

RONGWU XU

He/Him
Paul G. Allen School of Computer Science & Engineering
University of Washington

0xrwxu@gmail.com
rongwuxu.com
[Google Scholar](#)

EDUCATION

University of Washington , Seattle WA, USA	
PHD IN COMPUTER SCIENCE & ENGINEERING	2025 - 2030
- Paul G. Allen School of Computer Science & Engineering	
- Advisor: Tim Althoff	
Tsinghua University , Beijing, China	
MS IN COMPUTER SCIENCE	2022 - 2025
- Institute of Interdisciplinary Information Science	
- Advisor: Wei Xu	
- Thesis: <i>Evaluating and Enhancing Content Safety of Large Language Models</i>	
BENG IN COMPUTER SCIENCE	2019 - 2022
- Department of Computer Science & Technology	
BENG IN ELECTRONICS INFORMATION ENGINEERING (Pre-transfer)	2018 - 2019
- Department of Electronic Engineering	

RESEARCH

My primary interests lie in the field of **NLP, AI and ML**. I am actively exploring areas such as *experience-based RL, open-endedness*, etc.

My previous research examines how LM design and human interaction shape model capabilities, behavior, and alignment. I combined NLP, machine learning, interpretability, and behavioral psychology and developed engineered techniques, rigorous evaluations and real-world applications.

PUBLICATIONS AND MANUSCRIPTS

(* equal contribution, [†] corresponding author)

1. Difficulty-Based Preference Data Selection by DPO Implicit Reward Gap
Xuan Qi*, **Rongwu Xu***, Zhijing Jin
arXiv Preprint
2. AICrypto: A Comprehensive Benchmark For Evaluating Cryptography Capabilities of Large Language Models
Yu Wang, Yijian Liu, Liheng Ji, Han Luo, Wenjie Li, Xiaofei Zhou, Chiyun Feng, Puji Wang, Yuhan Cao, Geyuan Zhang, Xiaojian Li, **Rongwu Xu**, Yilei Chen, Tianxing He
arXiv Preprint
3. The Singapore Consensus on Global AI Safety Research Priorities
Yoshua Bengio, Tegan Maharaj, Luke Ong, ... **Rongwu Xu**, among others
arXiv Preprint
4. AI Awareness
Xiaojian Li, Haoyuan Shi, **Rongwu Xu[†]**, Wei Xu
arXiv Preprint
5. Humanity's Last Exam
Long Phan, Alice Gatti, Ziwen Han, ... **Rongwu Xu** ..., Dan Hendrycks
arXiv Preprint
6. Harnessing Large Language Models to Decode Public Opinions
Qingjie Zhang*, **Rongwu Xu***, Haoting Qian, Di Wang, Ye Yuxian, Yiming Li, Tianwei Zhang,

Wenyu Zhu, Chao Zhang, Hewu Li, Han Qiu
Preprint

7. Detecting LLM-Generated Spam Reviews by Integrating Language Model Embeddings and Graph Neural Network
Xin Liu*, **Rongwu Xu***, Xinyi Jia, Jason Liao, Jiao Sun, Wei Xu
Preprint
8. Rules Created by Symbolic Systems Cannot Constrain a Learning System
Shih-Wai Lin, **Rongwu Xu**, Xiaojian Li, Wei Xu
SSRN Preprint
9. DEBATEQA: Evaluating Question Answering on Debatable Knowledge
Rongwu Xu*, Xuan Qi*, Zehan Qi, Wei Xu, Zhijiang Guo
arXiv Preprint
10. LIFE BENCH: Evaluating Length Instruction Following in Large Language Models
Wei Zhang, Zhenhong Zhou, Junfeng Fang, **Rongwu Xu**, Yuanhe Zhang, Rui Wang, Ge Zhang, Xinfeng Li, Kun Wang, Li Sun, Lingjuan Lyu, Yang Liu, Sen Su
In Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems: Datasets and Benchmarks Track (**NeurIPS D&B 2025**)
11. Nuclear Deployed: Analyzing Catastrophic Risks in Decision-making of Autonomous LLM Agents
Rongwu Xu*, Xiaojian Li*, Shuo Chen*, Wei Xu
In Findings of the 63rd Annual Meeting of the Association for Computational Linguistics (**ACL 2025 Findings**)
12. Does Chain-of-Thought Reasoning Really Reduce Harmfulness from Jailbreaking?
Chengda Lu, Xiaoyu Fan, Yu Huang, **Rongwu Xu**, Jijie Li, Wei Xu
In Findings of the 63rd Annual Meeting of the Association for Computational Linguistics (**ACL 2025 Findings**)
13. On the Role of Attention Heads in Large Language Model Safety
Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, **Rongwu Xu**, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, Yongbin Li
In Proceedings of the Thirteenth International Conference on Learning Representations (**ICLR 2025**)
Oral
14. MR-BEN: A Comprehensive Meta-Reasoning Benchmark for Large Language Models
Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, **Rongwu Xu**, Zehan Qi, Wanru Zhao, Linling Shen, Jianqiao Lu, Haochen Tan, Yukang Chen, Hao Zhang, Zhan Shi, Bailin Wang, Zhijiang Guo, Jiaya Jia
In Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (**NeurIPS 2024**)
15. Knowledge Conflicts for LLMs: A Survey
Rongwu Xu*, Zehan Qi*, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, Wei Xu
In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2024**)
16. Walking in Others' Shoes: How Perspective-Taking Guides LLMs in Reducing Toxicity and Bias
Rongwu Xu, Zi'an Zhou, Tianwei Zhang, Zehan Qi, Su Yao, Ke Xu, Wei Xu, Han Qiu
In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2024**)
17. Course-Correction: Safety Alignment Using Synthetic Preferences
Rongwu Xu*, Yishuo Cai*, Zhenhong Zhou, Renjie Gu, Haiqin Wang, Yan Liu, Tianwei Zhang, Wei Xu, Han Qiu
In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track (**EMNLP 2024**)
18. LONG²RAG: Evaluating Long-Context & Long-Form Retrieval-Augmented Generation with Key Point Recall
Zehan Qi*, **Rongwu Xu***, Zhijiang Guo, Cunxiang Wang, Hao Zhang, Wei Xu

- In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2024 Findings**)
19. How Alignment and Jailbreak Work: Explain LLM Safety through Intermediate Hidden States
Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, **Rongwu Xu**, Fei Huang, Yongbin Li
In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2024 Findings**)
 20. Sing it, Narrate it: Quality Musical Lyrics Translation
Zhuorui Ye, Jinhan Li, **Rongwu Xu**
In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2024 Findings**)
 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
In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (**ACL 2024**)
Oral
Outstanding Paper Award
 22. Preemptive Answer "Attacks" on Chain-of-Thought Reasoning
Rongwu Xu*, Zehan Qi*, Wei Xu
In Findings of the 62nd Annual Meeting of the Association for Computational Linguistics (**ACL 2024 Findings**)
 23. Exploring Chinese Humor Generation: A Study on Two-part Allegorical Sayings
Rongwu Xu
In Proceedings of the 2024 International Joint Conference on Neural Networks (**IJCNN 2024**)
 24. Tempo: Confidentiality Preservation in Cloud-based Neural Network Training
Rongwu Xu and Zhixuan Fang
In Proceedings of the 2024 International Joint Conference on Neural Networks (**IJCNN 2024**)
 25. LSync: A Universal Timeline-synchronizing Solution for Live Streaming
Yifan Xu*, Fan Dang*, **Rongwu Xu**, Xinlei Chen, Yunhao Liu
In *IEEE/ACM Transactions on Networking*, 2024 (**ToN**)
 26. MISO: Legacy-compatible Privacy-preserving Single Sign-on using Trusted Execution Environments
Rongwu Xu, Sen Yang, Fan Zhang, Zhixuan Fang
In Proceedings of the 2023 IEEE 8th European Symposium on Security and Privacy (**EuroS&P 2023**)
 27. LSync: A Universal Event-synchronizing Solution for Live Streaming
Yifan Xu, Fan Dang, **Rongwu Xu**, Xinlei Chen, Yunhao Liu
In Proceedings of the IEEE INFOCOM 2022 - IEEE Conference on Computer Communications (**INFOCOM 2022**)
 28. LifeRec: A Mobile App for Lifelog Recording and Ubiquitous Recommendation
Jiayu Li, Hantian Zhang*, Zhiyu He*, **Rongwu Xu***, Pingfei Wu*, Min Zhang, Yiqun Liu, Shaoping Ma
In Proceedings of the 2022 Conference on Human Information Interaction and Retrieval (**CHIIR 2022**)

HONORS AND AWARDS

Paul G. Allen First-Year Graduate Student Fellowship	2025
Outstanding Graduate at Tsinghua University	2025
- Highest honor, Top 1.3%, 84 out of 6500+, Top 1 in IIIS	
Outstanding Master Thesis at Tsinghua University	2025
- Highest honor, Top 1.3%, 84 out of 6500+, Top 1 in IIIS	
Most Recognized Research Outcomes at Tsinghua University Nomination (Top 2 in IIIS)	2024
Tsinghua University Excellent Teaching Assistant (Top 2%, <200 out of 10000+)	2024
National Scholarship by Ministry of Education of China	2024
- Highest honor, Top 1%, Top 1 in IIIS	
Tsinghua University Outstanding Student Cadre (Top 1.5%, 121 out of 9000+)	2024
ACL 2024 Outstanding Paper Award	2024
- Top 0.79% of submitted papers, 35 out of 4407	
IJCNN 2024 Travel Grants	2024
Tsinghua-Yangtze River Delta International R&D Community Talent Scholarship	2023
Tsinghua University Overall Excellence Scholarship (Top 10%)	2023
Tsinghua University Overall Excellence Scholarship (Top 10%)	2022
Tsinghua University Technological Innovation Excellence Scholarship	2020
1 st in "Youth in Action" Social Practice at Tsinghua University	2020
Tsinghua-Panasonic Scholarship (Top 10%)	2019
Outstanding Volunteers in Beijing	2018
2 nd Prize (Preliminary) in Chinese Chemistry Olympiad (CChO)	2017
1 st Prize in Chinese Chemistry Olympiad (Beijing Regional Qualifiers)	2017

TALKS AND PRESENTATIONS

<i>Catastrophic Risks and Deception of LLM Agents</i>	
- Misalignment and Control Workshop (w. Concordia AI, FAR.AI, etc), Singapore	Apr 2025
<i>The Choice of Research</i>	
- Speech as a student representative at the 2024 IIIS opening ceremony , IIIS, Tsinghua	Sep 2024
<i>Investigating LLMs' Beliefs and Behaviors Under Persuasive Misinformation</i>	
- Oral report@ACL conference, Bangkok, Thailand	Aug 2024
- Short clip , IIIS, Tsinghua	Apr 2024
<i>Knowledge Conflicts for (RAG) LLMs</i>	
- Online talk , NICE (w. Soochow University)	Jul 2024
<i>Privacy-preserving Authentication using TEE</i>	
- Oral report, EuroS&P conference, Delft, The Netherlands	May 2023

EXPERIENCES

Research

University of Toronto , Toronto ON, Canada	Sep 2024 - Dec 2024
RESEARCH ASSISTANT (REMOTE)	
- Host: Zhijing Jin	
- Topic: AI Alignment	
Westlake University , Zhejiang, China	Aug 2023 - Mar 2024
RESEARCH ASSISTANT (REMOTE)	
- Host: Yue Zhang	
- Topic: Knowledge of LLMs	
Duke University , Durham NC, USA	Apr 2021 - Oct 2022
RESEARCH ASSISTANT	
- Host: Fan Zhang (Now at Yale University)	
- Topic: Privacy-preserving Authentication	

Industry

Shanghai Qi Zhi Institute , Shanghai, China RESEARCH INTERN - Mentor: Wei Xu - Topic: Clinical Large Multimodal Language Models	Sep 2024 - Sep 2025
TongYi Vision Intelligence Lab, Alibaba Inc. , Beijing, China RESEARCH INTERN - Mentor: Yu Liu - Topic: Generative Multimodal Models	May 2024 - Aug 2024
Shanghai Qi Zhi Institute , Shanghai, China RESEARCH INTERN - Mentor: Zhixuan Fang - Topic: Decentralized Finance	Dec 2022 - Jan 2023

Teaching

I was invited to share my teaching experience within my department (2024 & 2023) and was awarded as an excellent teaching assistant at Tsinghua University (2024).

<i>Introduction of Large Language Model Applications</i> , Tsinghua University - HEAD TEACHING ASSISTANT & ORGANIZER - Directed labs, organized curriculum, coordinated a team of 10 graduate students	Spring 2025, 2024
<i>Operating System and Distributed System</i> , Tsinghua University - TEACHING ASSISTANT - Held discussions, office hours; graded exams, assignments, and projects	Fall 2023
<i>Distributed System and Blockchain</i> , Tsinghua University - TEACHING ASSISTANT - Held discussions, office hours; graded exams, assignments, and projects	Spring 2022

Mentoring

I have supervised a total of 7 undergraduate/graduate students, many of whom have co-published articles under my mentorship.

Xuan Qi, Undergraduate, IIS, Tsinghua University	Mar 2024 - Present
Yishuo Cai, Undergraduate, Central South University → Peking Univ. (PhD)	Apr 2024 - Oct 2024
Zi'an Zhou, Undergraduate, Zhili College, Tsinghua University	Dec 2023 - May 2024
Tianqi Zhang, Undergraduate, Tsinghua Univ. → Tsinghua Univ. (Master's)	Jun 2023 - Apr 2024
Brian S. Lin, Undergraduate, Tsinghua Univ. → Tsinghua Univ. (Master's)	Jun 2023 - Apr 2024
Shujian Yang, Master's, SPEIT, Shanghai Jiao Tong University	Sep 2023 - Feb 2024
Xingyu Dang, Undergraduate, IIS, Tsinghua University	Oct 2021 - Jul 2022

PROFESSIONAL SERVICES

Reviewer

ACL Rolling Review - Tracks: Ethics, Bias, and Fairness, Human-Centered NLP (2025 -), NLP Applications, Resources and Evaluation	2024 - Present
Transactions on Knowledge and Data Engineering	2025
Transactions on Information, Forensics and Security	2025
IEEE Access	2025

Memberships

Member, Association for Computational Linguistics	Jul 2024 - Present
Member, ACL SIGSEC, Association for Computational Linguistics	Mar 2024 - Present
Member, International Neural Network Society (INNS)	Mar 2024 - Mar 2025
Member, Institute of Electrical and Electronics Engineers (IEEE)	Mar 2024 - Mar 2025

SOCIETAL LEADERSHIPS

I have held positions in various student organizations at Tsinghua University and IIS, and have accumulated rich student work experience.

President , IIIS Graduate Students Union, Tsinghua University	Jun 2024 - Jun 2025
- Responsible for organization and publicity	
Freshman Counselor , Graduate Class 2024, IIIS, Tsinghua University	May 2024 - Jun 2025
- Facilitated communication among new graduate students and fostered class cohesion	
Counselor , The 39 th Summer School for Graduate, Tsinghua University	Aug 2024
- Responsible for publicity	
- <i>Award: Second Prize for Publicity Work</i>	
Captain , Summer Social Practice for Graduate (Shenzhen-Hong Kong Line), IIIS, Tsinghua University	Apr 2024 - Jul 2024
- Coordinated team, liaised with government entities, led publicity efforts	
- <i>Award: Excellent Individual of IIIS Summer Social Practice</i>	
Member , IIIS Graduate Students Union, Tsinghua University	Jun 2023 - Jun 2024
- Organized the first student festival and various social activities	
- <i>Award: Excellent Individual of IIIS Graduate Students Union (2023-2024)</i>	