Tiejin Chen

734-210-2624 tiejin@asu.edu Website Google Scholar

Education

Arizona State University, Tempe 2023-2028 (Expected)
PhD in Computer Science *Tempe,AZ*

- **Advisor:** Hua Wei; **GPA:** 3.91/4.00
- **Research Interests:** Trustworthy LLMs, including uncertainty quantification, safety, explanation, and their applications, especially improving post-training like Supervised Fine-tuning and Reinforcement Learning.

University of Michigan, Ann Arbor 2021-2023
MS in Applied Statistics *Ann Arbor,MI*

- **GPA:** 3.98/4.00

Sichuan University 2016-2020
Bachelor in Statistics *Chengdu,Sichuan*

- **GPA:** 3.55/4.00 | 86.59/100; **Rank:** 4/33

Publications

[ICML'26] **Tiejin Chen**, Longchao Da, Xiaoou Liu, Hua Wei. *Position: Uncertainty Quantification in LLMs is Just Unsupervised Clustering* . Forty-Third International Conference on Machine Learning - Position Track

[ICML'26] Xiaoou Liu, **Tiejin Chen**, Dengjia Zhang, Yaqing Wang, Lu Cheng, Hua Wei. *Diagnosing Multi-step Reasoning Failures in Black-box LLMs via Stepwise Confidence Attribution* . Forty-Third International Conference on Machine Learning

[ACL'26] **Tiejin Chen**, Huaiyuan Yao, Jia Chen, Evangelos E. Papalexakis, Hua Wei. *Every Response Counts: Quantifying Uncertainty of LLM-based Multi-Agent Systems through Tensor Decomposition*. The 64th Annual Meeting of the Association for Computational Linguistics

[ACL Findings'26] **Tiejin Chen**, Xiaoou Liu, Longchao Da, Jia Chen, Evangelos E. Papalexakis, Hua Wei. *Uncertainty Quantification of Large Language Models through Multi-Dimensional Responses*. Findings of The 64th Annual Meeting of the Association for Computational Linguistics

[IV'26] **Tiejin Chen**, Ahmadreza Moradipari, Kyungtae Han, Hua Wei, Nejib Ammar. *Agentic AI for Trip Planning Optimization Application*. IEEE Intelligent Vehicles Symposium 2026

[EACL'26 - Oral] **Tiejin Chen***, Kaishen Wang*, Hua Wei. *Zer0-Jack: A Memory-efficient Gradient-based Jailbreaking Method for Black-box Multi-modal Large Language Models*. The 19th Conference of the European Chapter of the Association for Computational Linguistics; also presented at GenSafeAI Workshop @ NeurIPS 2024

[EACL Findings'26] **Tiejin Chen**, Xiaoou Liu, Vishnu Nandam, Kuan-Ru Liou, Hua Wei. *Conformal Feedback Alignment: Quantifying Answer-Level Reliability for Robust LLM Alignment*. Findings of The 19th Conference of the European Chapter of the Association for Computational Linguistics

[KDD Explor.] **Tiejin Chen**, Wenwang Huang, Linsey Pang, Dongsheng Luo, Hua Wei. *Are Classification Robustness and Explanation Robustness Really Strongly Correlated? An Analysis Through Input Loss Landscape*. ACM SIGKDD Explorations Newsletter(IF=4.0)

[QCE'25] Ningyi Xie, Xinwei Lee, **Tiejin Chen**, Yoshiyuki Saito, Nobuyoshi Asai, Dongsheng Cai. *An Adaptive Weighted QITE-VQE Algorithm for Combinatorial Optimization Problems*. IEEE International Conference on Quantum Computing and Engineering 2025

[ICCPS'25] Yiran Zhang, Khoa Vo, Longchao Da, **Tiejin Chen**, Xiaoou Liu, Hua Wei. *Reproducible and Low-cost Sim-to-Real Environment for Traffic Signal Control*. ACM/IEEE 16th International Conference on Cyber-Physical Systems

[ACL Findings'25] **Tiejin Chen**, Pingzhi Li, Kaixiong Zhou, Tianlong Chen, Hua Wei. *Unveiling Privacy Risks in Multi-modal Large Language Models: Task-specific Vulnerabilities and Mitigation Challenges*. Findings of the 63rd Annual Meeting of the Association for Computational Linguistics

[ACL Findings'25] **Tiejin Chen**, Pingzhi Li, Kaixiong Zhou, Tianlong Chen, Hua Wei. *Vision Language Model Helps Private Information De-Identification in Vision Data*. Findings of the 63rd Annual Meeting of the Association for Computational Linguistics

[KDD'25] **Tiejin Chen***, Xiaoou Liu*, Longchao Da, Chacha Chen, Zhen Lin, Hua Wei. *Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey*. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining

[RLC'25] Justin Turnau, Longchao Da, Khoa Vo, Ferdous Al Rafi, Shreyas Bachiraju, **Tiejin Chen**, Hua Wei. *Joint-Local Grounded Action Transformation for Sim-to-Real Transfer in Multi-Agent Traffic Control*. Reinforcement Learning Conference 2025

[SDM'25] **Tiejin Chen**, Longchao Da, Huixue Zhou, Pingzhi Li, Kaizhong Zhou, Tianlong Chen, Hua Wei. *Protecting Privacy against Membership Inference Attack with LLM Fine-tuning through Flatness*. SIAM International Conference on Data Mining 2025; short version presented at SeTLLM Workshop @ ICLR 2024

[ICLR Workshop'25] **Tiejin Chen***, Kuan-Ru Liou*, Mithun Shivakoti, Aaryan Gaur, Pragya Kumari, Meiqi Guo, Hua Wei. *Abg-SciQA: A Dataset for Understanding and Resolving Ambiguity in Scientific Questions*. ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models

[ITSC'24] **Tiejin Chen***, Prithvi Parag Shirke*, Bharatesh Chakravarthi, Arpitsinh Vaghela, Longchao Da, Duo Lu, Yezhou Yang, Hua Wei. *SynTraC: A Synthetic Dataset for Traffic Signal Control from Traffic Monitoring Cameras*. Proceedings of the 27th IEEE International Conference on Intelligent Transportation Systems

[Phys. Rev. A] Ningyi Xie, Jiahua Xu, **Tiejin Chen**, Yoshiyuki Saito, Nobuyoshi Asai, Dongsheng Cai. *Performance Upper Bound of the Grover-Mixer Quantum Alternating Operator Ansatz*. Physical Review A

[BlackHat'24] Zicheng Wang, **Tiejin Chen**, Qinrun Dai, Yueqi Chen, Hua Wei, Qingkai Zeng. *When eBPF Meets Machine Learning: On-the-fly OS Kernel Compartmentalization*. BlackHat USA 2024 Briefing

[IJMLC] Longchao Da, Kuan-Ru Liou, **Tiejin Chen**, Xuesong Zhou, Xiangyong Luo, Yezhou Yang, Hua Wei. *Open-TI: Open Traffic Intelligence with Augmented Language Model*. International Journal of Machine Learning and Cybernetics (IF=5.6)

[AAAI'24] Kai Ye, **Tiejin Chen**, Hua Wei, Liang Zhan. *Uncertainty Regularized Evidential Regression*. Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence

Preprints

[arXiv'25] Longchao Da, **Tiejin Chen**, Zhuoheng Li, Shreyas Bachiraju, Huaiyuan Yao, Xiyang Hu, Zhengzhong Tu, Yue Zhao, Dongjie Wang, Ram Pendyala, Benjamin Stabler, Yezhou Yang, Xuesong Zhou, Hua Wei. *Generative AI in Transportation Planning: A Survey*. arXiv:2503.07158

[arXiv'24] Longchao Da, **Tiejin Chen**, Lu Cheng, Hua Wei. *LLM Uncertainty Quantification through Directional Entailment Graph and Claim Level Response Augmentation*. arXiv:2407.00994

[arXiv'23] **Tiejin Chen***, Yuanpu Cao*, Yujia Wang*, Cho-Jui Hsieh, Jinghui Chen. *Federated Learning with Projected Trajectory Regularization*. arXiv:2312.14380

[arXiv'22] **Tiejin Chen***, Yicheng Tao*. *Learning Sparsity and Randomness for Data-driven Low Rank Approximation*. arXiv:2212.08186

Service

Teaching Assistant: CSE 355, Intro Theoretical Computer Sci, Spring 2025. CSE 572, Data Mining, Spring 2026.
Reviewer: PAKDD 2024, ICLR 2024 SeTLLM Workshop, WACV 2025, Neurips 2024 SafeGenAI Workshop, ICLR 2025, SDM 2025, CVPR2025, ICCV 2025, ACL 2025 Demo Track, EACL 2026, ICLR 2026, KDD 2026.

Award

Sichuan University Comprehensive Scholarship
ASU Experiential Learning Grant
ASU Graduate Student Government Travel Grant
NSF travel award
Ferdinand A. Stanchi Fellowship

Experience

Amazon 08/2025 – 11/2025
Applied Scientist Intern *Sunnyvale*

- Post-training on small LLM on various tasks to replace Claude's Model in the production pipeline;
- Reduce the overall cost to 1%-5% of the original cost with comparable performance.

Toyota Motor North America, Inc. (InfoTech Lab) 05/2025 – 08/2025
STEM Co-op *Mountain View*

- Build the multi-agent systems that generate the personalized trip plan based on the user's requirement;
- Research about the orchestration of multi-agent systems and write a patent about how to detect similar agents inside orchestration and remove the ambiguity provided by the similar agents;
- Involved in 4 patents about the hybrid autoregressive language models and diffusion language model, continue learning for LLMs, and increase the performance of orchestration of multi-agent systems.
- The submitted paper is accepted by IV 2026.

Points Technology 03/2021 – 08/2021
Algorithm Intern *Shanghai*

- Get to learn federated learning. Reproduce the vertical logistic regression in federated learning way by numpy. Learn some basic knowledge of homomorphic encryption and secret sharing;
- Research about the recommendation system. Reproduce the SVD,FM,FunkSVD,BiasSVD algorithm with numpy, reproduce AutoRec. Denoisy AutoRec,NFM,AFM,AFN,NFM,FiBiNet,DeepFm etc. deep learning recommendation algorithm by Pytorch;
- Design a vertical DeepFm algorithm. Work with team to realize the vertical DeepFm.

Research Projects

Reinforcement Learning From AI Feedback with Uncertainty 12/2024 – 08/2025

- This project is funded by the AWS Research Award;
- Read papers about Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF);
- Enhance the performance of pre-trained LLMs on various tasks including reasoning and summarization using RLAIF;
- Implement a novel algorithm that utilizes the uncertainties from conformal prediction to further enhance the performance of existing RLAIF technologies.
- The Paper is accepted to EACL 2026.

Jailbreak Black-box Multi-modal LLMs 05/2024 – 09/2024

- Submitted paper is under review;
- Propose a novel zeroth-order gradient-based method to jailbreak black-box MLLMs by generating malicious images;
- Propose patch coordinate descent in our method, significantly reducing memory consumption, enabling attacks on 13B models using a single GPU;
- Our method achieves high attack success rates, surpassing transfer-based methods, performing on par with white-box approaches and achieving 95% attack success rates on MiniGPT-4 ;
- Using a showcase to demonstrate that it is possible for our method to directly attack commercial MLLMs such as GPT-4o.

Privacy Benchmark of Multi-Modal LLMs 03/2024 – 06/2024

- Read papers about Multi-modal LLMs (MLLMs) and safety aspects of Multi-modal LLMs including jailbreak and defending methods of Multi-modal LLMs;
- Build a dataset that evaluates the potential privacy leakages of Multi-modal LLMs. The dataset contains different tables as images and evaluates whether MLLMs will output private information in the images or private information in their memory induced by the input images;
- Evaluate the influence of different tasks for potential privacy leakages under various MLLMs including GPT-4o, Idefics2, etc. The results show that different tasks have different risks of privacy leakage.
- Write the paper that is currently under review by ARR.

Ambiguity Related Tasks Evaluation of LLMs 02/2024 – 08/2024

- Get familiar with datasets on ambiguity and its corresponding tasks, including ambiguity detection and clarification question generation;
- Build a new dataset on ambiguity based on the articles from different areas of science, from different tests using LLMs. The dataset is challenging and contains a new task of ambiguity-type classification.
- Evaluate the ambiguity dataset with different closed-source LLMs such as Gemini, GPT and Claude by their APIs. The results show that even the most state-of-the-art LLMs fail in the ambiguity-related tasks;
- Evaluate the dataset with different open-source LLMs and fine-tune LLMs on the dataset to improve the ability of ambiguity-related tasks of LLMs.
- Write a paper that is currently under review by ARR.

Privacy Fine-tuning for LLMs 09/2023 – 02/2024

- Submitted paper is accepted by SDM 2025;
- Explore the weight loss landscape of differential private (DP) models and discover that the DP-trained model has a sharper weight loss landscape;
- Come up with three methods from the perspective of cross-layers, within-layers and cross-models to flatten the weight loss landscape during differential private training;
- Experimental results on both black-box and white-box settings show that the methods can bridge the performance gap between privacy LLMs and normally trained LLMs;

Dataset Condensation 04/2022 – 02/2023

- Supervised by Prof.Jinghui Chen at Pennsylvania State University
- Research about Dataset Condensation which aims to create a much smaller dataset than the original one, and a network trained on this new dataset can have similar performance with networks trained on the original dataset;
- Explore method which aims to have state-of-the-art performance; Try to combine Dataset Condensation with a continual learning method such as AGEM;
- Research about utilizing dataset condensation to extract global information under federated learning and using global information to reduce the influence of Non-I.I.D federated learning.

Algorithm Competition: Adversarial Robustness of Deep Learning Based on ImageNet 08/2022 – 11/2022

- Attended the algorithm competition, which aims to get a high average accuracy on ImageNet under different white box attacks such as AutoAttack with different radius of perturbation;
- Replaced ReLU in Wide-ResNet with a more smoothing activation function such as SiLU to make the loss landscape smoother, which is beneficial to the robustness of the deep learning model;
- Added Non-local means denoising filters to ResNet, which can reduce the effect of perturbation from white box attacks;
- Adversarially trained several ResNet and EfficientNet under AutoAttack with different radius on ImageNet, and trained an ensemble model with all models and a pertain Swin Transformer to get a final model;
- Ranked 5th among all participants and won a prize of about 6000 dollars.

Technical Skills

Technologies: Python, OpenRLHF, Amazon AWS, PyTorch, Transformers, Large Language Models, Reinforcement Learning, Large Language Model Fine-tuning, Multi-agent System (AutoGen).
Hobbies: Mystery Novels, Oscar Prediction
Extracurricular Activities: Deputy director of Reasoning Association for organizing mystery games and organizing Sichuan University to join in the national BBS mystery contest.