Rongwu Xu

Personal Info

Pronouns:	He/Him/His
Location:	FIT building, Tsinghua University, Beijing, China
Place of Birth:	Beijing, China
E-mail:	0xrwxu@gmail.com (primary) or xrw22@mails.tsinghua.edu.cn
Homepage:	rongwuxu.site

Research

I delve into the *intersection* of trustworthy AI [TAI] and natural language processing [NLP] , with a focus on safety and ethics issues in large language models (LLMs). My concentration encompasses two primary areas:

- **Investigating and Enhancing Trust in LLMs**. I investigate and address the inherent safety, ethics, and bias concerns present in LLMs. My ultimate goal is to develop strategies that mitigate these issues, contributing to the creation of more trustworthy AI systems that align with human values.
- **Utilizing NLP in Public Discourse and Moderation**. I apply the advancements in NLP to analyze public opinions and to enhance content moderation. This involves leveraging cutting-edge language techniques to understand, interpret, and manage the vast array of information and viewpoints expressed in digital platforms.

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

Education

06/2025 (Expected)	M.S. in Computer Science **Tsinghua University** Beijing, China Advisor: Prof. Wei Xu
08/2022–current	Graduate Student at IIIS (Headed by Andrew C. Yao, the Turing award laureate'2000)
06/2022	B.E. in Computer Science **Tsinghua University** Beijing, China
09/2018–06/2022	Bachelor Student at Department of Computer Science and Technology (DCST)

Publications and Manuscripts

11. [TAI] [NLP] Preemptive Answer "Attacks" on Chain-of-Thought Reasoning.
Rongwu Xu*, Zehan Qi*, Wei Xu.
Preprint.

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

9. [TAI] [NLP] Perspective-taking Gives More Ethical Responses.
Rongwu Xu, Zi'an Zhou, Tianwei Zhang, Zehan Qi, Su Yao, Ke Xu, Wei Xu, Han Qiu.
Preprint.

8. [NLP] A Survey on Knowledge Conflicts in the Era of LLMs.
Rongwu Xu*, Zehan Qi*, Cunxiang Wang, Hongru Wang, Yue Zhang, Wei Xu.
Preprint.

7. [TAI] [NLP] 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.
Preprint.

6. [NLP] Exploring Chinese Humor Generation: A Study on Two-part Allegorical Sayings.
Rongwu Xu.
Preprint.

5. [TAI] Tempo: Confidentiality Preservation in Cloud-based Neural Network Training.
Rongwu Xu and Zhixuan Fang.
Preprint.

4. LSync: A Universal Timeline-synchronizing Solution for Live Streaming.
Yifan Xu*, Fan Dang*, **Rongwu Xu**, Xinlei Chen, Yunhao Liu.
Under review. Submitted to IEEE/ACM ToN. Journal version of [INFOCOM'2022].

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* (**Euro S&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)

Awards

2023	Overall Excellence Scholarship	Tsinghua Univ., China
2020	Technological Innovation Scholarship	Tsinghua Univ., China
2020	First Prize in "Youth in Action" Social Practice	Tsinghua Univ., China
2019	Tsinghua-Panasonic Scholarship	Tsinghua Univ., China

Talk and Presentation

May. 2023	Privacy-preserving authentication	Oral report@Euro S&P conference

Experience

Mentoring

Dec. 2023 -	Zi'an Zhou	Undergrad, Zhili College (Information and Computational Science)@Tsinghua Univ.
Sep. 2023 - Feb. 2024	Shujian Yang	Master, SPEIT (Information Engineering)@SJTU.
Jun. 2023 -	Tianqi Zhang	Undergrad, CS@Tsinghua Univ.
Jun. 2023 - Jan. 2024	Brian S. Lin	Undergrad, CS@Tsinghua Univ.
Oct. 2021 - Jul. 2022	Xingyu Dang	Undergrad, IIIS@Tsinghua Univ.

Teaching

SS 2024 **Teaching Assistant & Organizer**, Tsinghua University.
Introduction of Large Language Model Applications
Co-designed labs and organized the curriculum.

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

SS 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
Mentor: Prof. Fan Zhang.

(Research-oriented) Internship

Dec. 2022 - Jan. 2023 **Student Researcher**, 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

- **Research:** Proficient in computer system, deep learning, data analysis, applied cryptography, and software engineering.
- **Programming language:** Proficient in C++/Go/Python/Java/JavaScript, basic coding in Verilog/System Verilog/ASM/Rust/P4/MATLAB.
- **Other tools:** Proficient in Git/Docker/Linux OS/LaTeX/Markdown/Wireshark/Microsoft Office.