

# RONGWU XU

He/Him  
Tsinghua University  
FIT Building, Tsinghua University, Beijing, China  
[0xrxu@gmail.com](mailto:0xrxu@gmail.com)  
[rongwuxu.com](http://rongwuxu.com)

## EDUCATION

<b>University of Washington</b>  , Seattle WA, USA	
PHD IN COMPUTER SCIENCE & ENGINEERING	2025 - 2030
- Paul G. Allen School of Computer Science & Engineering	
- Advisor: Prof. Tim Althoff	
<b>Tsinghua University</b>  , Beijing, China	
MS IN COMPUTER SCIENCE	2022 - 2025
- Institute of Interdisciplinary Information Science	
- Advisor: Prof. Wei Xu	
- Thesis: <i>Evaluating and Enhancing Content Safety of Large Language Models</i> (Examiners: Prof. Minlie Huang and Prof. Tianxing He)	
BENG IN COMPUTER SCIENCE	2019 - 2022
- Department of Computer Science & Technology	
BENG IN ELECTRONICS INFORMATION ENGINEERING (Pre-transfer)	2018 - 2019
- Department of Electronic Engineering	
<b>Beijing No.4 High School</b>  , Beijing, China	
HIGH SCHOOL DIPLOMA	2015 - 2018
- Class of Olympiad (Chemistry)	

## RESEARCH

My primary interests lie in the field of natural language processing (NLP). My current exploration focuses on the following aspects of large language models (LLMs):

- **AI Safety and Alignment:** Identifying potential safety and ethics risks associated with AI R&D and developing strategies to align AI systems with human values, behaviors, and expectations.
- **Machine Behavior:** Investigating the similarities and differences between AI models and human behaviors, and utilizing psychology-inspired experiments to test and understand machines.
- **AI and Psychology:** Studying both the understanding the psychological impacts of AI systems on humans (Psychology of AI, a subset of Psychology of Technology) and the application of AI in psychological research (AI for Psychology, a subset of AI for Science).

Beyond these, I am interested in evaluation and application of LLMs.

## PUBLICATIONS AND MANUSCRIPTS

(\* equal contribution, <sup>†</sup> corresponding author)

To date, I have (co)authored **25** research papers, including **15** as first/co-first/corresponding author. Most of my works are accepted to top NLP conferences including ACL/EMNLP. I received an ACL 2024 **Outstanding Paper Award** as the first author.

See [Google Scholar](#) for a complete list and bibliometric.

1. LIFE Bench: Evaluating Length Instruction Following in Large Language Models  
Wei Zhang\*, Zhenhong Zhou\*, Junfeng Fang\*, **Rongwu Xu\***, Kun Wang\*, Yuanhe Zhang, Rui Wang, Ge Zhang, Xinfeng Li, Li Sun, Lingjuan Lyu, Yang Liu, Sen Su  
*arXiv Preprint*

2. AI Awareness  
Xiaojian Li, Haoyuan Shi, **Rongwu Xu**<sup>†</sup>, Wei Xu  
*arXiv Preprint*
3. Humanity's Last Exam  
Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, ... **Rongwu Xu** ..., Summer Yue, Alexandr Wang, Dan Hendrycks  
*arXiv Preprint*
4. 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*
5. 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*
6. Rules Created by Symbolic Systems Cannot Constrain a Learning System  
Shih-Wai Lin, **Rongwu Xu**, Xiaojian Li, Wei Xu  
*SSRN Preprint*
7. DEBATEQA: Evaluating Question Answering on Debatable Knowledge  
**Rongwu Xu**\*, Xuan Qi\*, Zehan Qi, Wei Xu, Zhijiang Guo  
*arXiv Preprint*

————— Below are published papers —————

8. 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)
9. 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)
10. 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**
11. 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)
12. 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)
13. 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)

14. 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)
15. LONG<sup>2</sup>RAG: 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)
16. 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)
17. 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)
18. 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**
19. 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)
20. 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)
21. 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)
22. 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)
23. 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)
24. 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)
25. 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 (SELECTED)

---

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
<b>National Scholarship</b> (Top 1%, by Ministry of Education of China)	2024
Tsinghua University Outstanding Student Cadre (Top 1.5%, 121 out of 9000+)	2024
<b>ACL 2024 Outstanding Paper Award</b> (Top 0.79%, 35 out of 4407)	2024
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 <sup>st</sup> in "Youth in Action" Social Practice at Tsinghua University	2020
Tsinghua-Panasonic Scholarship (Top 10%)	2019
Outstanding Volunteers in Beijing	2018
2 <sup>nd</sup> Prize (Preliminary) in Chinese Chemistry Olympiad (CChO)	2017
1 <sup>st</sup> 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>	
- <a href="#">Speech as a student representative at the 2024 IIIS opening ceremony</a> , IIIS, Tsinghua	Sep 2024
<i>Investigating LLMs' Beliefs and Behaviors Under Persuasive Misinformation</i>	
- Oral report@ACL conference, Bangkok, Thailand	Aug 2024
- <a href="#">Propaganda film</a> , IIIS, Tsinghua	Apr 2024
<i>Knowledge Conflicts for (RAG) LLMs</i>	
- <a href="#">Online talk</a> , 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 Appointments

---

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

### Working

<b>Shanghai Qi Zhi Institute</b> , Shanghai, China RESEARCH INTERN - Mentor: Prof. Wei Xu - Topic: Clinical Large Multimodal Language Models	Sep 2024 - Present
<b>TongYi Vision Intelligence Lab, Alibaba Inc.</b> , Beijing, China RESEARCH INTERN - Mentor: Yu Liu - Topic: Generative Multimodal Models	May 2024 - Aug 2024
<b>Shanghai Qi Zhi Institute</b> , Shanghai, China RESEARCH INTERN - Mentor: Prof. 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, IIIS, 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, IIIS, 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
IEEE Access	2025

### Memberships

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

## SOCIETAL LEADERSHIPS

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

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

## SKILLS AND EXPERTISE

---

- **Research:** Experienced in deep learning/data analysis, also ability with computer systems/applied cryptography/software engineering.
- **Programming Language:** Proficient in C++/Go/Python/Java/JavaScript/L<sup>A</sup>T<sub>E</sub>X, also ability with Verilog/System Verilog/ASM/Rust/P4/HTML/Matlab.
- **Technological:** Proficient in Pytorch/NumPy/Matplotlib/Git/Linux OS/Markdown, also ability with Docker/Wireshark/Microsoft Office.
- **Other Expertise:** Good communication skills, love collaborating with people, experienced in mentoring students, works well in a team.
- **Standard Language Test (TOEFL):** 106 (May 2021).