刘瑗玮

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安全性与可靠性。研究成果已在顶级会议发表多篇论文，获得600 余次学术引用，目前主要关注以下三个研究方向：

- 大模型安全与对齐：通过创新的对齐技术（如自奖励对比学习）和系统的红队测试方法增强模型安全性，研究高效的偏好优化算法（如 TIS-DPO）。
- 大语言模型水印技术：致力于设计不可伪造的公开可验证水印和语义不变水印，实现内容溯源与版权保护，防范模型滥用。
- 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 教授 (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 教授 (ACM Fellow, 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

第一作者论文汇总

如果您对我作为第一作者的研究贡献感兴趣（这对某些机构的招聘可能很重要），以下是我第一作者论文发表的汇总 (不包含共同一作排名靠后的论文)：

- 顶级会议论文： **4** 篇 ICLR 论文，1 篇 ACL 论文，1 篇 SIGKDD 论文，1 篇 EMNLP 论文，以及 1 篇 ACL(Findings) 论文
- 期刊论文： 1 篇 ACM Computing Surveys 论文（影响因子：23.8）

核心研究项目

TIS-DPO: Token-level Importance Sampling for Direct Preference Optimization With Estimated Weights 2025

- ***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*
- **ICLR 2025** (谷歌学术计算机科学会议排名第**3**位)

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

Can Watermarked LLMs be Identified by Users via Crafted Prompts? 2025

- ***Aiwei Liu***, *Sheng Guan, Yiming Liu, Leyi Pan, Yifei Zhang, Liancheng Fang, Lijie Wen, Philip S. Yu, Xuming Hu*
- **ICLR 2025** (谷歌学术计算机科学会议排名第**3**位)

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 300 颗星标，11 位开源贡献者，展示了显著的社区影响力

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

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-途游奖学金（一等） 2024
- 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标兵 2019
- 国家奖学金 2018
- 海南航空奖学金 2017

学术服务

会议审稿人

- 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

全部会议论文

Can Watermarked LLMs be Identified by Users via Crafted Prompts? 2025

- ***Aiwei Liu***, *Sheng Guan, Yiming Liu, Leyi Pan, Yifei Zhang, Liancheng Fang, Lijie Wen, Philip S. Yu, Xuming Hu*
- 已被 ICLR 2025 接收

TIS-DPO: Token-level Importance Sampling for Direct Preference Optimization With Estimated Weights 2025

- ***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*
- 已被 ICLR 2025 接收

Mitigating Modality Prior-induced Hallucinations in Multimodal Large Language Models via Deciphering Attention Causality 2025

- *Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang,* ***Aiwei Liu****, Xuming Hu*
- 已被 ICLR 2025 接收

WaterSeeker: Efficient Detection of Watermarked Segments in Large Documents 2025

- *Leyi Pan,* ***Aiwei Liu****, Yijian Lu, Zitian Gao, Yichen Di, Lijie Wen, Irwin King, Philip S. Yu*
- 已被 NAACL 2025 Findings 接收

Entropy-Based Decoding for Retrieval-Augmented Large Language Models 2025

- *Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li,* ***Aiwei Liu****, Irwin King*
- 已被 NAACL 2025 接收

ChatCite: LLM Agent with Human Workflow Guidance for Comparative Literature Summary 2024

- *Yutong Li, Lu Chen,* ***Aiwei Liu****, Kai Yu, Lijie Wen*
- 已被 COLING 2024 接收

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

MarkLLM: An Open-Source Toolkit for LLM Watermarking 2024

- *Leyi Pan,* ***Aiwei Liu****, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King*
- 已被 EMNLP 2024 Demo 接收

Preventing and Detecting Misinformation Generated by Large Language Models 2024

- ***Aiwei Liu****, Qiang Sheng, Xuming Hu*
- SIGIR 2024 Tutorial

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

An Entropy-based Text Watermarking Detection Method 2024

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

Can Watermarks Survive Translation? 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

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

A Semantic Invariant Robust Watermark for Large Language Models 2024

- ***Aiwei Liu****, Leyi Pan, Xuming Hu, Shiao Meng, Lijie Wen*
- 已被 ICLR 2024 接收

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*
- 已被 ICLR 2024 接收

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

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

Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction 2023

- *Xuming Hu, Junzhe Chen,* ***Aiwei Liu****, Shiao Meng, Lijie Wen, Philip S Yu*
- 已被 MM 2023 接收

EnTDA: Entity-to-Text based Data Augmentation with Semantic Coherence and Entity Preserving for various NER 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

GDA: Generative Data Augmentation Techniques for Relation Extraction Tasks 2023

- *Xuming Hu,* ***Aiwei Liu****, Zeqi Tan, Xin Zhang, Chenwei Zhang, Irwin King, Philip S. Yu*
- 已被 ACL 2023 Findings 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

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

Semantics Matters: AMR-based Path Aggregation Relational Network for Aspect-based Sentiment Analysis 2023

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

Gaussian Prior Reinforcement Learning for Nested Named Entity Recognition 2023

- *Yawen Yang, Xuming Hu, Fukun Ma, Shu'ang Li,* ***Aiwei Liu****, Lijie Wen, Philip S. Yu*
- 已被 ICASSP 2023 接收

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

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

CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking 2022

- *Xuming Hu, Zhijiang Guo, Guanyu Wu,* ***Aiwei Liu****, Lijie Wen, Philip S. Yu*
- 已被 NAACL 2022 接收

全部期刊论文

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*
- 发表于 ACM Computing Surveys

Reading Broadly to Open Your Mind: 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*
- 发表于 TKDE

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*
- 发表于 TASLP 2023

全部预印本

TabGEN-RAG: Iterative Retrieval for Tabular Data Generation with Large Language Models 2024

- *Liancheng Fang,* ***Aiwei Liu****, Hengrui Zhang, Henry Peng Zou, Weizhi Zhang, Philip S. Yu*
- 发表于 TRL@NeurIPS 2024 Workshop

A Survey of AIOps for Failure Management in the Era of Large Language Models 2024

- *Lingzhe Zhang, Tong Jia, Mengxi Jia, Yifan Wu,* ***Aiwei Liu****, Yong Yang, Zhonghai Wu, Xuming Hu, Philip S. Yu, Ying Li*

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

- ***Aiwei Liu****, Xuming Hu, Lijie Wen, Philip S. Yu*

Interpretable Contrastive Monte Carlo Tree Search Reasoning 2024

- *Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu,* ***Aiwei Liu****, Xuming Hu, Lijie Wen*

Recent Advances of Multimodal Continual Learning: A Comprehensive Survey 2024

- *Dianzhi Yu, Xinni Zhang, Yankai Chen,* ***Aiwei Liu****, Yifei Zhang, Philip S Yu, Irwin King*

Less is More: Extreme Gradient Boost Rank-1 Adaption for Efficient Finetuning of LLMs 2024

- *Yifei Zhang, Hao Zhu,* ***Aiwei Liu****, Han Yu, Piotr Koniusz, Irwin King*

Exploring Response Uncertainty in MLLMs: An Empirical Evaluation Under Misleading Scenarios 2024

- *Yunkai Dang, Mengxi Gao, Yibo Yan, Xin Zou, Yanggan Gu,* ***Aiwei Liu****, Xuming Hu*

Cold-Start Recommendation towards the Era of Large Language Models (LLMs): A Comprehensive Survey and Roadmap 2025

- *Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou,* ***Aiwei Liu****, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, Feiran Huang, Sheng Zhou, Jiajun Bu, Allen Lin, James Caverlee, Fakhri Karray, Irwin King, Philip S Yu*