Tiankai Yang

raymondyangtk@gmail.com
raymondy.github.io
github.com/RaymondY
linkedin.com/in/tiankai-yang
Google Scholar

(213) 275-6503
CS Department, GCS Hall
Los Angeles, CA
University of Southern California
Department of Computer Science

Research Interests

1. Post-training Alignment for Trustworthy LLMs
Safety Alignment, Preference Learning

2. Robust and Reliable LLM Inference
Hallucination Mitigation, Jailbreak Detection, OOD Detection, Multimodal Robustness, Model Selection & Routing

3. Trustworthy LLM Agents and Agentic Systems
Agent Safety, Runtime Reliability, Multi-Agent Orchestration

Education

University of Southern California, Los Angeles, CA Aug 2024 – Present
Ph.D. in Computer Science
· Affiliation: FORTIS Lab; Advisor: Prof. Yue Zhao

University of Southern California, Los Angeles, CA Jan 2022 – Dec 2023
M.S. in Machine Learning and Data Science

Nankai University, Tianjin, China Sep 2017 – Jul 2021
B.E. in Software Engineering

Selected Preprints & Under Submission

* *denotes joint first authors. See Google Scholar for the full list.*

· **Cat-DPO: Category-Adaptive Safety Alignment**
Tiankai Yang*, Yi Nian*, Xinyuan Li, Ruiyao Xu, Henry Peng Zou, Kaize Ding, Xiyang Hu, Yan Liu, Yue Zhao.
Under submission. arXiv preprint arXiv:2604.17299

· **No Attacker Needed: Unintentional Cross-User Contamination in Shared-State LLM Agents**
Tiankai Yang, Jiate Li, Yi Nian, Shen Dong, Ruiyao Xu, Ryan A. Rossi, Kaize Ding, Yue Zhao.
Under submission. arXiv preprint arXiv:2604.01350

· **Learning to Route LLMs from Bandit Feedback: One Policy, Many Trade-offs**
Wang Wei, Tiankai Yang, Hongjie Chen, Yue Zhao, Franck Dernoncourt, Ryan A. Rossi, Hoda Eldardiry.
Under submission. arXiv preprint arXiv:2510.07429

· **StealthRank: LLM Ranking Manipulation via Stealthy Prompt Optimization**
Yiming Tang, Yi Fan, Chenxiao Yu, Tiankai Yang, Yue Zhao, Xiyang Hu.
Under submission. arXiv preprint arXiv:2504.05804

Selected Publications

* *denotes joint first authors. See Google Scholar for the full publication list.*

- **CoAct: Co-Active LLM Preference Learning with Human-AI Synergy**
  Ruiyao Xu, Mihir Parmar, Tiankai Yang, Zhengyu Hu, Yue Zhao, Kaize Ding.
  ACL, Oral, 2026

- **A Personalized Conversational Benchmark: Towards Simulating Personalized Conversations**
  Li Li, Peilin Cai, Ryan A. Rossi, Franck Dernoncourt, Branislav Kveton, Junda Wu, Tong Yu, Linxin Song, Tiankai Yang, Yuehan Qin, Nesreen K. Ahmed, Samyadeep Basu, Subhojyoti Mukherjee, Ruiyi Zhang, Zhengmian Hu, Bo Ni, Yuxiao Zhou, Zichao Wang, Yue Huang, Yu Wang, Xiangliang Zhang, Philip S. Yu, Xiyang Hu, Yue Zhao.
  NeurIPS MTI-LLM Workshop, Spotlight, 2025

- **AD-AGENT: A Multi-agent Framework for End-to-end Anomaly Detection**
  Tiankai Yang*, Junjun Liu*, Michael Siu*, Jiahang Wang, Zhuangzhuang Qian, Chanjuan Song, Cheng Cheng, Xiyang Hu, Yue Zhao.
  IJCNLP-AACL Findings, 2025

- **Treble Counterfactual VLMs: A Causal Approach to Hallucination**
  Shawn Li, Jiashu Qu, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, Yue Zhao.
  EMNLP Findings, 2025

- **AD-LLM: Benchmarking Large Language Models for Anomaly Detection**
  Tiankai Yang*, Yi Nian*, Shawn Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan A. Rossi, Kaize Ding, Xia Hu, Yue Zhao.
  ACL Findings, 2025

- **DPU: Dynamic Prototype Updating for Multimodal Out-of-Distribution Detection**
  Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, Yue Zhao.
  CVPR, Highlight, 2025

- **PyOD 2: A Python Library for Outlier Detection with LLM-powered Model Selection**
  Sihan Chen*, Zhuangzhuang Qian*, Michael Siu*, Tiankai Yang, Xingcan Hu, Jiaqi Li, Shawn Li, Yuehan Qin, Zhuo Xiao, Wanghao Ye, Yichi Zhang, Yushun Dong, Yue Zhao.
  The Web Conference (Demo Track), 2025

Work Experience

LinkedIn — May 2026 – Aug 2026 (Incoming)
AI/ML Engineer Intern, Generative AI — Sunnyvale, CA

USC Media Communications Lab — May 2023 – Aug 2023
Research Assistant for Prof. C.-C. Jay Kuo — Los Angeles, CA

Services

Program Committee for Conferences

- ICLR 2026; ICML 2025, 2026; ACL ARR 2025 (*Outstanding Reviewer*), 2026; NeurIPS 2026; AAAI 2026; COLM 2026.

Journal Reviewer

- IEEE Transactions on Computational Social Systems