Rongwu Xu

Personal

Pronouns:	He/Him/His
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 fields of natural language processing (NLP) and trustworthy AI, as well as their intersections, *e.g.*, factuality, safety, and ethics in *up-to-date* NLP systems. Considering these intersections, my current exploration focuses on:

- **Investigating and Enhancing Trust in LLMs**. I identify risks present in large language models (LLMs), including misinformation, safety, ethical concerns. My ultimate goal is to contribute to more trustworthy NLP systems that align with human values.
- **Utilizing NLP in Public Discourse and Moderation**. I apply the advancements in NLP to analyze public opinions. This involves understanding, interpreting, and managing the vast array of information expressed in digital platforms.

My vision is to harness the power of AI and NLP to create trustworthy and human-centric digital environments.

Education

Jun. 2025	MRes in Computer Science **Tsinghua University** Beijing, China Advisor: Prof. Wei Xu
Aug. 2022	Graduate Student at IIIS (Headed by Andrew C. Yao, the Turing award laureate'2000)
Jun. 2022	BEng in Computer Science **Tsinghua University** Beijing, China
Sept. 2018	Bachelor Student at Department of Computer Science and Technology (DCST)
Jun. 2018	**Beijing No.4 High School** Beijing, China
Sept. 2015	High School Student at the Class of Olympiad (Chemistry)

Publications and Manuscripts

11. 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

10. Understandable and Singable Musical Lyrics Translation
Jinhan Li, Zhuorui Ye, **Rongwu Xu**
Preprint

9. 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*, Cunxiang Wang, Hongru Wang, Yue Zhang, Wei Xu
arXiv Preprint

7. 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

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

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

4. 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

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

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

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

(* equal contribution)

Honors and Awards

2023 Overall Excellence Scholarship@Tsinghua University
2020 Technological Innovation Scholarship@Tsinghua University
2020 1st in "Youth in Action" Social Practice@Tsinghua University
2019 Tsinghua-Panasonic Scholarship@Tsinghua University
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

Apr. 2024	Investigating large language models' 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 -	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 -	Brian S. Lin	Undergrad, CS@Tsinghua Univ.
Oct. 2021 - Jul. 2022	Xingyu Dang	Undergrad, IIIS@Tsinghua Univ.

Teaching

Spring 2024	**Teaching Assistant & Organizer**, Tsinghua University *Introduction of Large Language Model Applications* Co-designed labs, 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. Host: Prof. Fan Zhang

Internship

Apr. 2024 -	**TongYi Vision Intelligence Lab, Alibaba Inc.**, Beijing, China TBA TBA
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/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, experi-

enced in mentoring students, works well in a team.

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

Jun. 2024 - Current	**2024 Graduate Freshman Assistant**, IIIS, Tsinghua University
Apr. 2024 - Current	**Social Practice Stewardship**, IIIS, Tsinghua University
Nov. 2023 - Current	**Member**, IIIS Students Congress, Tsinghua University

— Last updated on Thursday 16th May, 2024 —