

许融武

基本信息

性别:	男
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.com

履历

学习经历

北京市第四中学高中学习	2015 年 – 2018 年
清华大学电子工程系电子信息工程专业学习	2018 年 – 2019 年
清华大学计算机系计算机科学与技术专业学习	2019 年 – 2022 年
- 2022 年 6 月获计算机科学与技术工学学士学位	
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	2022 年 – 2025 年
- 导师: 徐威教授	
- 论文: 评测和增强大型语言模型的内容安全性 (评阅人: 黄民烈教授, 贺天行助理教授)	
华盛顿大学 Paul G. Allen 计算机学院计算机科学与工程专业博士研究生学习	2025 年 – 2030 年
- 导师: Tim Althoff 副教授	

非学位学习经历

多伦多大学科研助理 (远程)	2024 年 9 月 – 至今
- 主持人: Zhijing Jin 助理教授	
- 课题: 人工智能对齐	
西湖大学科研助理 (远程)	2023 年 8 月 – 2024 年 3 月
- 主持人: 张岳教授	
- 课题: 大语言模型的知识能力	
杜克大学科研助理 (清华大学计算机系海外暑期实习)	2021 年 4 月 – 2022 年 10 月
- 主持人: Fan Zhang 助理教授 (现于耶鲁大学)	
- 课题: 隐私保护的身份认证	

工作经历

上海期智研究院实习生	2022 年 12 月 – 2023 年 1 月
- 课题: 去中心化金融	
- 在去中心化金融进行研究。使用图神经网络进行套利预测算法的探索。	
阿里巴巴通义基础视觉研究室实习生	2024 年 5 月 – 2024 年 8 月
- 课题: 生成式多模态模型	
- 在视觉语言任务和世界模型方向开展研究。	
上海期智研究院实习生	2024 年 9 月 – 至今
- 课题: 临床多模态大语言模型	
- 在医疗人工智能方向进行研究, 参与构建用于临床问诊的多模态医疗大模型。	

教学经历

我在交叉信息院内多次分享我的教学经验 (2023、2024) 并被评为清华大学校级优秀助教。

分布式系统和区块链（清华大学交叉信息院）助教 - 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。	2022 年 2 月 – 2022 年 6 月
操作系统与分布式系统（清华大学交叉信息院）助教 - 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。	2023 年 9 月 – 2024 年 1 月
大语言模型应用概论（清华大学交叉信息院）领头助教 & 组织者 - 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 位同学从零开始编写了课程实验的代码。	2024 年 2 月 – 2024 年 6 月 2025 年 2 月 – 2025 年 6 月

学生干部经历

我在清华大学、交叉信息院担任过多个学生组织的职务，积累了丰富的学生工作经验。

清华大学交叉信息院研究生会干事 - 参与组织首届院学生节和联谊活动。 - 奖项：院研会优秀个人	2023 年 6 月 – 2024 年 6 月
清华大学交叉信息院暑期社会实践（深圳支线）支队长 - 负责组织、动员、联络（政府）和宣传工作。 - 奖项：社会实践优秀个人	2024 年 4 月 – 2024 年 7 月
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）辅导员 - 分管宣传工作，集体获评 宣传工作二等奖。	2024 年 8 月
清华大学交叉信息院德育助理（新生助理） - 负责 2024 级研究生新生入学事宜和新生班集体（交叉研 41、交叉研 42）组建。	2024 年 5 月 – 至今
清华大学交叉信息院研究生会主席 - 分管组织和宣传工作。	2024 年 6 月 – 至今

主要奖励与荣誉

清华大学最受师生关注的年度亮点成果提名（系内唯二）	2024 年
清华大学优秀助教（前 2%，<200/10000+）	2024 年
国家奖学金（前 1%）	2024 年
清华大学优秀学生干部（前 1.5%，121/9000+）	2024 年
ACL 2024 杰出论文奖（前 0.79%，35/4407）	2024 年
IJCNN 2024 会议旅行津贴	2024 年
清华-长三角国际研发社区英才奖学金	2023 年
清华大学综合优秀奖学金（前 10%）	2023 年
清华大学综合优秀奖学金（前 10%）	2022 年
清华大学科技创新优秀奖学金	2020 年
清华大学“青年行”社会实践一等奖学金	2020 年
清华大学清华-松下奖学金（前 10%）	2019 年
北京市优秀志愿者	2018 年
中国化学奥林匹克竞赛（初赛）二等奖	2017 年
北京市化学奥林匹克竞赛一等奖	2017 年

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：

- **人工智能安全和对齐**：识别与人工智能研发相关的潜在安全和道德风险，并制定策略以使 AI 系统与人类价值观、行为和期望保持一致。
- **机器行为学**：研究 AI 模型和人类行为之间的异同，并利用受心理学启发的实验来测试和理解机器。
- **人工智能与心理学**：研究理解人工智能系统对人类的心理影响（Psychology of AI, Psychology of Technology 的一个子集）和人工智能在心理学研究中的应用（AI for Psychology, AI for Science 的一个子集）。

除了这些，我还对大型语言模型的评估和应用感兴趣。

发表论文和手稿

(* 表示同等贡献, [†] 表示通讯作者)

截至目前, 我一共完成了 **24** 篇学术论文, 包括 **13** 篇第一作者 (含共同一作) 和通讯作者文章。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**。

完整论文列表和文献计量学信息请参考[谷歌学术](#)。

1. AI Awareness
Xiaojian Li, Haoyuan Shi, **Rongwu Xu**[†], Wei Xu
arXiv Preprint
2. Humanity's Last Exam
Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, ... **Rongwu Xu** ..., Summer Yue, Alexandr Wang, Dan Hendrycks
arXiv Preprint
3. Harnessing Large Language Models to Decode Public Opinions
Qingjie Zhang*, **Rongwu Xu***, Haoting Qian, Di Wang, Ye Yuxian, Yiming Li, Tianwei Zhang, Wenyu Zhu, Chao Zhang, Hewu Li, Han Qiu
Preprint
4. Detecting LLM-Generated Spam Reviews by Integrating Graph Neural Network and Language Model Embeddings
Xin Liu, **Rongwu Xu**, Xinyi Jia, Jason Liao, Jiao Sun, Wei Xu
Preprint
5. Rules Created by Symbolic Systems Cannot Constrain a Learning System
Shih-Wai Lin, **Rongwu Xu**, Xiaojian Li, Wei Xu
SSRN Preprint
6. DEBATEQA: Evaluating Question Answering on Debatable Knowledge
Rongwu Xu*, Xuan Qi*, Zehan Qi, Wei Xu, Zhijiang Guo
arXiv Preprint

以下为已发表论文

7. Nuclear Deployed: Analyzing Catastrophic Risks in Decision-making of Autonomous LLM Agents
Rongwu Xu*, Xiaojian Li*, Shuo Chen*, Wei Xu
In Findings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025 Findings)
8. Does Chain-of-Thought Reasoning Really Reduce Harmfulness from Jailbreaking?
Chengda Lu, Xiaoyu Fan, Yu Huang, **Rongwu Xu**, Jijie Li, Wei Xu
In Findings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025 Findings)
9. 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
In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)
Oral
10. 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 2024)
11. 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)*
12. 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 2024 Findings)*
 13. 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 2024 Findings)*
 14. 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 2024 Findings)*
 15. 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 2024)*
 16. 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 2024)*
 17. 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)*
Oral
Outstanding Paper Award
 18. 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 Findings)*
 19. Exploring Chinese Humor Generation: A Study on Two-part Allegorical Sayings
Rongwu Xu
In Proceedings of the *2024 International Joint Conference on Neural Networks (IJCNN 2024)*
 20. Tempo: Confidentiality Preservation in Cloud-based Neural Network Training
Rongwu Xu and Zhixuan Fang
In Proceedings of the *2024 International Joint Conference on Neural Networks (IJCNN 2024)*
 21. 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*, 2024 (**ToN**)
 22. MISO: Legacy-compatible Privacy-preserving Single Sign-on using Trusted Execution Environments
Rongwu Xu, Sen Yang, Fan Zhang, Zhixuan Fang
In Proceedings of the *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P 2023)*
 23. LSync: A Universal Event-synchronizing Solution for Live Streaming
Yifan Xu, Fan Dang, **Rongwu Xu**, Xinlei Chen, Yunhao Liu
In Proceedings of the *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications (INFOCOM 2022)*

24. 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 the 2022 Conference on Human Information Interaction and Retrieval (**CHIIR 2022**)

讲演与演示

LLM 代理的灾难性风险和欺骗

- Misalignment and Control Workshop (由 Concordia AI, FAR.AI 等协办), 新加坡 2025 年 4 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

- ACL 会议口头报告, 泰国曼谷 2024 年 8 月

- 宣传片@清华大学交叉信息院 2024 年 4 月

(检索增强) 大语言模型的知识冲突

- Bilibili 在线讲解@NICE (由苏州大学协办) 2024 年 7 月

隐私保护的身份验证

- EuroS&P 会议口头报告, 荷兰代尔夫特 2023 年 5 月

学术服务

审稿人

ACL Rolling Review 2024 年 – 至今

- 方向: 伦理、公平性与偏见、人本自然语言处理 (自 2025 年起)、NLP 应用、资源与评估

IEEE Access 2025 年

会员

会员, 计算语言学协会 (ACL) 2024 年 7 月 – 至今

会员, 国际神经网络学会 (INNS) 2024 年 3 月 – 至今

会员, 电气与电子工程师协会 (IEEE) 2024 年 3 月 – 至今

会员, ACL 安全与隐私特别兴趣小组 (SIGSEC) 2024 年 3 月 – 至今

指导学生

我累计指导过本科生和研究生达 7 人次, 其中不少同学在我的带领下合作发表了文章。

齐轩, 本科生, 清华大学交叉信息院 2024 年 3 月 – 至今

蔡艺硕, 本科生, 中南大学 → 北京大学 (博士) 2024 年 4 月 – 2024 年 10 月

周子安, 本科生, 清华大学致理书院 (信息与计算科学方向) 2023 年 12 月 – 2024 年 5 月

张天祺, 本科生, 清华大学 → 清华大学 (硕士) 2023 年 6 月 – 2024 年 4 月

Brian S. Lin (蔡诗怀), 本科生, 清华大学 → 清华大学 (硕士) 2023 年 6 月 – 2024 年 4 月

杨殊鉴, 硕士生, 上海交通大学巴黎卓越工程师学院 2023 年 9 月 – 2024 年 2 月

党星宇, 本科生, 清华大学交叉信息院 2021 年 10 月 – 2022 年 7 月

技能和专长

- **科研技能:** 我在深度学习、数据分析方面富有经验; 我亦具有计算机系统、应用密码学和软件工程的基本功底。
- **其它专长:** 我拥有良好的沟通技巧, 善于与人合作; 我在指导学生方面富有经验; 我在团队中领导、协作能力强。
- **英语水平:** 托福 106 (2021 年 5 月)。

- 最后更新于 Thursday 15th May, 2025 -