

RONGWU XU

PERSONAL

PRONOUNS: He/Him
LOCATION: FIT building, Tsinghua University, Beijing, China
E-MAIL: 0xrwxu@gmail.com (primary) or xrw22@mails.tsinghua.edu.cn
HOMEPAGE: rongwuxu.site

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)**:

- *Data*: Working on data-driven learning, synthetic data, data generation and data assessment. Data is everywhere in LLMs' pipelines.
- *Evaluation*: Constructing reliable frameworks and datasets to assess LLMs' language generation, contextual understanding, factuality, robustness, among others.
- *Social implication*: Identifying risks in LLMs such as misinformation and ethical issues, aiming to build NLP systems that reflect human values.

I have supervised a total of **10** undergraduate/master's students, many of whom have co-published articles under my mentorship.

EDUCATION

Jun. 2025	MRES IN COMPUTER SCIENCE Tsinghua University  Beijing, China
Aug. 2022	Graduate Student at IIIS (Directed by Andrew C. Yao , Turing award laureate 2000)
Jun. 2022	BENG IN COMPUTER SCIENCE Tsinghua University  Beijing, China
Sept. 2019	Bachelor Student at Department of Computer Science and Technology (DCST)
Aug. 2019	BENG IN ELECTRONIC ENGINEERING Tsinghua University  Beijing, China
Sept. 2018	Bachelor Student at Department of Electronic Engineering (EE)
Jun. 2018	Beijing No.4 High School  Beijing, China
Sept. 2015	High School Student at the Class of Olympiad (Chemistry)

PUBLICATIONS AND MANUSCRIPTS

TL;DR To date, I have authored **16** research papers, including **10** first/co-first author papers. A total of **8** papers have been published in international conferences and journals. I received an ACL 2024 **Outstanding Paper Award** as the first author.

1. DEBATEQA: Evaluating Question Answering on Debatable Knowledge
Rongwu Xu*, Xuan Qi*, Zehan Qi, Wei Xu, Zhijiang Guo
Preprint

2. 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
arXiv Preprints
3. 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
Preprint
4. 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
arXiv Preprints
5. How Alignment and Jailbreak Work: Explain LLM Safety through Intermediate Hidden States
Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, **Rongwu Xu**, Fei Huang, Yongbin Li
arXiv Preprint
6. Sing it, Narrate it: Quality Musical Lyrics Translation
Jinhan Li, Zhuorui Ye, **Rongwu Xu**
Preprint
7. 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
Preprint
8. Knowledge Conflicts for LLMs: A Survey
Rongwu Xu*, Zehan Qi*, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, Wei Xu
arXiv Preprint

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

9. 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
10. 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
11. Exploring Chinese Humor Generation: A Study on Two-part Allegorical Sayings
Rongwu Xu
In Proceedings of International Joint Conference on Neural Networks (IJCNN), 2024
12. 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

13. 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
14. 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
15. 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
16. 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 | ACL 2024 Outstanding Paper Award |
| 2023 | Tsinghua-Yangtze River Delta International R&D Community Talent Scholarship |
| 2023 | Tsinghua University Overall Excellence Scholarship |
| 2020 | Tsinghua University Technological Innovation Scholarship |
| 2020 | 1 st in "Youth in Action" Social Practice@Tsinghua University |
| 2019 | Tsinghua-Panasonic Scholarship |
| 2018 | Outstanding Volunteers in Beijing |
| 2017 | 2 nd Prize (Preliminary) in Chinese Chemistry Olympiad (CChO) |
| 2017 | 1 st Prize in Chinese Chemistry Olympiad (Beijing Regional Qualifiers) |

TALKS AND PRESENTATIONS

- | | | |
|-----------|--|--|
| Aug. 2024 | Investigating LLMs' beliefs and behaviors under misinformation | Oral report@ACL conference |
| Jul. 2024 | Knowledge conflicts for (RAG) LLMs | Online talk@NICE |
| 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

It is my pleasure to collaborate with the following brilliant students:

- | | | |
|------------------------|----------------|---|
| Apr. 2024 - | Yishuo Cai | Undergrad, SE@Central South Univ. |
| Mar. 2024 - | Yishu Yin | Undergrad, CS@Tsinghua Univ. |
| Mar. 2024 - | Priscilla Chen | Undergrad, EECS@UC Berkeley |
| Mar. 2024 - Apr. 2024 | Xinghan Li | Undergrad, IIIS@Tsinghua Univ. |
| Mar. 2024 - | Xuan Qi | Undergrad, IIIS@Tsinghua Univ. |
| Dec. 2023 - May. 2025 | Zi'an Zhou | Undergrad, Zhili College@Tsinghua Univ. |
| Sept. 2023 - Feb. 2024 | Shujian Yang | Master, SPEIT@Shanghai Jiao Tong Univ. |
| Jun. 2023 - Apr. 2024 | Tianqi Zhang | Undergrad, CS@Tsinghua Univ. |
| Jun. 2023 - Apr. 2024 | Brian S. Lin | Undergrad, CS@Tsinghua Univ. |
| Oct. 2021 - Jul. 2022 | Xingyu Dang | Undergrad, IIIS@Tsinghua Univ. |

Teaching

- Spring 2024 **Head Teaching Assistant & Organizer**, Tsinghua University
Introduction of Large Language Model Applications
Directed Co-designed labs by mobilizing 10 graduate students, organized the curriculum and assisted labs in class.
- Fall 2023 **Teaching Assistant**, Tsinghua University
Operating System and Distributed System
Held discussion and office hours, graded exams, assignments and projects.
- Spring 2022 **Teaching Assistant**, Tsinghua University
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

- May. 2024 - **TongYi Vision Intelligence Lab, Alibaba Inc.**, Beijing, China
Conducted research in visual language tasks and world model.
Mentor: Yu Liu
- Dec. 2022 - Jan. 2023 **Shanghai Qi Zhi Institute**, Shanghai, China
Conducted research in decentralized finance (DeFi). Worked on MEV arbitrage forecasting algorithms using graph neural networks (GNNs).
Mentor: Prof. Zhixuan Fang

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^AT_EX, 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.

SERVICES AND MEMBERSHIPS

Academic

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

Societal

- Aug. 2024 **Counselor**, The 39th Summer School for Graduate, Tsinghua University
In charge of publicity.
- Jun. 2024 - Current **President**, IIS Graduate Students Union, Tsinghua University
In charge of organization and publicity.
- May. 2024 - Current **Freshman Counselor**, IIS, Tsinghua University
Class 2024 (graduate)
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), IIS, Tsinghua University
Responsible for organization, mobilization, liaison (government) and advocacy.
Awarded *excellent individual* of IIS summer social practice.
- Sept. 2023 - Jun. 2024 **Member**, IIS Graduate Students Union, Tsinghua University
Participated in the organization of the first student festival and other social activities.
Awarded *excellent individual* of IIS Graduate Students Union (2023—2024).