Rongwu Xu

Personal

Pronouns:	He/Him
Location:	FIT Building, Tsinghua University, Beijing, China
E-mail:	0xrwxu@gmail.com
Homepage:	rongwuxu.com

Education

Jun 2025 – Aug 2022 | MS in Computer Science, **Tsinghua University**
Advisor: Prof. Wei Xu, Vice Dean
Graduate Student at IIIS (Directed by Andrew C. Yao, Turing award laureate 2000)

Jun 2022 – Sep 2019 | BEng in Computer Science, **Tsinghua University**
Bachelor Student at Department of Computer Science and Technology (DCST)

Aug 2019 – Sep 2018 | BEng in Electronic Engineering, **Tsinghua University**
Bachelor Student at Department of Electronic Engineering (EE)

Jun 2018 – Sep 2015 | High School Diploma, **Beijing No.4 High School**
High School Student at the 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)**:

- *Machine Psychology:* What are the key similarities and differences between AI models and human mind? How to use psychology-inspired (behavioral, neuro) experiments to test and understand LLMs?
- *Human-AI Alignment:* What are the key ingredients to make AI systems be effectively aligned with human values, behavioral patterns and expectations beyond machine learning algorithms?
- *AI Safety and Ethics:* What are the primary safety and ethical concerns associated with the deployment and application of LLMs, and how can these be addressed to ensure responsible and trustworthy use?

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

Publications and Manuscripts

To date, I have authored **22** research papers, including **12** first/co-first author pieces. 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. ”Nuclear Deployed!”: Analyzing Catastrophic Risks in Decision-making of Autonomous LLM Agents
Rongwu Xu*, Xiaojian Li*, Shuo Chen*, Wei Xu
arXiv Preprint

2. 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

3. Detecting LLM-Generated Spam Reviews by Integrating Graph Neural Network and Language Model Embeddings
Xin Liu, **Rongwu Xu**, Xinyi Jia, Jason Liao, Jiao Sun, Wei Xu
Preprint

4. Does Chain-of-Thought Reasoning Really Reduce Harmfulness from Jailbreaking?
Chengda Lu, Xiaoyu Fan, Yu Huang, **Rongwu Xu**, Jijie Li, Wei Xu
Preprint

5. Rules Created by Symbolic Systems Cannot Constrain a Learning System
Shih-Wai Lin, **Rongwu Xu**, Xiaojian Li, Wei Xu
SSRN Preprint

6. DebateQA: Evaluating Question Answering on Debatable Knowledge
Rongwu Xu*, Xuan Qi*, Zehan Qi, Wei Xu, Zhijiang Guo
arXiv Preprint, major revision at TACL

Below are published papers

7. 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**, **Oral**), 2025

8. 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**-Industry), 2024

9. 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

10. Long^{2}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**-Findings), 2024

11. 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**-Findings), 2024

12. 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**-Findings), 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**-Main), 2024

14. 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**-Main), 2024

15. 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**-Main, **Oral**), 2024
Outstanding Paper Award

16. 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**-Findings), 2024

17. Exploring Chinese Humor Generation: A Study on Two-part Allegorical Sayings
Rongwu Xu
In Proceedings of *International Joint Conference on Neural Networks* (**IJCNN**), 2024

18. Tempo: Confidentiality Preservation in Cloud-based Neural Network Training
Rongwu Xu and Zhixuan Fang
In Proceedings of *International Joint Conference on Neural Networks* (**IJCNN**), 2024

19. 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* (**ToN**), 2024

20. MISO: Legacy-compatible Privacy-preserving Single Sign-on using Trusted Execution Environments
Rongwu Xu, Sen Yang, Fan Zhang, Zhixuan Fang
In Proceedings of *IEEE European Symposium on Security and Privacy* (**EuroS&P**), 2023

21. LSync: A Universal Event-synchronizing Solution for Live Streaming
Yifan Xu, Fan Dang, **Rongwu Xu**, Xinlei Chen, Yunhao Liu
In Proceedings of *IEEE Conference on Computer Communications* (**INFOCOM**), 2022

22. 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 *ACM SIGIR Conference on Human Information Interaction and Retrieval* (**CHIIR**), 2022

(* equal contribution)

Honors and Awards (Selected)

2024	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** (Top 1%, by Ministry of Education of China)
2024	Tsinghua University Outstanding Student Cadre (Top 1.5%, 121 out of 9000+)
2024	**ACL 2024 Outstanding Paper Award** (Top 0.79%, 35 out of 4407)
2024	IJCNN 2024 Travel Grants
2023	Tsinghua-Yangtze River Delta International R&D Community Talent Scholarship
2023	Tsinghua University Overall Excellence Scholarship (Top 10%)
2022	Tsinghua University Overall Excellence Scholarship (Top 10%)
2020	Tsinghua University Technological Innovation Excellence Scholarship
2020	1st in "Youth in Action" Social Practice at Tsinghua University
2019	Tsinghua-Panasonic Scholarship (Top 10%)
2018	Outstanding Volunteers in Beijing
2017	2nd Prize (Preliminary) in Chinese Chemistry Olympiad (CChO)
2017	1st Prize in Chinese Chemistry Olympiad (Beijing Regional Qualifiers)

Talks and Presentations

Sep 2024	The Choice of Research	Speech as a student representative at the 2024 IIIS opening ceremony
Aug 2024	Investigating LLMs' beliefs and behaviors under misinformation	Oral report@ACL conference
Jul 2024	Knowledge conflicts for (RAG) LLMs	Online talk@NICE (w. Soochow University)
Apr 2024	Investigating LLMs' beliefs and behaviors under misinformation	Propaganda film@IIIS, Tsinghua
May 2023	Privacy-preserving authentication	Oral report@EuroS&P conference

Experiences

Mentoring

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

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

Teaching

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

Spring 2025
Spring 2024 **Head Teaching Assistant & Organizer**, Tsinghua University
Course: Introduction of Large Language Model Applications
Directed and Co-designed labs by mobilizing 10 graduate students, organized the curriculum and assisted labs in class.

Fall 2023 **Teaching Assistant**, Tsinghua University
Course: Operating System and Distributed System
Held discussion and office hours, graded exams, assignments and projects.

Spring 2022 **Teaching Assistant**, Tsinghua University
Course: Distributed System and Blockchain
Held discussion and office hours, graded exams, assignments and projects.

Exchanging

Apr 2021 - Oct 2022 **Research Assistant** (Remote), Duke University
Granted summer overseas internship (undergrad) by Tsinghua
Conducted research in privacy-preserving authentication. The research findings were published in the EuroS&P conference.
Host: Prof. Fan Zhang

Internship

Sep 2024 - Present **Research Intern**, Shanghai Qi Zhi Institute, Shanghai, China
Topic: Clinical Large Multimodal (Language) Models
Conducted research in AI for healthcare. Developed multimodal large language models tailored for clinical chatbots.
Mentor: Prof. Wei Xu

May 2024 - Aug 2024 **Research Intern**, TongYi Vision Intelligence Lab, Alibaba Inc., Beijing, China
Topic: Generative Multimodal Models
Conducted research in vision language tasks and world model.
Mentor: Yu Liu

Dec 2022 - Jan 2023 **Research Intern**, Shanghai Qi Zhi Institute, Shanghai, China
Topic: Decentralized Finance
Conducted research in MEV arbitrage forecasting algorithms using graph neural networks (GNNs).
Mentor: Prof. Zhixuan Fang

Professional Services and Memberships

Services

2024 - Present **Reviewer**, ACL Rolling Review
Tracks: Ethics, Bias, and Fairness, Human-Centered NLP (2025 -), NLP Applications, Resources and Evaluation

Memberships

Jul 2024 - Present **Member**, Association for Computational Linguistics
Mar 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

Societal Leaderships

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

Jun 2024 - Present **President**, IIIS Graduate Students Union, Tsinghua University
In charge of organization and publicity.

May 2024 - Present **Freshman Counselor** (Class 2024 (graduate)), IIIS, Tsinghua University
Responsible for the communication of new students and the collective formation of class.

Apr 2024 - Jul 2024 **Captain**, summer social practice (graduate, Shenzhen-Hong Kong line), IIIS, Tsinghua University
Responsible for organization, mobilization, liaison (government) and advocacy.
Award: Excellent Individual of IIIS summer social practice.

Jun 2023 - Jun 2024 **Member**, IIIS Graduate Students Union, Tsinghua University
Participated in the organization of the first student festival and other social activities.
Award: Excellent Individual of IIIS Graduate Students Union (2023-2024).

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/LaTeX, 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).

- Last updated on Tuesday 25th February, 2025 -