Kaijie Zhu

+86-152-7182-9115 kaijiezhu11@gmail.com Google Scholar

Education

Institute of Automation, Chinese Academy of Sciences **Sep. 2021 – June 2024 (expected)**
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, Detecting AIGC
- **Large Language Models:** Evaluation

Publications

- *Improving Generalization of Adversarial Training via Robust Critical Fine-Tuning.* **Kaijie Zhu**, Jindong Wang, Xixu Hu, Xing Xie, Ge Yang [ICCV 2023]
- *DyVal: Graph-informed Dynamic Evaluation of Large Language Models.* **Kaijie Zhu***, Jiaao Chen*, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, Xing Xie [Submitted to ICLR 2024]
- *PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts.* **Kaijie Zhu**, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, Xing Xie [Submitted to USENIX Security 2024] **36** **315**
- *A survey on evaluation of large language models.* 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 [Submitted to TIST] **98**
- *CompeteAI: Understanding the Competition Behaviors in Large Language Model-based Agents.* Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, **Kaijie Zhu**, Hao Chen, Xing Xie
- *Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus.* Cheng Li, Jindong Wang, **Kaijie Zhu**, Yixuan Zhang, Wenxin Hou, Jianxun Lian, Xing Xie

Experience

Microsoft Research Asia **Oct. 2022 – Current**
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 | **315** **Mar. 2023 – Current**

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

robustlearn | **357** **Oct. 2022 – Current**

- Collected latest research in robust machine learning, including adversarial/backdoor attack and defense, out-of-distribution generalization, and safe transfer learning.

SearchAnything | **175** **June 2023**

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

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

Collaborator Professors

- **Jindong Wang** – Microsoft Research
- **Xing Xie** – Mircosoft Research
- **Janice Yixuan Zhang** – William & Mary College
- **Neil Zhenqiang Gong** – Duke University
- **Diyi Yang** – Stanford University