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
政治面貌：中共预备党员
工作地点：清华大学 FIT 楼 6 层交叉信息院
电子邮件：xrw22@mails.tsinghua.edu.cn
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**就读高中
2018 年 9 月—2022 年 8 月　**清华大学计算机系**就读本科
2022 年 9 月—至今　**清华大学交叉信息院**就读硕士研究生

主要工作经历

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—至今　**支队长**，清华大学交叉信息院暑期社会实践（深圳—香港线路）
负责组织、动员、联络（政府）和宣传工作。

2024 年 5 月—至今　**新生助理**，清华大学交叉信息院 2024 级研究生
负责新生入学沟通和党团班集体组建。

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。

主要教学经历

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

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。

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。

主要奖励与荣誉

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**和**计算社会科学**领域，以及它们的交叉点。

我目前的探索集中在**大型语言模型**的如下方面：

- **自然语言处理：**
  - 大型语言模型的评测：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。
  - 大型语言模型的安全性和对齐：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。
  - 大型语言模型的可解释性：可解释性是理解模型“思考决策”过程的关键。研究包括但不限于可视化技术、注意力机制分析、以及模型预测的逻辑解释。更好地理解超级模型的工作原理，也能提高用户信任并促进相应改进。
- **计算社会科学：**
  - 社会舆论理解：我应用自然语言处理的最新技术来分析公众舆论。这涉及理解、解释和管理数字平台中表达的大量信息

我取得了一系列创新性研究成果。我累计合作完成论文 13 篇（含手稿），其中包含 8 篇第一（含共同第一）作者文章。所有论文在国际会议和期刊上发表共 8 篇。我累计指导过本科/硕士学生达 10 人次，其中不少在我带领下合作发表了文章。

发表论文和手稿

13. 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
Preprint

12. 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

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

10. 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

9. Knowledge Conflicts for LLMs: A Survey
Rongwu Xu*, Zehan Qi*, Cunxiang Wang, Hongru Wang, Yue Zhang, Wei Xu
arXiv Preprint

以下为已发表论文

8. 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

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 Proceedings of *International Joint Conference on Neural Networks* (**IJCNN**), 2024

5. 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

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 Proceedings of *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 Proceedings of *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 Proceedings of *ACM SIGIR Conference on Human Information Interaction and Retrieval* (**CHIIR**), 2022

(* 表示同等贡献)