刘瑷玮

手机：18851830977 | 邮箱：liuaw20@mails.tsinghua.edu.cn | 网站：exlaw.github.io

👤 个人简介

我是清华大学软件学院在读博士生（即将毕业），主要研究大语言模型的可靠性与安全性。研究成果已在顶级会议发表多篇论文，获得400 余次学术引用。目前主要关注以下三个研究方向：

🏷️ **大语言模型水印技术**：致力于在模型输出中嵌入可检测的特征标记，实现内容溯源与版权保护，防范模型滥用。

🛡️ **模型安全性提升**：通过对齐技术与红队测试等方法增强模型安全性，系统性地发现和解决潜在安全隐患。

🗃️ **自然语言转 SQL**：研究如何提高自然语言到 SQL 的转换准确性，重点解决复杂数据库架构下的查询转换问题。

🎓 教育背景

🏛️ 清华大学软件学院博士，软件工程专业，导师：闻立杰副教授　　　　　　2020.09 - 至今
🏛️ 南京大学软件学院本科，软件工程专业　　　　　　　　　　　　　　　2016.09 - 2020.07

GPA：4.6 / 5.00　　　排名：5/220

💼 实习经历

🏛️ 伊利诺伊大学芝加哥分校 | **BDSC Lab**　访问学者　　　　　　　　2024.7 - 至今
- 导师：Philip S. Yu (ACM Fellow, IEEE Fellow)
- 主要贡献：研究带水印大语言模型的隐私性，特别是用户对其的可识别性

🏛️ 香港中文大学 | **MISC Lab**　访问学者　　　　　　　　　　　　　2023.7 - 2024.5
- 导师：Irwin King (IEEE Fellow)
- 主要贡献：开发了一种用于大语言模型的不可伪造的公开可验证水印，发表在 ICLR 2024。并撰写了一篇关于大语言模型时代文本水印的综述论文，已被 ACM Computing Surveys 接收。

 苹果公司 | **AIML** 团队　研究实习生　　　　　　　　　　　　　　2023.3 - 2024.9
- 导师：曹蒙博士
- 主要贡献：1) 开发了基于大语言模型的自动属性识别方法，用于提示词难度评估 2) 提出了一种无需人工标注偏好数据的大语言模型安全对齐方法，发表在 ACL 2024 3) 提出 TIS DPO：基于估计权重的令牌级重要性采样直接偏好优化方法

🏷️ 科研成果（大模型水印方向）

🔗 **A Semantic invariant Robust Watermark for Large Language Models**　　2024
- *Aiwei Liu*, Leyi Pan, Xuming Hu, Shiao Meng, Lijie Wen
- **ICLR 2024** (谷歌学术计算机科学会议排名第**3**位)

🔗 **An Unforgeable Publicly Verifiable Watermark for Large Language Models**　2024
- *Aiwei Liu*, Leyi Pan, Xuming Hu, Shu'ang Li, Lijie Wen, Irwin King, Philip S. Yu
- **ICLR 2024** (谷歌学术计算机科学会议排名第**3**位)

🔗 **A Survey of Text Watermarking in the Era of Large Language Models**　　2024
- *Aiwei Liu**, Leyi Pan*, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, Philip S. Yu
- **ACM Computing Surveys** (影响因子: **23.8**, 计算机科学理论与方法领域排名第 **1/143** 位)

⌘ MarkLLM: An Open-Source Toolkit for LLM Watermarking   2024

- *Leyi Pan,* ***Aiwei Liu***[†] *(项目负责人), Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, Philip S. Yu*
- 已被 EMNLP 2024 Demo Track 接收
- GitHub 仓库已获得超过 250 颗星标，展示了显著的社区影响力

⌘ An Entropy-based Text Watermarking Detection Method   2024

- *Yijian Lu,* ***Aiwei Liu****, Dianzhi Yu, Jingjing Li, Irwin King*
- 已被 ACL 2024 接收

⌘ On the Cross-lingual Consistency of Text Watermark for Large Language Models   2024

- *Zhiwei He, Binglin Zhou, Hongkun Hao,* ***Aiwei Liu****, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, Rui Wang*
- 已被 ACL 2024 接收 (口头报告)

⌘ Can Watermarked LLMs be Identified by Users via Crafted Prompts?   2024

- ***Aiwei Liu****, Sheng Guan, Yiming Liu, Leyi Pan, Yifei Zhang, Liancheng Fang, Lijie Wen, Philip S. Yu, Xuming Hu*
- arXiv 预印本，已投稿至 **ICLR 2025**

⛊ 科研成果（大模型安全性提升方向）

⌘ Direct Large Language Model Alignment Through Self-Rewarding Contrastive Prompt Distillation   2024

- ***Aiwei Liu****, Haoping Bai, Zhiyun Lu, Xiang Kong, Simon Wang, Jiulong Shan, Meng Cao, Lijie Wen*
- 已被 ACL 2024 接收

⌘ Character-level White-Box Adversarial Attacks against Transformers via Attachable Subwords Substitution   2022

- ***Aiwei Liu****, Honghai Yu, Xuming Hu, Shu'ang Li, Li Lin, Fukun Ma, Yawen Yang, Lijie Wen*
- 已被 EMNLP 2022 接收

⌘ TIS-DPO: Token-level Importance Sampling for Direct Preference Optimization With Estimated Weights   2024

- ***Aiwei Liu****, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Xiaoming Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, Philip S. Yu, Meng Cao*
- arXiv 预印本，已投稿至 **ICLR 2025**

⛁ 科研成果（自然语言转 SQL 方向）

⌘ Semantic Enhanced Text-to-SQL Parsing via Iteratively Learning Schema Linking Graph   2022

- ***Aiwei Liu****, Xuming Hu, Li Lin, Lijie Wen*
- 已被 SIGKDD 2022 接收

⌘ Exploring the Compositional Generalization in Context Dependent Text-to-SQL Parsing 2023

- ***Aiwei Liu****, Wei Liu, Xuming Hu, Shuang Li, Fukun Ma, Yawen Yang, Lijie Wen*
- 已被 ACL 2023 Findings 接收

⌘ A Comprehensive Evaluation of ChatGPT's Zero-shot Text-to-SQL Capability   2023

- ***Aiwei Liu****, Xuming Hu, Lijie Wen, Philip S Yu*
- 该工作已获得超过100 次引用，展示了显著的学术影响力

奖项与荣誉

- 🏆 清华之友-沈阳浑南英才奖学金（二等） 2022
- 🏆 南京大学优秀毕业生 2020
- 🏆 中国电子科技集团奖学金 2019
- 🏆 南京大学优秀共青团干 2019
- 🏆 国家奖学金 2018
- 🏆 海南航空奖学金 2017

🎙 学术服务

会议评审

- The International Conference on Learning Representations (ICLR)
- The Annual Meeting of the Association for Computational Linguistics (ACL)
- The Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)
- The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)
- The Annual Conference of the European Chapter of the Association for Computational Linguistics (EACL)
- The ACM WWW International World Wide Web Conference (WWW)
- The ACM International Conference on Multimedia (MM)

工作坊（workshop）组织

- AAAI 2025 Workshop on Preventing and Detecting LLM Generated Misinformation (PDLM)联合组织者

📖 教学经历

大语言模型生成误导信息的预防与检测 July 2024
SIGIR 2024 Tutorial

- 作为主讲人在第 47 届国际 ACM SIGIR 会议上进行大语言模型生成误导信息预防与检测技术的教程讲解
- 教程网站：https://sigir24-llm-misinformation.github.io/

创新人才与大学文化 2021
助教，清华大学

- 协助课程教学，组织学生讨论，为课程相关项目提供支持

操作系统 2018
助教，南京大学

- 指导实验课程，批改作业，为学生理解复杂的操作系统概念提供一对一辅导

其他发表论文

🔗 **GDA: Generative Data Augmentation Techniques for Relation Extraction Tasks** 2023

- *Xuming Hu*, **Aiwei Liu*** (共同一作), Zeqi Tan, Xin Zhang, Chenwei Zhang, Irwin King, Philip S. Yu*
- 已被 ACL 2023 Findings 接收

🔗 **RAPL: A Relation-Aware Prototype Learning Approach for Few-Shot Document-Level Relation Extraction** 2023

- *Shiao Meng, Xuming Hu, **Aiwei Liu**, Shu'ang Li, Fukun Ma, Yawen Yang, Lijie Wen*
- 已被 EMNLP 2023 接收

🔗 **AMR-based network for aspect-based sentiment analysisn** 2023

- *Fukun Ma, Xuming Hu, **Aiwei Liu**, Yawen Yang, Philip S. Yu, Lijie Wen*
- 已被 ACL 2023 接收

⚲ CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking   2022
- *Xuming Hu, Zhijiang Guo, Guanyu Wu, **Aiwei Liu**, Lijie Wen, Phlip S.Yu*
- 已被 NAACL 2022 接收

⚲ A Multi-level Supervised Contrastive Learning Framework for Low-Resource Natural Language Inference   2023
- *Shu'ang Li, Xuming Hu, Li Lin, **Aiwei Liu**, Lijie Wen, Philip S. Yu*
- IEEE/ACM Transactions on Audio, Speech, and Language Processing

⚲ Improving Open Relation Extraction With Search Documents Under Self-Supervisions
2023
- *Xuming Hu, Zhaochen Hong, Chenwei Zhang, **Aiwei Liu**, Shiao Meng, Lijie Wen, Irwin King, Philip S. Yu*
- IEEE Transactions on Knowledge and Data Engineering

⚲ Refiner: Restructure Retrieval Content Efficiently to Advance Question-Answering Capabilities   2024
- *Zhonghao Li, Xuming Hu, **Aiwei Liu**, Kening Zheng, Sirui Huang, Hui Xiong*
- 已被 EMNLP 2024 Findings 接收

⚲ Entity-to-Text based Data Augmentation for various Named Entity Recognition Tasks
2023
- *Xuming Hu, Yong Jiang, **Aiwei Liu**, Zhongqiang Huang, Pengjun Xie, Fei Huang, Lijie Wen, Philip S. Yu*
- 已被 ACL 2023 Findings 接收

⚲ Enhancing Cross-lingual Natural Language Inference by Soft Prompting with Multilingual Verbalizer   2023
- *Shuang Li, Xuming Hu, **Aiwei Liu**, Yawen Yang, Fukun Ma, Philip S. Yu, Lijie Wen*
- 已被 ACL 2023 Findings 接收

⚲ Gaussian prior reinforcement learning for nested named entity recognition   2023
- *Yawen Yang, Xuming Hu, Fukun Ma, Shuang Li, **Aiwei Liu**, Lijie Wen, Philip S. Yu*
- 已被 ICASSP 接收

⚲ On the Robustness of Document-Level Relation Extraction Models to Entity Name Variations   2024
- *Shiao Meng, Xuming Hu, **Aiwei Liu**, Fukun Ma, Yawen Yang, Shuang Li, Lijie Wen*
- 已被 ACL 2024 Findings 接收

其他预印本论文

⚲ WaterSeeker: Efficient Detection of Watermarked Segments in Large Documents   2024
- *Leyi Pan, **Aiwei Liu**, Yijian Lu, Zitian Gao, Yichen Di, Lijie Wen, Irwin King, Philip S. Yu*
- 已投稿至 NAACL 2025

⚲ Interpretable Contrastive Monte Carlo Tree Search Reasoning   2024
- *Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, **Aiwei Liu**, Xuming Hu, Lijie Wen*
- 已投稿至 ICLR 2025

⚲ Entropy-Based Decoding for Retrieval-Augmented Large Language Models   2024
- *Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, **Aiwei Liu**, Irwin King*
- NeurIPS 2024 - Workshop on Foundation Model Interventions

⚲ ChatCite: LLM Agent with Human Workflow Guidance for Comparative Literature Summary

2024

- *Yutong Li, Lu Chen, **Aiwei Liu**, Kai Yu, Lijie Wen*
- Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges