Kaijie Zhu

☎ +1-820-587-3174 ✉ kaijiezhu11@gmail.com 🎓 Google Scholar 🌐 Website

Education

University of California, Santa Barbara — **Sep. 2024 – June 2029 (expected)**
Ph.D, Computer Science, GPA: 4.00/4.00 — *California, US*

Institute of Automation, Chinese Academy of Sciences — **Sep. 2021 – June 2024**
Master, Computer Science, GPA: 3.86/4.00 — *Beijing, China*

Huazhong University of Science and Technology — **Sep. 2017 – June 2021**
Bachelor, ACM Class in Computer Science, GPA: 3.95/4.00 — *Wuhan, Hubei, China*

Research Interests

- **Trustworthy Machine Learning:** Adversarial robustness, indirect prompt injection
- **Large Language Models:** Dynamic evaluation, Agents

Publications

- **Kaijie Zhu**, Xianjun Yang, Jindong Wang, Wenbo Guo, William Wang. COCO: Defending Indirect Prompt Injection via Contrasive Output Comparison [Submitted to ICML 2025]

- **Kaijie Zhu**, Xixu Hu, Jindong Wang, Xing Xie, Ge Yang. *Improving Generalization of Adversarial Training via Robust Critical Fine-Tuning.* [ICCV 2023]

- **Kaijie Zhu**, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, Xing Xie. *PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts.* [CCS 2024 LAMPS] 🎓 **200+**

- **Kaijie Zhu***, Jiaao Chen*, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, Xing Xie. *DyVal: Graph-informed Dynamic Evaluation of Large Language Models.* [ICLR 2024 (Spotlight)]

- **Kaijie Zhu**, Jindong Wang, Qinlin Zhao, Ruochen Xu, Xing Xie. *Dynamic Evaluation of Large Language Models by Meta Probing Agents* [ICML 2024]

- **Kaijie Zhu**, Qinlin Zhao, Hao Chen, Jindong Wang, Xing Xie. *PromptBench: A Unified Library for Evaluation of Large Language Models* [JMLR MLOSS] ★ **2.5k**

- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, **Kaijie Zhu**, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S Yu, Qiang Yang, Xing Xie. *A survey on evaluation of large language models.* [ACM TIST] 🎓 **1.9k**

- Cheng Li, Jindong Wang, **Kaijie Zhu**, Yixuan Zhang, Wenxin Hou, Jianxun Lian, Xing Xie. *Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus.* [ICML 2024]

- Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, **Kaijie Zhu**, Hao Chen, Xing Xie. *CompeteAI: Understanding the Competition Behaviors in Large Language Model-based Agents.* [ICML 2024 (Oral)]

- Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, **Kaijie Zhu**, Yijia Xiao, Jindong Wang. *AgentReview: Exploring Peer Review Dynamics with LLM Agents.* [EMNLP 2024 (Oral)]

Experience

Microsoft Research Asia — **Mar. 2023 – Apr. 2024**
Research Intern Advisors: Jindong Wang, Xing Xie — *Beijing, China*

- Developed a robust fine-tuning strategy to enhance the generalization ability of adversarially trained models.
- Introduced PromptBench: a benchmark to evaluate the robustness of LLMs on adversarial prompts.
- Proposed a graph-informed dynamic evaluation for LLMs in reasoning tasks to mitigate test data contamination.

Projects

promptbench | ★ **2.5k** — **Mar. 2023 – Current**

- Developed a flexible evaluation pipeline for large language models as the main contributor.
- Incorporated prompt engineering, dynamic evaluation for accelerating research in LLMs.

Awards

- **Excellent Graduate Student (Top 5%)**, Huazhong University of Science and Technology, 2021
- **Outstanding Student (Top 5%)**, Huazhong University of Science and Technology, 2019
- **Certified Software Professional Test (Top 1%)**, China Computer Federation (CCF), 2019