 26] Can Jin*, Rui Wu*, Tong Che*, Qixin Zhang, Hongwu Peng, Jiahui Zhao, Zhenting Wang, Wenqi Wei, Ligong Han, Zhao Zhang, Yuan Cao, **Ruixiang Tang**, Dimitris N Metaxas. "Reasoning over Precedents Alongside Statutes: Case-Augmented Deliberative Alignment for LLM Safety." In *Annual Meeting of the Association for Computational Linguistics*.

[CVPR' 26] Yuefei Chen, Jiang Liu, Xiaodong Lin, **Ruixiang Tang**. "Distilling Counterfactual Reasoning from Language to Vision: Causal Graph-Guided Post-Training for Video Understanding." In *Findings of Computer Vision and Pattern Recognition Conference*.

[ICME' 26] Yanchuan Tang*, Taowen Wang*, Yuefei Chen, Boxuan Zhang, Qiang Guan, **Ruixiang Tang**. "Shifting Uncertainty to Critical Moments: Towards Reliable Uncertainty Quantification for VLA Model." In *IEEE International Conference on Multimedia and Expo. (Best Paper Candidate)*

[AAAI' 26] Yanshu Li, Jianjiang Yang, Ziteng Yang, Bozheng Li, Hongyang He, Zhengtao Yao, Ligong Han, Yingjie Victor Chen, Songlin Fei, Dongfang Liu, **Ruixiang Tang**. "Make LVLMS Focus: Context-Aware Attention Modulation for Better Multimodal In-Context Learning." In *AAAI Conference on Artificial Intelligence, 2026 Oral Presentation*

[AAAI' 26] Yanshu Li, Jianjiang Yang, Zhennan Shen, Ligong Han, Haoyan Xu, **Ruixiang Tang**. "Catp: Contextually adaptive token pruning for efficient and enhanced multimodal in-context learning." In *AAAI Conference on Artificial Intelligence, 2026 Oral Presentation*

-
- [AAAI’ 26] Yuefei Chen, Vivek K. Singh, Jing Ma, **Ruixiang Tang**. “CounterBench: Evaluating and Improving Counterfactual Reasoning in Large Language Models.” In *AAAI Conference on Artificial Intelligence, 2026*
- [ICLR’ 26] Runjia Zeng, Qifan Wang, Qiang Guan, **Ruixiang Tang**, Lifu Huang, Zhenting Wang, Xueling Zhang, Cheng Han, Dongfang Liu “TokenSeek: Memory Efficient Fine Tuning via Instance-Aware Token Ditching.” In *International Conference on Learning Representations, 2026*
- [EACL’ 26] Yu-Neng Chuang*, Guanchu Wang*, Chia-Yuan Chang, **Ruixiang Tang**, Shaochen Zhong, Fan Yang, Andrew Wen, Mengnan Du, Xuanning Cai, Vladimir Braverman, Xia Hu “From Plausible to Faithful: Optimizing the Faithfulness of LLM Explanations.” In *European Chapter of the Association for Computational Linguistics, 2026*
- [EMNLP’ 25] Yanshu Li, Tian Yun, Jianjiang Yang, Pinyuan Feng, Jinfa Huang, **Ruixiang Tang**. “TACO: Enhancing Multimodal In-context Learning via Task Mapping-Guided Sequence Configuration.” In *Proceedings of Empirical Methods in Natural Language Processing, 2025*.
- [EMNLP’ 25] Hongyi Liu, Shaochen Zhong, Xintong Sun, Minghao Tian, Mohsen Hariri, Zirui Liu, **Ruixiang Tang**, Zhimeng Jiang, Jiayi Yuan, Yu-Neng Chuang, Li Li, Soo-Hyun Choi, Rui Chen, Vipin Chaudhary, Xia Hu. “LoRATK: LoRA Once, Backdoor Everywhere in the Share-and-Play Ecosystem.” In *Findings of Empirical Methods in Natural Language Processing, 2025*.
- [EMNLP’ 25] Lingxi Zhang, Yu-Neng Chuang, Guanchu Wang, **Ruixiang Tang**, Xuanning Cai, Rajesh Shenoy, Xia Hu. “A Decoupled Multi-Agent Framework for Complex Text Style Transfer.” In *Findings of Empirical Methods in Natural Language Processing, 2025*.
- [AACL’ 25] Haein Kong, A M Muntasir Rahman, **Ruixiang Tang**, Vivek Singh. “SafePersuasion: A Dataset, Taxonomy, and Baselines for Analysis of Rational Persuasion and Manipulation.” In *Findings of Asia-Pacific Chapter of the Association for Computational Linguistics, 2025*.
- [COLM’ 25] Yanshu Li, Hongyang He, Yi Cao, Qisen Cheng, Xiang Fu, **Ruixiang Tang**. “M²IV: Towards Efficient and Fine-grained Multimodal InContext Learning in Large Vision-Language Models.” In *Conference on Language Modeling*.
- [ICCV’ 25] Taowen Wang, Dongfang Liu, James Chenhao Liang, Wenhao Yang, Qifan Wang, Cheng Han, Jiebo Luo, **Ruixiang Tang**. “Exploring the Adversarial Vulnerabilities of Vision-Language-Action Models in Robotics.” In *International Conference on Computer Vision*.
- [ICCV’ 25] Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, **Ruixiang Tang** Vladimir Pavlovic. “EAZY: Eliminating Hallucinations in LVLMs by Zeroing out Hallucinatory Image Tokens” In *International Conference on Computer Vision*.
- [ACL’ 25] Huaizhi Ge, Yiming Li, Qifan Wang, Yongfeng Zhang, **Ruixiang Tang**. “When Backdoors Speak: Understanding LLM Backdoor Attacks Through Model-Generated Explanations.” In *Annual Meeting of the Association for Computational Linguistics*.
- [ACL’ 25] Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, **Ruixiang Tang**, William Yang Wang, Yongfeng Zhang. “Disentangling Memory and Reasoning Ability in Large Language Models.” In *Annual Meeting of the Association for Computational Linguistics*.
- [ICML’ 25] Mingyu Jin*, Kai Mei*, Wujiang Xu, Mingjie Sun, **Ruixiang Tang**, Mengnan Du, Zirui Liu, Yongfeng Zhang. “Massive Values in Self-Attention Modules are the Key to Contextual Knowledge Understanding” In *International Conference on Machine Learning*.
- [ICLR’ 25] Yiyang Liu*, James Chenhao Liang*, **Ruixiang Tang**, Yugyung Lee, MAJID RABBANI, Sohail Dianat, Raghuveer Rao, Lifu Huang, Dongfang Liu, Qifan Wang, Cheng Han. “Re-Imagining

Multimodal Instruction Tuning: A Representation View.” In *International Conference on Learning Representations*.

[KDD Explorations’ 25] Zihao Li, **Ruixiang Tang**, Lu Cheng, Shuaiqiang Wang, Dawei Yin, Mengnan Du. “DBR: Divergence-Based Regularization for Debiasing Natural Language Understanding Models.” In *SIGKDD Explorations Letter*.

[CACM’ 24] **Ruixiang Tang**, Yu-Neng Chuang, Xia Hu. “The Science of Detecting LLM-Generated Texts”. In *Communications of the ACM*. **(Featured Cover Article)**

[TKDD’ 24] Jinfeng Yang*, Hongye Jin*, Xiaotian Han*, **Ruixiang Tang***, Qizhang Feng*, Haoming Jiang, Xia Hu. “Harnessing The Power of LLMs in Practice: A Survey on ChatGPT and Beyond.” In *Transactions on Knowledge Discovery from Data*.

[NAACL’ 24] **Ruixiang Tang***, Yu-Neng Chuang*, Xia Hu. “Secure Your Model: A Effective Key Prompt Protection Mechanism for Large Language Models.” In *Findings of Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

[TKDD’ 24] Yu-Neng Chuang, Kwei-Herng Lai, **Ruixiang Tang**, Mengnan Du, Na Zou, Xia Hu. “Mitigating Relational Bias on Knowledge Graphs.” In *Transactions on Knowledge Discovery from Data*.

[EMNLP’ 24] Guanchu Wang*, Yu-Neng Chuang*, **Ruixiang Tang**, Shaochen Zhong, Jiayi Yuan, Hongye Jin, Zirui Liu, Vipin Chaudhary, Shuai Xu, James Caverlee, Xia Hu. “Secured Weight Release for Large Language Models via Taylor Expansion?” In *Proceedings of Empirical Methods in Natural Language Processing, 2024*.

[EMNLP’ 24] Yuqing Zhou, **Ruixiang Tang**, Ziyu Yao, Ziwei Zhu. “Navigating the Shortcut Maze: A Comprehensive Analysis of Shortcut Learning in Text Classification by Language Models” In *Findings of Empirical Methods in Natural Language Processing, 2024*.

[EMNLP’ 24] Wenyue Hua, Xianjun Yang, Mingyu Jin, Wei Cheng, **Ruixiang Tang**, Yongfeng Zhang. “TrustAgent: Towards Safe and Trustworthy LLM-based Agents through Agent Constitution” In *Findings of Empirical Methods in Natural Language Processing, 2024*.

[JBI’ 24] Yu-Neng Chuang, **Ruixiang Tang**, Xiaoqian Jiang, Xia Hu. “SPeC: A Soft Prompt-Based Calibration on Mitigating Performance Variability in Clinical Notes Summarization.” *Journal of Journal of Biomedical Informatics*.

[EMNLP’ 23] **Ruixiang Tang**, Gord Lueck, Rodolfo Quispe, Huseyin A Inan, Janardhan Kulkarni, Xia Hu. “Privacy Analysis in Summarization Models: Did Your Model Memorize My Information?” In *Findings of Conference on Empirical Methods in Natural Language Processing*.

[NeurIPS’ 23] **Ruixiang Tang***, Jiayi Yuan*, Yiming Li, Zirui Liu, Rui Chen, Xia Hu. “Setting the Trap: Capturing and Defeating Backdoor Threats in PLMs through Honeypots”. In *Conference on Neural Information Processing Systems*.

[NeurIPS’ 23] Zirui Liu*, Guanchu Wang*, Shaochen Zhong, Zhaozhuo Xu, Daochen Zha, **Ruixiang Tang**, Zhimeng Jiang, Kaixiong Zhou, Vipin Chaudhary, Shuai Xu, Xia Hu. “Winner-Take-All Column Row Sampling for Memory Efficient Adaptation of Language Model”. In *Conference on Neural Information Processing Systems*.

[KDD Explorations’ 23] **Ruixiang Tang**, Qizhang Feng, Ninghao Liu, Fan Yang, Xia Hu. “Did You Train on My Dataset? Towards Public Dataset Protection with Clean-Label Backdoor Watermarking.” In *SIGKDD Explorations Letter*.

-
- [ECML’ 23] **Ruixiang Tang**, Mengnan Du, Xia Hu. “Deep Serial Number: Computational Watermarking for DNN Intellectual Property Protection.” In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.
- [ACL’ 23] **Ruixiang Tang**, Dehan Kong, Longtao Huang, Hui Xue. “Large Language Models Can be Lazy Learners: Analyze Shortcuts in In-Context Learning” In *Findings of Annual Meeting of the Association for Computational Linguistics*.
- [CIKM’ 23] **Ruixiang Tang**, Hongye Jin, Curtis Wigington, Mengnan Du, Rajiv Jain, Xia Hu. “Exposing Model Theft: A Robust and Transferable Watermark for Thwarting Model Extraction Attacks.” In *ACM Conference on Information and Knowledge Management, short paper*.
- [CIKM’ 23] Yu-Neng Chuang, Guanchu Wang, Chia-Yuan Chang, Kwei-Herng Lai, **Ruixiang Tang**, Fan Yang, Alfredo Costilla Reyes, Kaixiong Zhou, Xiaoqian Jiang, Xia Hu. “DiscoverPath: A Knowledge Refinement and Retrieval System for Interdisciplinarity on Biomedical Research.” In *ACM Conference on Information and Knowledge Management, demo paper*. (**Best Demo Paper Honorable Mention**)
- [AMIA’ 23] **Ruixiang Tang***, Xiaotian Han*, Xiaoqian Jiang, Xia Hu. “Does Synthetic Data of LLMs Help Clinical Text Mining?” In *AMIA Annual Symposium*.
- [AMIA’ 23] Jiayi Yuan, **Ruixiang Tang**, Xiaoqian Jiang, Xia Hu. “Large language models for health-care data augmentation: An example on patient-trial matching.” In *AMIA Annual Symposium*. (**Best Student Paper Award**)
- [AMIA’ 23] Syed A. Rizvi, **Ruixiang Tang**, Xiaoqian Jiang, Xiaotian Ma, Xia Hu. “Local Contrastive Learning for Medical Image Recognition.” In *AMIA Annual Symposium*.
- [ICHI’ 23] **Ruixiang Tang***, Shenghan Zhang*, Haoxuan Li*, Sirui Ding*, Laila Rasmy, Degui Zhi, Na Zou, Xia Hu. “PheME: A deep ensemble framework for improving phenotype prediction from multi-modal data”, In *IEEE International Conference on Healthcare Informatics*.
- [AMIA’ 22] Sirui Ding, **Ruixiang Tang**, Zha, D, Na Zou, Zhang Kai, Xiaoqian Jiang, Xia Huia. “Fairly Predicting Graft Failure in Liver Transplant for Organ Assigning.” In *AMIA Annual Symposium*. (**Best Student Paper (Shortlist) Award**)
- [WWW’ 21] **Ruixiang Tang**, Mengnan Du, Yuening Li, Zirui Liu. Na Zou, Xia Hu. “Mitigating Gender Bias in Captioning Systems.” In *ACM Web Conference*.
- [ICLR’ 21] Qizhang Feng, Ninghao Liu, Fan Yang, **Ruixiang Tang**, Mengnan Du, Xia Hu. “DEGREE: Decomposition Based Explanation for Graph Neural Network.” In *International Conference on Learning Representations*.
- [AAAI’ 21] Mengnan Du, **Ruixiang Tang**, Weijie Fu, Xia Hu, 2021. “Towards Debiasing DNN Models from Spurious Feature Influence.” In *AAAI Conference on Artificial Intelligence*.
- [NeurIPS’ 21] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, **Ruixiang Tang**, Ahmed Awadallah, Xia Hu. “Fairness via Representation Neutralization.” In *Conference on Neural Information Processing Systems*.
- [FBD’ 21] **Ruixiang Tang***, Ninghao Liu*, Fan Yang, Na Zou, Xia Hu. “Defense Against Explanation Manipulation”, In *Frontiers in Big Data, section Machine Learning and Artificial Intelligence*.
- [TCSS’ 21] Fangsheng Wu, Mengnan Du, Chao Fan, **Ruixiang Tang**, Yang Yang, Ali Mostafavi, Xia Hu. “Understanding Social Biases Behind Location Names in Contextual Word Embedding Models.” In *IEEE Transactions on Computational Social Systems*.

-
- [KDD’ 20] **Ruixiang Tang**, Mengnan Du, Ninghao Liu, Fan Yang, Xia Hu. “An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks.” In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- [ISBI’ 20] Chunze Lin, **Ruixiang Tang**, Darryl D Lin, Langechuan Liu, Jiwen Lu, Yunqiang Chen, Dashan Gao, Jie Zhou. “Deep Feature Disentanglement Learning for Bone Suppression in Chest Radiographs.” In *IEEE International Symposium on Biomedical Imaging*.
- [SPIE’ 19] **Ruixiang Tang**, Fakrul Islam Tushar, Songyue Han, Rui Hou, Geoffrey D Rubin, Joseph Y Lo. “Classification of chest CT using case-level weak supervision.” In *SPIE Medical Imaging: Computer-Aided Diagnosis. International Society for Optics and Photonics*.
- [SPIE’ 19] Wanyi Fu, Shobhit Sharma, Taylor Smith, Rui Hou, Ehsan Abadi, Vignesh Selvakumaran, **Ruixiang Tang**, Joseph Y Lo, W Paul Segars, Anuj J Kapadia, Justin B Solomon, Geoffrey D Rubin, Ehsan Samei. “Multi-organ segmentation in clinical-computed tomography for patient-specific image quality and dose metrology.” In *SPIE Medical Imaging: Physics of Medical Imaging. International Society for Optics and Photonics*.
- [SASHIMI’ 19] Chunze Lin, **Ruixiang Tang**, Darryl D Lin, Langechuan Liu, Jiwen Lu, Yunqiang Chen, Dashan Gao, Jie Zhou. “Breast Mass Detection in Mammograms via Blending Adversarial Learning.” In *International Workshop on Simulation and Synthesis in Medical Imaging*.
-

Preprint and Under Review

- [Preprint’ 26] Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, Mubbasir Kapadia, Vladimir Pavlovic, Jiang Liu, Mengdi Wang, Yiyu Shi, Dimitris N. Metaxas, **Ruixiang Tang**. “MemEye: A Visual-Centric Evaluation Framework for Multimodal Agent Memory.”
- [Preprint’ 26] Boxuan Zhang, Jianing Zhu, Zeru Shi, Dongfang Liu, **Ruixiang Tang**. “AgentForesight: Online Auditing for Early Failure Prediction in Multi-Agent Systems.”
- [Preprint’ 26] Yihao Quan, Zeru Shi, Jinman Zhao, **Ruixiang Tang**. “Reinforcing Consistency in Video MLLMs with Structured Rewards.”
- [Preprint’ 26] Rui Wu, **Ruixiang Tang**. “From Rebound to Remedy: Understanding and Mitigating Reward Hacking via Representation Engineering.”
- [Preprint’ 26] Boxuan Zhang, Jianing Zhu, Jiang Liu, Zhangyang Wang, **Ruixiang Tang**. “Micro-Defects Expose Macro-Fakes: Detecting AI-Generated Images via Local Distributional Shifts.”
- [Preprint’ 26] Zeru Shi, Kai Mei, Yihao Quan, Dimitris N. Metaxas, **Ruixiang Tang**. “Enhancing Multi-modal Reasoning via Self-Revisiting Image Evidence.”
- [Preprint’ 26] Yuefei Chen, Yihao Quan, **Ruixiang Tang**. “Where Fake Citations Are Made: Tracing Field-Level Hallucination to Specific Neurons in LLMs.”
- [Preprint’ 25] Zeru Shi, Yingjia Wan, Zhenting Wang, Qifan Wang, Fan Yang, Elisa Kreiss, **Ruixiang Tang**. “Meaningless Tokens, Meaningful Gains: How Activation Shifts Enhance LLM Reasoning.”
- [Preprint’ 25] Rui Wu, Yihao Quan, Zeru Shi, Zhenting Wang, Yanshu Li, **Ruixiang Tang**. “Read the Scene, Not the Script: Outcome-Aware Safety for LLMs.”
- [Preprint’ 25] Zicong He, Boxuan Zhang, Weihao Liu, **Ruixiang Tang**, Cheng Lu. “What Shapes a Creative Machine Mind? Comprehensively Benchmarking Creativity in Foundation Models.”

[Preprint’ 25] Zeru Shi, Zhenting Wang, Yongye Su, Weidi Luo, Fan Yang, **Ruixiang Tang**, Yongfeng Zhang. “Auto-Prompt Generation is Not Robust: Prompt Optimization Driven by Pseudo Gradient.”

TEACHING EXPERIENCE

Instructor , Department of Computer Science, Rutgers University CS 439 Introduction to Data Science	<i>Spring 2026</i>
Instructor , Department of Computer Science, Rutgers University CS 550 Massive Data Mining	<i>Fall 2025</i>
Instructor , Department of Computer Science, Rutgers University CS 439 Introduction to Data Science	<i>Spring 2025</i>
Instructor , Department of Computer Science, Rutgers University CS 598 Trustworthy AI Foundations	<i>Fall 2024</i>
Guest Lecturer , Department of Data Science, New Jersey Institute of Technology Department Seminar: “Safe and Reliable Large Language Models.”	<i>Fall 2024</i>

MENTORSHIP

Ph.D. Students

Boxuan Zhang, Rutgers University, Project: Agent Safety, Agent Memory	<i>September 2025 - Present</i>
Zeru Shi, Rutgers University, Project: Mechanistic Interpretation; Reinforcement Learning; LLM/VLM Reasoning	<i>September 2025 - Present</i>
Yihao Quan, Rutgers University, Project: Hallucination Detection/Mitigation; Spatial Reasoning; Unified Model	<i>September 2025 - Present</i>
Yuefei Chen (Co-advised), Rutgers University, Project: Large language model for causal inference	<i>September 2024 - Present</i>

Master Students

Yanshu Li, Brown University, Project: Vision In-context Learning (AAAI 2026, COLM 2025, EMNLP 2025)	<i>September 2024 - Present</i>
Huaizhi Ge, Columbia University, Project: Defending Large Language Models Backdoor Attacks (ACL 2025)	<i>September 2024 - September 2025</i>

Undergraduate Students

Rui Wu, Rutgers University, Project: LLM Safety Alignment	<i>September 2024 - Present</i>
Leroy Souz, Rutgers University, Project: Detecting AI-generated Audio (Semifinal for the 2025 Rice Business Plan Competition.)	<i>September 2024 - September 2025</i>
Syed A. Rizvi, University of Houston Current: Enrolled in the Ph.D. program in Computer Science at Yale University Project: Medical Image-text Pretraining, Interpretation, AMIA 2023	<i>August 2022 - May 2023</i>

MEDIA COVERAGE

“Rutgers Computer Science Research Team Awarded AMD AI & HPC Cluster Compute Resources” Rutgers Computer Science News	2026
“Rutgers Climate and Energy Fellowships provide support for early career researchers to participate in RCEI program offerings” Rutgers Climate and Energy Institute	2025
“AI vs. copyright:How large vision-language models are changing IP protection” Devdiscourse	2025
“The Science of Detecting LLM-Generated Text” Communications of the ACM	2024
“Large Language Models in Healthcare” Rice CS News	2023
“Mitigating Gender Bias in Captioning Systems” NSF Fairness in AI Program	2021
“TrojanNet – a simple yet effective attack on machine learning models”, The Daily Swig	2020

SELECTED ACADEMIC SERVICES

Editorial Board: Associate Editor, Smart Health Journal

Senior Program Committee Member: ACL Rolling Review (ARR), CIKM’ 23-26.

Program Committee Member: NeurIPS, ICLR, ICML, ARR, AAAI, etc.

Journal Reviewer: PNAS, Nature Human Behaviour, TPAMI, TKDD, JAMIA, npj Digital Medicine, IEEE Transactions on Information Forensics and Security, etc.

Recognition: Top Reviewer of NeurIPS 2023

INDUSTRIAL RESEARCH EXPERIENCE

Applied Scientist Intern at Microsoft, USA *May 2023-August 2023*

Project: Medical Dialogue Generation, Personal Identifying Information (PII) Detection.

Mentors: Gord Lueck, Rodolfo Quispe, Huseyin Inan, and Janardhan Kulkarni.

Research Intern at Microsoft Research, USA. *May 2022-August 2022*

Project: Medical Dialogue Summarization, Privacy Risk Analysis in Large Language Models.

Mentors: Gord Lueck, Rodolfo Quispe, Huseyin Inan, and Janardhan Kulkarni.

Research Scientist Intern at Adobe Research, USA *May 2021-August 2021*

Project: Watermarking Computer Vision Foundation Models and APIs.

Mentors: Curtis Wigington and Rajiv Jain.