Kaijie Zhu

+1-820-587-3174 kaijiezhu@ucsb.edu Google Scholar Website

Education

University of California, Santa Barbara **Sep. 2024 – June 2028 (expected)**
Ph.D, Computer Science, GPA: 4.00/4.00, Advisors: William Wang, Wenbo Guo *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

- **Agent:** Evaluating (propose new benchmarks), improving (via SFT and RL), and securing (preventing indirect prompt injection) agent for tool use.

Publications

Reinforcement Learning & Agent for Tool use

- **Kaijie Zhu**, Yuzhou Nie, Yijiang Li, Yiming Huang, Jialian Wu, Jiang Liu, Ximeng Sun, Zhenfei Yin, Lun Wang, Zicheng Liu, Emad Barsoum, William Yang Wang, Wenbo Guo. *TermiGen: High-Fidelity Environment and Robust Trajectory Synthesis for Terminal Agents.* [Submitted to ICML 2026, HuggingFace Daily Paper #2]
- Xian Wu*, **Kaijie Zhu***, Wenbo Guo, William Wang. *rePIRL: Learn PRM with Inverse RL for LLM Reasoning.* [Submitted to ICML 2026]
- Yuheng Tang*, **Kaijie Zhu***, et al. *DevOps-Gym: Benchmarking AI Agents in Software DevOps Cycle.* [ICLR 2026]
- **Kaijie Zhu**, Xianjun Yang, Jindong Wang, Wenbo Guo, William Wang. *MELON: Indirect Prompt Injection Defense via Masked Re-execution and Tool Comparison.* [ICML 2025]
- Zekun Li, Shinda Huang, Jiangtian Wang, Nathan Zhang, Antonis Antoniades, Wenyue Hua, **Kaijie Zhu**, Sirui Zeng, William Yang Wang, Xifeng Yan. *AgentOrca: A Dual-System Framework to Evaluate Language Agents on Operational Routine and Constraint Adherence.* [ACL 2025]
- 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)]

Dynamic Evaluation of LLMs

- **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] **2.6k**
- 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] **2.4k**

Adversarial Robustness

- **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] **300+**
- **Kaijie Zhu**, Xixu Hu, Jindong Wang, Xing Xie, Ge Yang. *Improving Generalization of Adversarial Training via Robust Critical Fine-Tuning.* [ICCV 2023]

Experience

AMD **Oct. 2025 – March 2026**

Part-time Research Intern *Advisors: Jialian Wu* *Santa Clara, CA*

- Agent training for bash tool use: TermiGen.

Microsoft Research **June 2025 – Sep. 2025**

Research Intern *Advisors: Sheng Zhang, Hoifung Poon* *Redmond, WA*

- Proposed a new data synthetic method and ranking loss to mitigate the label noise for process reward model training.

Microsoft Research Asia **March 2023 – April 2024**

Research Intern *Advisors: Jindong Wang, Xing Xie* *Beijing, China*

- [ICCV'23] Developed a robust fine-tuning strategy to enhance the generalization ability of adversarially trained models.
- [JMLR'24] Introduced PromptBench: a benchmark to evaluate the robustness of LLMs on adversarial prompts.
- [ICLR'24 Spotlight, ICML'24] Proposed a graph-informed dynamic evaluation for LLMs in reasoning tasks to mitigate test data contamination.

Awards

- **KAUST Rising Stars in AI Symposium 2025**
- **Microsoft Research Star of Tomorrow**, Microsoft, 2024
- **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

Projects

promptbench | **2.7k** **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.

SearchAnything | **300+** **June 2023**

- Created a semantic local search tool for retrieving texts and images, powered by state-of-the-art AI models.

Tutorials & Invited Talks

- **CVPR 2025 Tutorial:** Multi-modal Models Evaluation: Methods and Challenges
- **AAAI 2025 Tutorial:** Evaluation of LLMs: Methods and Challenges
- **Microsoft Research Asia 2024 LLM Evaluation Symposium**