Yiwei Wang
Assistant Professor, Computer Science Department, University of California, Merced
https://wangywust.github.io `yiweiwang2@ucmerced.edu`

Research Interests

Trustworthy and robust large language models, safety alignment, adversarial robustness, vision-language models, and interpretability of foundation models.

Education

Ph.D., Computer Science — National University of Singapore
M.Phil., Electronic and Computer Engineering — HKUST
B.Sc., Information Engineering — Southeast University

Academic Appointments

Assistant Professor — University of California, Merced
Postdoctoral Researcher — UCLA NLP Group

Selected Publications

- Yiwei Wang, Muhao Chen, Nanyun Peng, Kai-Wei Chang. *Vulnerability of Large Language Models to Output Prefix Jailbreaks: Impact of Positions on Safety.* NAACL 2025.
- Cheng Wang, Yiwei Wang, Bryan Hooi, Yujun Cai, Nanyun Peng, Kai-Wei Chang. *Con-ReCall: Detecting Pre-training Data in LLMs via Contrastive Decoding.* COLING 2025.
- Bingxuan Li, Yiwei Wang, Tao Meng, Nanyun Peng, Kai-Wei Chang. *Control Large Language Models via Divide-and-Conquer.* EMNLP 2024.
- Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Xueqi Cheng. *SLANG: New Concept Comprehension of Large Language Models.* EMNLP 2024.
- Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Xueqi Cheng. *Adaptive Token Biaser: Knowledge Editing via Biasing Key Entities.* EMNLP 2024.
- Zhicheng Yang, Yiwei Wang, Yinya Huang, Zhijiang Guo, Wei Shi, Liang Feng, Jing Tang. *OptiBench meets ReSocratic: Measure and Improve LLMs for Optimization Modeling.* ICLR 2025.
- Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Xueqi Cheng. *Is Factuality Enhancement a Free Lunch For LLMs? Better Factuality Can Lead to Worse Context-Faithfulness.* ICLR 2025.
- Shaochen Zhong, et al. *MQuAKE-Remastered: Multi-Hop Knowledge Editing Can Only Be Advanced With Reliable Evaluations.* ICLR 2025.
- Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, Bryan Hooi. *Primacy Effect of ChatGPT.* EMNLP 2023.

Academic Services

- **Area Chair**: NeurIPS 2025, ICLR 2025, ICML 2025, ACL Rolling Review (ARR) 2024–2025, COLM 2025, IJCAI 2025, ICME 2025.
- **Reviewer (Conferences)**: NeurIPS, ICLR, ICML, EMNLP, ACL, NAACL, CVPR, ICCV, AAAI, KDD, WWW, AISTATS, ECML-PKDD, IJCNN.
- **Reviewer (Journals)**: TMLR, TPAMI, TKDE, Neurocomputing, Pattern Recognition, IEEE Transactions on Big Data, IEEE Transactions on Systems, Man and Cybernetics: Systems, Computer Science Review, Cybernetics and Systems, IET Image Processing.

Teaching

- **Instructor:** Large Language Models, University of California at Merced, 2025
- **Teaching Assistant (TA):** Big Data Systems for Data Science (NUS, 2021), Knowledge Discovery and Data Mining (NUS, 2020–2021), Programming Methodology (NUS, 2021), Parallel Computing (NUS, 2020), Signal Processing and Communications (HKUST, 2018)

Selected Honors

- Dean's Graduate Research Excellence Award, National University of Singapore, 2022
- SDSC Dissertation Research Fellowship (Top 10 across Singapore), 2021