


RONGWU XU

PERSONAL

PRONOUNS: He/Him
LOCATION: FIT Building, Tsinghua University, Beijing, China
E-MAIL: 0xrw Xu@gmail.com
HOMEPAGE: rongwuxu.com

EDUCATION

| | |
|----------|--|
| Jun 2025 | MS IN COMPUTER SCIENCE, Tsinghua University  <i>Advisor: Prof. Wei Xu, Vice Dean</i> |
| Aug 2022 | Graduate Student at IIIS (Directed by Andrew C. Yao, Turing award laureate 2000) |
| Jun 2022 | BENG IN COMPUTER SCIENCE, Tsinghua University  |
| Sep 2019 | Bachelor Student at Department of Computer Science and Technology (DCST) |
| Aug 2019 | BENG IN ELECTRONIC ENGINEERING, Tsinghua University  |
| Sep 2018 | Bachelor Student at Department of Electronic Engineering (EE) |
| Jun 2018 | HIGH SCHOOL DIPLOMA, Beijing No.4 High School  |
| Sep 2015 | 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)**:

- *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.

PUBLICATIONS AND MANUSCRIPTS

To date, I have authored **17** research papers, including **10** 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 (**Citations**: 129, **h-index**: 5).

1. DEBATEQA: Evaluating Question Answering on Debatable Knowledge
Rongwu Xu*, Xuan Qi*, Zehan Qi, Wei Xu, Zhijiang Guo
arXiv Preprint, major revision at TACL
2. 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
arXiv Preprint

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

3. 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
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
 In Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024
5. 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
 In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP-Findings), 2024
6. 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
7. 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
8. 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
9. 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
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. 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
12. Exploring Chinese Humor Generation: A Study on Two-part Allegorical Sayings
Rongwu Xu
 In Proceedings of International Joint Conference on Neural Networks (IJCNN), 2024

13. 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
14. 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
15. 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
16. 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
17. 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 2024 2024 2024 2024 2023 2023 2022 2020 2020 2019 2018 2017 2017 | Most Recognized Research Outcomes at Tsinghua University Nomination (Top-2 in IIIS) Tsinghua University Excellent Teaching Assistant (Top 2%, <200 out of 10000+) National Scholarship (Top 1%, by Ministry of Education of China) Tsinghua University Outstanding Student Cadre (Top 1.5%, 121 out of 9000+) ACL 2024 Outstanding Paper Award (Top 0.79%, 35 out of 4407) IJCNN 2024 Travel Grants Tsinghua-Yangtze River Delta International R&D Community Talent Scholarship Tsinghua University Overall Excellence Scholarship (Top 10%) Tsinghua University Overall Excellence Scholarship (Top 10%) Tsinghua University Technological Innovation Excellence Scholarship 1 st in “Youth in Action” Social Practice at Tsinghua University Tsinghua-Panasonic Scholarship (Top 10%) Outstanding Volunteers in Beijing 2 nd Prize (Preliminary) in Chinese Chemistry Olympiad (CChO) 1 st Prize in Chinese Chemistry Olympiad (Beijing Regional Qualifiers) |
|--|---|

TALKS AND PRESENTATIONS

- | | | |
|--|---|---|
| Sep 2024 Aug 2024 Jul 2024 Apr 2024 May 2023 | The Choice of Research Investigating LLMs’ beliefs and behaviors under misinformation Knowledge conflicts for (RAG) LLMs Investigating LLMs’ beliefs and behaviors under misinformation Privacy-preserving authentication | Speech as a student representative at the 2024 IIIS opening ceremony Oral report@ACL conference Online talk@NICE (w. Soochow University) Propaganda film@IIIS , Tsinghua 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 2024 | Head Teaching Assistant & Organizer , Tsinghua University <i>Course:</i> 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 <i>Course:</i> Operating System and Distributed System Held discussion and office hours, graded exams, assignments and projects. |
| Spring 2022 | Teaching Assistant , Tsinghua University <i>Course:</i> Distributed System and Blockchain Held discussion and office hours, graded exams, assignments and projects. |

Exchanging

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

Internship

| | |
|---------------------|--|
| Sep 2024 - Present | Research Intern , Shanghai Qi Zhi Institute, Shanghai, China <i>Topic:</i> Clinical Large Multimodal (Language) Models Conducted research in AI for healthcare. Developed multimodal large language models tailored for clinical chatbots. <i>Mentor:</i> Prof. Wei Xu |
| May 2024 - Aug 2024 | Research Intern , TongYi Vision Intelligence Lab, Alibaba Inc., Beijing, China <i>Topic:</i> Generative Multimodal Models Conducted research in vision language tasks and world model. <i>Mentor:</i> Yu Liu |
| Dec 2022 - Jan 2023 | Research Intern , Shanghai Qi Zhi Institute, Shanghai, China <i>Topic:</i> Decentralized Finance Conducted research in MEV arbitrage forecasting algorithms using graph neural networks (GNNs). <i>Mentor:</i> Prof. Zhixuan Fang |

PROFESSIONAL SERVICES AND MEMBERSHIPS

Services

2024 **Reviewer**, ACL Rolling Review

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. <i>Award</i> : 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. <i>Award</i> : 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/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 (May 2021).