

# Aiwei Liu

☎ +86 18851830977 | @ liuaw20@mails.tsinghua.edu.cn | 🌐 exlaw.github.io | 📍 Tsinghua University, Beijing, China

## 👤 ABOUT ME

I am a **final-year Ph.D. candidate** at Tsinghua University, specializing in robust and trustworthy **large language models (LLMs)**. My research has made significant contributions to the field, as evidenced by over **400 citations**. Specifically, I focus on three critical areas:

- 🔹 **LLM Watermarking:** Embedding detectable features in LLM outputs to protect copyright and prevent the misuse of LLMs.
- 🔹 **LLM Safety Improvement:** Enhancing LLM safety through alignment techniques and red-teaming methods, including adversarial testing, to improve safety against potential vulnerabilities.
- 🔹 **Robust Natural Language to SQL:** Enhancing the reliability and accuracy of converting natural language queries into SQL, particularly for complex database schemas.

## 🎓 EDUCATION

<b>🏛️ Tsinghua University</b> <i>Ph.D. in Software Engineering; Advisor: Associate Professor Lijie Wen</i>	Beijing, China <i>Sep 2020 – present</i>
<b>🏛️ Nanjing University</b> <i>B.Eng. in Software Engineering; GPA: 4.6/5.0, ranking: 5/220</i>	Nanjing, China <i>Sep 2016 – Jun 2020</i>

## 🔑 SELECTED WORKS IN LLM WATERMARKING

- 🔑 **A Semantic invariant Robust Watermark for Large Language Models**
  - Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, Lijie Wen
  - **ICLR 2024** (Ranked **3** in all computer science conferences by Google Scholar)
- 🔑 **An Unforgeable Publicly Verifiable Watermark for Large Language Models**
  - Aiwei Liu, Leyi Pan, Xuming Hu, Shu'ang Li, Lijie Wen, Irwin King, Philip S. Yu
  - **ICLR 2024** (Ranked **3** in all computer science conferences by Google Scholar)
- 🔑 **A Survey of Text Watermarking in the Era of Large Language Models**
  - **Aiwei Liu\***, Leyi Pan\*, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, Philip S. Yu
  - **ACM Computing Surveys** (Impact Factor: **23.8**, ranked **1/143** in Computer Science Theory & Methods)
- 🔑 **MarkLLM: An Open-Source Toolkit for LLM Watermarking**
  - Leyi Pan, **Aiwei Liu**, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, Philip S. Yu
  - In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Demo)
  - **GitHub repository** has garnered over 250 stars, demonstrating significant community interest and impact.
- 🔑 **An Entropy-based Text Watermarking Detection Method**
  - Yijian Lu, **Aiwei Liu**, Dianzhi Yu, Jingjing Li, Irwin King
  - In Proceedings of Association for Computational Linguistics (ACL), 2024
- 🔑 **On the Cross-lingual Consistency of Text Watermark for Large Language Models**
  - Zhiwei He, Binglin Zhou, Hongkun Hao, **Aiwei Liu**, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, Rui Wang
  - In Proceedings of Association for Computational Linguistics (ACL), 2024 (**Oral**)
- 🔑 **Can Watermarked LLMs be Identified by Users via Crafted Prompts?**
  - **Aiwei Liu**, Sheng Guan, Yiming Liu, Leyi Pan, Yifei Zhang, Liancheng Fang, Lijie Wen, Philip S. Yu, Xuming Hu
  - arXiv preprint, submitted to **ICLR 2025**

## 🏆 SELECTED WORKS IN LLM SAFETY IMPROVEMENT

- 🔗 **Direct Large Language Model Alignment Through Self-Rewarding Contrastive Prompt Distillation**
  - **Aiwei Liu**, Haoping Bai, Zhiyun Lu, Xiang Kong, Simon Wang, Jiulong Shan, Meng Cao, Lijie Wen
  - In Proceedings of Association for Computational Linguistics (ACL), 2024
- 🔗 **Character-level White-Box Adversarial Attacks against Transformers via Attachable Subwords Substitution**
  - **Aiwei Liu**, Honghai Yu, Xuming Hu, Shu'ang Li, Li Lin, Fukun Ma, Yawen Yang, Lijie Wen
  - In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022
- 🔗 **TIS-DPO: Token-level Importance Sampling for Direct Preference Optimization With Estimated Weights**
  - **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 preprint, submitted to **ICLR 2025**

## 📚 SELECTED WORKS IN ROBUST NATURAL LANGUAGE TO SQL

- 🔗 **Semantic Enhanced Text-to-SQL Parsing via Iteratively Learning Schema Linking Graph**
  - **Aiwei Liu**, Xuming Hu, Li Lin, Lijie Wen
  - In Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2022
- 🔗 **Exploring the Compositional Generalization in Context Dependent Text-to-SQL Parsing**
  - **Aiwei Liu**, Wei Liu, Xuming Hu, Shuang Li, Fukun Ma, Yawen Yang, Lijie Wen
  - In The Findings of Association for Computational Linguistics (Findings of ACL), 2023
- 🔗 **A Comprehensive Evaluation of ChatGPT's Zero-shot Text-to-SQL Capability**
  - **Aiwei Liu**, Xuming Hu, Lijie Wen, Philip S Yu
  - With over **100 citations**, this work has demonstrated significant impact in the field.

## 👜 WORK EXPERIENCE

### 🏛️ University of Illinois Chicago

*Visiting Scholar; Advisor: Prof **Philip S. Yu** (ACM Fellow, IEEE Fellow)*

Chicago, USA

*July 2024 – Present*

- Project: Privacy of Large Language Models
- Investigated the privacy of watermarked LLMs, specifically their identifiability by users.

### 🏛️ The Chinese University of Hong Kong

*Visiting Scholar Advisor: Prof **Irwin King** (IEEE Fellow)*

Hong Kong, China

*July 2023 – May 2024*

- Project: Watermark for Large Language Models
- 1) Developed an unforgeable publicly verifiable watermark for Large Language Models 2) Write a comprehensive survey about the text watermarking in the era of LLMs.

### 🍏 Apple AIML Group

*Research Intern: Mentored by Dr. Meng Cao*

Beijing, China

*Mar 2023 – Sep 2024*

- Project: Prompt Difficulty Evaluation, Safety alignment for LLM
- 1) Developed a LLM-based automatic attributes identification methods for prompt difficulty evaluation. 2) A safety alignment method for LLM that does not require manual annotation of preference data. 3) TIS DPO: Token-level Importance Sampling for Direct Preference Optimization With Estimated Weights

## 🏆 SELECTED AWARDS

Tsinghua Excellent Student Award (SecondClass)	2022
Outstanding Graduates Nanjing University	2020
China Electronics Technology Group Scholarship	2019
National Scholarship	2018
Hainan Airlines Scholarship	2017

### Program Committee/ Reviewer

- 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)

### Workshop Organization

- Co-organizer of the [AAAI 2025 Workshop on Preventing and Detecting LLM Generated Misinformation \(PDLM\)](#)

### 📖 TEACHING EXPERIENCE

---

#### Preventing and Detecting Misinformation Generated by Large Language Models July 2024

*SIGIR 2024 Tutorial*

- **Lead presenter** for a tutorial on techniques for preventing and detecting LLM-generated misinformation at the 47th International ACM SIGIR Conference.
- Tutorial website: <https://sigir24-llm-misinformation.github.io/>

#### Innovation Talent and University Culture 2021

*Teaching Assistant, Tsinghua University*

- Assisted in course delivery, facilitated student discussions, and provided support for course-related projects

#### Operating Systems 2018

*Teaching Assistant, Nanjing University*

- Conducted lab sessions, graded assignments, and provided one-on-one support to students in understanding complex OS concepts

### OTHER PUBLICATIONS

---

#### 🔗GDA: Generative Data Augmentation Techniques for Relation Extraction Tasks

- Xuming Hu\*, **Aiwei Liu\***, Zeqi Tan, Xin Zhang, Chenwei Zhang, Irwin King, Philip S. Yu
- In The Findings of Association for Computational Linguistics (Findings of ACL), 2023

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

- Shiao Meng, Xuming Hu, **Aiwei Liu**, Shu'ang Li, Fukun Ma, Yawen Yang, Lijie Wen
- In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023

#### 🔗AMR-based network for aspect-based sentiment analysis

- Fukun Ma, Xuming Hu, **Aiwei Liu**, Yawen Yang, Philip S. Yu, Lijie Wen
- In Proceedings of Association for Computational Linguistics (ACL), 2023

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

- Xuming Hu, Zhijiang Guo, Guanyu Wu, **Aiwei Liu**, Lijie Wen, Philip S. Yu
- In The 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2022

#### 🔗A Multi-level Supervised Contrastive Learning Framework for Low-Resource Natural Language Inference

- 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

- Xuming Hu, Zhaochen Hong, Chenwei Zhang, **Aiwei Liu**, Shiao Meng, Lijie Wen, Irwin King, Philip S. Yu
- IEEE/ACM Transactions on Audio, Speech, and Language Processing

🔗 **Refiner: Restructure Retrieval Content Efficiently to Advance Question-Answering Capabilities**

- Zhonghao Li, Xuming Hu, **Aiwei Liu**, Kening Zheng, Sirui Huang, Hui Xiong
- In The Findings of Empirical Methods in Natural Language Processing (EMNLP), 2024

🔗 **Entity-to-Text based Data Augmentation for various Named Entity Recognition Tasks**

- Xuming Hu, Yong Jiang, **Aiwei Liu**, Zhongqiang Huang, Pengjun Xie, Fei Huang, Lijie Wen, Philip S. Yu
- In The Findings of Association for Computational Linguistics (Findings of ACL), 2023

🔗 **Enhancing Cross-lingual Natural Language Inference by Soft Prompting with Multilingual Verbalizer**

- Shuang Li, Xuming Hu, **Aiwei Liu**, Yawen Yang, Fukun Ma, Philip S. Yu, Lijie Wen
- In The Findings of Association for Computational Linguistics (Findings of ACL), 2023

🔗 **Gaussian prior reinforcement learning for nested named entity recognition**

- Yawen Yang, Xuming Hu, Fukun Ma, Shuang Li, **Aiwei Liu**, Lijie Wen, Philip S. Yu
- IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

🔗 **On the Robustness of Document-Level Relation Extraction Models to Entity Name Variations**

- Shiao Meng, Xuming Hu, **Aiwei Liu**, Fukun Ma, Yawen Yang, Shuang Li, Lijie Wen
- In The Findings of Association for Computational Linguistics (Findings of ACL), 2024

OTHER PREPRINTS

---

🔗 **WaterSeeker: Efficient Detection of Watermarked Segments in Large Documents**

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

🔗 **Interpretable Contrastive Monte Carlo Tree Search Reasoning**

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

🔗 **Entropy-Based Decoding for Retrieval-Augmented Large Language Models**

- Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, **Aiwei Liu**, Irwin King

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

- Yutong Li, Lu Chen, **Aiwei Liu**, Kai Yu, Lijie Wen