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 600 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 Alignment: Developing novel alignment techniques including self-rewarding contrastive learning and token-level importance sampling to enhance LLM's adherence to human values and preferences without manual annotation.

Robust Natural Language to SQL: Enhancing the reliability and accuracy of converting natural language queries into SQL, particularly for complex database schemas.

Education

Tsinghua University — Beijing, China
*Ph.D. in Software Engineering; Advisor: Associate Professor **Lijie Wen*** — *Sep 2020 – present*

Nanjing University — Nanjing, China
B.Eng. in Software Engineering; GPA: 4.6/5.0, ranking: 5/220 — *Sep 2016 – Jun 2020*

Work Experience

University of Illinois Chicago — Chicago, USA
*Visiting Scholar; Advisor: Prof **Philip S. Yu** (ACM Fellow, IEEE Fellow)* — *July 2024 – Dec 2024*

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

The Chinese University of Hong Kong — Hong Kong, China
*Visiting Scholar Advisor: Prof **Irwin King** (ACM Fellow, IEEE Fellow)* — *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 — Beijing, China
Research Intern: Mentored by Dr. Meng Cao — *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

First-author Publication Summary

If you are interested in my **first-author** research contributions (maybe important to some institutions in hiring), here is a summary of my **first-author** publications:

Top-tier Conference Papers: 4 papers in ICLR (2024, 2025), 1 paper in ACL, 1 paper in KDD, 1 paper in EMNLP, and 1 paper in ACL(Findings) as first author

Journal Papers: 1 paper in ACM Computing Surveys (Impact Factor: 23.8) as first author

★ Research Highlights

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
- **ICLR 2025** (Ranked **3** in all computer science conferences by Google Scholar)

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
- **ICLR 2025** (Ranked **3** in all computer science conferences by Google Scholar)

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)

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
- **ACL 2024**

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**[†] (Project Lead), Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, Philip S. Yu
- **EMNLP 2024 Demo**
- **GitHub repository** has garnered over 300 stars, demonstrating significant community interest and impact.

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
- **EMNLP 2022**

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

- **Aiwei Liu**, Xuming Hu, Li Lin, Lijie Wen
- **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
- **ACL 2023 Findings**

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

- **Aiwei Liu**, Xuming Hu, Lijie Wen, Philip S Yu
- With over **120 citations**, this work has demonstrated significant impact in the field.

Selected Awards

Tsinghua Excellent Student Award (First Class)	*2024*
Tsinghua Excellent Student Award (SecondClass)	*2022*
Outstanding Graduates Nanjing University	*2020*
China Electronics Technology Group Scholarship	*2019*
National Scholarship	*2018*
Hainan Airlines Scholarship	*2017*

Professional Services

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

Conference Papers

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
- Accepted at ICLR 2025

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
- Accepted at ICLR 2025

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

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

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
- Accepted at NAACL 2025 Findings

Entropy-Based Decoding for Retrieval-Augmented Large Language Models

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

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

- Yutong Li, Lu Chen, **Aiwei Liu**, Kai Yu, Lijie Wen
- Accepted at COLING 2024

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

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
- Accepted at EMNLP 2024 Demo

Preventing and Detecting Misinformation Generated by Large Language Models
- **Aiwei Liu**, Qiang Sheng, Xuming Hu
- SIGIR 2024 Tutorial

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
- Accepted at ACL 2024 Findings

An Entropy-based Text Watermarking Detection Method
- Yijian Lu, **Aiwei Liu**, Dianzhi Yu, Jingjing Li, Irwin King
- Accepted at 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
- Accepted at ACL 2024

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
- Accepted at ACL 2024

A Semantic Invariant Robust Watermark for Large Language Models
- **Aiwei Liu**, Leyi Pan, Xuming Hu, Shiao Meng, Lijie Wen
- Accepted at ICLR 2024

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
- Accepted at ICLR 2024

RAPL: A Relation-Aware Prototype Learning Approach for Few-Shot Document-Level Relation Extraction
- Shiao Meng, Xuming Hu, **Aiwei Liu**, Shuang Li, Fukun Ma, Yawen Yang, Lijie Wen
- Accepted at EMNLP 2023

Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction
- Xuming Hu, Junzhe Chen, **Aiwei Liu**, Shiao Meng, Lijie Wen, Philip S Yu
- Accepted at MM 2023

EnTDA: Entity-to-Text based Data Augmentation with Semantic Coherence and Entity Preserving for various
- Xuming Hu, Yong Jiang, **Aiwei Liu**, Zhongqiang Huang, Pengjun Xie, Fei Huang, Lijie Wen, Philip S. Yu
- Accepted at ACL 2023 Findings

GDA: Generative Data Augmentation Techniques for Relation Extraction Tasks
- Xuming Hu, **Aiwei Liu**, Zeqi Tan, Xin Zhang, Chenwei Zhang, Irwin King, Philip S. Yu
- Accepted at ACL 2023 Findings

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
- Accepted at ACL 2023 Findings

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
- Accepted at ACL 2023 Findings

Semantics Matters: AMR-based Path Aggregation Relational Network for Aspect-based Sentiment Analysis
- Fukun Ma, Xuming Hu, **Aiwei Liu**, Yawen Yang, Shuang Li, Philip S. Yu, Lijie Wen
- Accepted at ACL 2023

Gaussian Prior Reinforcement Learning for Nested Named Entity Recognition
- Yawen Yang, Xuming Hu, Fukun Ma, Shu'ang Li, **Aiwei Liu**, Lijie Wen, Philip S. Yu
- Accepted at ICASSP 2023

Semantic Enhanced Text-to-SQL Parsing via Iteratively Learning Schema Linking Graph
- **Aiwei Liu**, Xuming Hu, Li Lin, Lijie Wen
- Accepted at SIGKDD 2022

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
- Accepted at EMNLP 2022

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

Journal Papers

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
- Published in ACM Computing Surveys, 2024

Reading Broadly to Open Your Mind: 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
- Published in TKDE, 2023

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
- Published in TASLP, 2023

Preprints

TabGEN-RAG: Iterative Retrieval for Tabular Data Generation with Large Language Models
- Liancheng Fang, **Aiwei Liu**, Hengrui Zhang, Henry Peng Zou, Weizhi Zhang, Philip S. Yu
- Submitted to TRL@NeurIPS 2024 Workshop

A Survey of AIOps for Failure Management in the Era of Large Language Models
- Lingzhe Zhang, Tong Jia, Mengxi Jia, Yifan Wu, **Aiwei Liu**, Yong Yang, Zhonghai Wu, Xuming Hu, Philip S. Yu, Ying Li
- Preprint

A Comprehensive Evaluation of ChatGPT's Zero-Shot Text-to-SQL Capability
- **Aiwei Liu**, Xuming Hu, Lijie Wen, Philip S. Yu
- Preprint

Interpretable Contrastive Monte Carlo Tree Search Reasoning
- Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, **Aiwei Liu**, Xuming Hu, Lijie Wen
- Preprint

Recent Advances of Multimodal Continual Learning: A Comprehensive Survey
- Dianzhi Yu, Xinni Zhang, Yankai Chen, **Aiwei Liu**, Yifei Zhang, Philip S Yu, Irwin King
- Preprint

Less is More: Extreme Gradient Boost Rank-1 Adaption for Efficient Finetuning of LLMs
- Yifei Zhang, Hao Zhu, **Aiwei Liu**, Han Yu, Piotr Koniusz, Irwin King
- Preprint

Exploring Response Uncertainty in MLLMs: An Empirical Evaluation Under Misleading Scenarios
- Yunkai Dang, Mengxi Gao, Yibo Yan, Xin Zou, Yanggan Gu, **Aiwei Liu**, Xuming Hu
- Preprint

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