许融武

基本信息

性别：男
出生年月：2000 年 2 月
出生地点：北京市西城区
工作地点：清华大学 FIT 楼 6 层交叉信息院
电子邮件：0xrwxu@gmail.com
个人网站：rongwuxu.site

履历

学习经历

2015 年 9 月—2018 年 8 月　**北京市第四中学**高中学习
2018 年 9 月—2019 年 8 月　**清华大学电子工程系**电子信息工程专业学习
2019 年 9 月—2022 年 8 月　**清华大学计算机系**计算机科学与技术专业学习
2022 年 6 月获计算机科学与技术工学学士学位
2022 年 9 月—至今　**清华大学交叉信息院**计算机科学与技术专业硕士研究生学习
交叉信息院由 2000 年图灵奖得主姚期智院士领导

学生干部经历

2023 年 9 月—2024 年 6 月　**干事**，清华大学交叉信息院研究生会
参与组织首届院学生节和联谊活动。
2024 年 6 月：获评院研会优秀个人。
2024 年 4 月—2024 年 7 月　**支队长**，清华大学交叉信息院暑期社会实践（深圳支线）
负责组织、动员、联络（政府）和宣传工作。
2024 年 9 月：获评社会实践优秀个人。
2024 年 8 月　**辅导员**，清华大学第十八届研究生新生骨干培训班暨第三十九期暑期团校（研究生班）
分管宣传工作，集体获评宣传工作二等奖。
2024 年 5 月—至今　**新生助理**，清华大学交叉信息院
负责 2024 级研究生新生入学事宜和班集体（交叉研 241、交叉研 242）组建。
2024 年 6 月—至今　**主席**，清华大学交叉信息院研究生会
分管组织和宣传工作。

教学经历

2022 年 2 月—2022 年 6 月　**助教**，清华大学交叉信息院
课程：分布式系统和区块链
我主持了讨论和答疑，完成考试，作业和课程项目评分等工作。
2023 年 9 月—2024 年 1 月　**助教**，清华大学交叉信息院
课程：操作系统与分布式系统
我主持了讨论、答疑和带领习题课，完成考试，作业和课程项目评分等工作。
2024 年 2 月—2024 年 6 月　**领头助教 & 组织者**，清华大学交叉信息院
课程：大语言模型应用概论
本课程是当年的清华本科生新开设课程。作为 2 位领头助教之一，我配合教授完成了课程大纲设计，统筹课程具体事宜。并组织动员了 10 位同学从零开始编纂了课程实验的代码。

实习经历

2022 年 12 月—2023 年 1 月　**实习生**，上海期智研究院
在去中心化金融进行研究。使用了图神经网络研究预测算法。

2024 年 5 月—至今　**实习生**，阿里巴巴通义基础视觉研究室
在视觉语言任务和世界模型进行研究。

交换经历

2021 年 4 月—2022 年 10 月　**科研助理**（远程），美国杜克大学
清华大学计算机系海外暑期实习
在隐私保护和认证方向进行研究。使用可信执行环境研究用户身份认证。研究成果以论文发表在 EuroS&P 会议上。

主要奖励与荣誉

2024　**国家奖学金** (前 1%)
2024　清华大学优秀学生干部 (前 <1.5%, 121/9000+)
2024　**ACL 2024 杰出论文奖** (前 0.79%, 35/4407)
2024　IJCNN 2024 会议旅行津贴
2023　清华-长三角国际研发社区英才奖学金
2023　清华大学综合优秀奖学金
2020　清华大学科技创新奖学金
2020　清华大学 "青年行" 社会实践一等奖学金
2019　清华大学清华-松下奖学金
2018　北京市优秀志愿者
2017　中国化学奥林匹克竞赛（初赛）二等奖
2017　北京市化学奥林匹克竞赛一等奖

科研工作

我的主要兴趣在于**自然语言处理**领域。我目前的探索集中在**大型语言模型**的如下方面：

- 数据：数据无处不在，我的研究重点在于数据驱动的学习。我特别关注如何利用合成数据进行模型微调、探索数据生成技术，以及实施数据评估方法。
- 评测：如何评估能力日渐增强的生成式语言模型（"超级模型"）？通过构造可靠的评测框架与数据集，我评估包括它们的语言生成能力、上下文理解能力、事实性和鲁棒性等。
- 社会影响：我识别大型语言模型中存在的风险，包括错误信息、安全和道德问题。最终目标是构建更符合人类价值观的自然语言处理系统。

我累计指导过本科/硕士学生达 **10** 人次，其中不少同学在我的带领下合作发表了文章。

发表论文和手稿

截至目前，我一共完成 **16** 篇学术论文, 包括 **10** 篇第一作者（含共同一作）文章。我的大部分作品被包括 ACL/EMNLP 在内的顶级 NLP 国际会议接收。我以第一作者身份获得了 ACL 2024 的**杰出论文奖**。

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. 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 Preprint

以下为已发表论文

3. 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

4. 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

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
In Findings of *The 2024 Conference on Empirical Methods in Natural Language Processing* (**EMNLP**-Findings), 2024

6. Sing it, Narrate it: Quality Musical Lyrics Translation
Jinhan Li, Zhuorui Ye, **Rongwu Xu**
In Findings of *The 2024 Conference on Empirical Methods in Natural Language Processing* (**EMNLP**-Findings), 2024

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
In Proceedings of *The 2024 Conference on Empirical Methods in Natural Language Processing* (**EMNLP**-Main), 2024

8. 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

9. 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

10. 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

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

(* 表示同等贡献)

讲演与演示

2024 年 8 月	做学术的选择	在 2024 年交叉信息院开学典礼上作为在校生代表的发言
2024 年 8 月	大语言模型面对误信息的行为和信念	口头报告 @ACL 会议
2024 年 7 月	（检索增强）大语言模型的知识冲突	在线讲解@NICE
2024 年 4 月	大语言模型面对误信息的行为和信念	宣传片@ 清华大学交叉信息院
2023 年 5 月	隐私保护的身份验证	口头报告 @EuroS&P 会议

学术服务和组织

学术服务

2024 **审稿人**, ACL Rolling Review

学术组织

2024 年 7 月—至今 **会员**, Association for Computational Linguistics
2024 年 3 月—至今 **会员**, International Neural Network Society (INNS)
2024 年 3 月—至今 **会员**, Institute of Electrical and Electronics Engineers (IEEE)
2024 年 3 月—至今 **会员**, ACL SIGSEC, Association for Computational Linguistics

— 更新于 Sunday 22nd September, 2024 —