

EDUCATION	Department of Computer Science, UC Davis <i>BS. in Computer Science and Engineering</i> • Advisors: Prof. Muhao Chen, Prof. Zhe Zhao • Research area: Statistical ML, NLP, Trustworthy AI	Davis, CA 2022 - 2026 (<i>expected</i>)
AWARDS AND HONORS	• (International) Amazon Trustworthy AI Fellow, 250k • (National) 2025 CRA Outstanding Undergraduate Researcher Award Honorable Mention • (Regional) UC Davis Provost Undergraduate Fellowship, 1.5k • (Regional) UC Davis Hong Kong Undergraduate Scholarship, 4k • (Regional) UC Davis Vincent V.C. Woo Undergraduate Scholarship, 3k • (Regional) UC Davis James Leung Family Endowed UG Scholarship, 2k • (Regional) UC Davis Dean Witter / Student Development Fund, \$500 • (Regional) UC Davis 2024 Fall Travel Award, \$500 • Dean's Honor Roll, x3, • Chairman's Award for Excellence, 45/45 IB, KGV High School	2024.10 2024.12 2024.04 2023.09 2025.01 2025.01 2024.11 2024.11 2022.9-Present 2022.09
PUBLICATIONS	1. Terry Tong, Fei Wang Zhe Zhao, Muhao Chen, "BadJudge: Backdoor Vulnerabilities of LLM-as-a-Judge" <i>In Proceedings of the 13th International Conference on Learning Representations (ICLR)</i>, 2025 • Identified vulnerability in the NLP evaluation pipeline with automatic judges, single token on 1% data leads to 3x of scores. • Proposed a principled defense that integrates seamlessly into the LLM Judge training pipeline, reducing ASR to $\sim 0\%$, and achieving SOTA on evaluation abilities. • Came up with LLM Judge vulnerability idea, designed all experiments and wrote all the code, led paper writing of all sections with comments from co-authors and PI. 2. Terry Tong, Jiashu Xu, Qin Liu, Muhao Chen, "Securing Multi-turn Conversational Language Models from Distributed Backdoor Attacks", <i>In the 40th Conference on Empirical Methods in Natural Language Processing (EMNLP-Findings)</i>, 2024 https://arxiv.org/abs/2407.04151 • Exposed a new threat to multi-turn chatbots, attack success rate (ASR) of 97%+ on all experimented triggers and hard to defend as output space is larger. • Proposed a novel linear-time defense that drops ASR to as low as 0.36%. • Came up with idea of attack and defense, designed all experiments and wrote all the code, led paper writing of all sections with comments from co-authors and PI. 3. Qin Liu, Wenjie Mo, Terry Tong, Jiashu Xu, Fei Wang, Chaowei Xiao, Muhao Chen, "Mitigating Backdoor Threats to Large Language Models: Advancement and Challenges", <i>In Proceedings of the 60th Annual Allerton Conference on Control, Communication, and Computing Systems (Allerton)</i>, 2024. https://arxiv.org/abs/2409.19993 • Led the backdoor attack section, identifying threats at training-time and test-time, spanning instance-level attacks and other stealthier variants.	

ONGOING
PROJECTS

1. **Terry Tong**, Fei wang, Anshuman Chaabra, Zhe Zhao, Muhao Chen, “ISVP: Influence Selection via Proxy for Efficient Preference Adaptation”, **To be submitted to NeurIPS 2025**
 - Contributed a novel derivation for preference-data influence functions.
 - Designed the methodology of approximating the test-time loss with unobserved preference data.
 - Leading the project, came up with the idea and proposed solution, currently designing experiments.
2. Hadi Askari, Shivanshu Gupta, **Terry Tong**, Fei Wang, Anshuman Chaabra , Muhao Chen, “Influence Functions for Indirect and Noisy In-Context Learning”, Under Review at **ACL**, 2025.
 - Led the literature review, led the EKFac influence function section, encompassing both experiments and code.

ACADEMIC
SERVICES

Reviewers for: *EMNLP-2024 (5 papers)*, ICLR-2025, NAACL-2025, ACL-2025
Supported Grants / Proposals :
Amazon Trustworthy AI Challenge Fellowship (250k)

SKILLS

Languages: English, Chinese
Programming: Python (Pytorch, Transformers, TRL, Accelerate/PEFT/Quantization/Deepspeed, Hydra, Optuna, Wandb, Pandas, Numpy, Matplotlib/Seaborn, StatsModels/Scikit-Learn/Scipy), C++ / C, Java.