Ruiyang Sun /rueɪ jaŋ swən/

Undergraduate Student; *AI Researcher*

ruiyangsun02@gmail.com ☎ +86 18609715159 rockmagma02
sun_ruiyang@stu.pku.edu.cn `https://www.ruiyangsun.com`
@RuiyangSun02 Google Scholar LinkedIn

Education

Sep 2021 — Present	**B.Sc. in Artificial Intelligence** School of Intelligence Science and Technology, ***Peking University*** Double Degree Program
Sep 2020 — Present	**B.Sc. in Geophysics** School of Earth and Space Sciences, ***Peking University*** Major in Geophysics

Research Experience

Sep 2021 — Sep 2022	**Research Intern** – Institute of Theoretical and Applied Geophysics, ***Peking University*** – Focus on computational **Seismic Waveform Inversion**
Sep 2022 — Dec 2023	**Research Intern** – PKU Alignment and Interaction Research Lab – Advised by Prof. **Yaodong Yang** – Focus on **Safe-RL** and **AI alignment** – Contributed to 4 papers, including one **co-first author** paper in **ICLR 2024 Spotlight** – Contributed to 2 open-source projects, gained **2000+ stars**

Working Experience

Jul 2023 — Sep 2023	**LLM Alignment Engineer Intern** – Beijing Baichuan Intelligent Technology Co., Ltd. – Provided **RLHF** technological support for **Baichuan-2** – Focus on scale up **RL Training for LLM** – Simultaneously operated exceed **500 A100 GPUs**

Research Publications

1. Josef Dai*, Xuehai Pan*, **Ruiyang Sun***, Jiaming Ji*, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang†, ***Safe RLHF: Safe Reinforcement Learning from Human Feedback***, in *The Twelfth International Conference on Learning Representations*, ser. ICLR 2024 Spotlight, 2023. DOI: `10.48550/arxiv.2310.12773`.

2. Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, **Ruiyang Sun**, Yizhou Wang, and Yaodong Yang, ***BEAVERTAILS: towards improved safety alignment of llm via a human-preference dataset***, in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS 2023, Red Hook, NY, USA: Curran Associates Inc., 2023. 🔗DOI: `10.48550/arxiv.2307.04657`.
3. Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, **Ruiyang Sun**, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu, ***Baichuan 2: Open Large-scale Language Models***, *arXiv*, 2023. 🔗DOI: `10.48550/arxiv.2309.10305`. eprint: `2309.10305`.
4. Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, **Ruiyang Sun**, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang, ***Safety Gymnasium: A Unified Safe Reinforcement Learning Benchmark***, in *Advances in Neural Information Processing Systems*, ser. NIPS 2023, vol. 36, Curran Associates, Inc., 2023, pp. 18 964–18 993. 🔗[Online]. Available: `https://arxiv.org/abs/2310.12567`.
5. Jiaming Ji, Jiayi Zhou, Borong Zhang, Juntao Dai, Xuehai Pan, **Ruiyang Sun**, Weidong Huang, Yiran Geng, Mickel Liu, and Yaodong Yang, ***OmniSafe: An Infrastructure for Accelerating Safe Reinforcement Learning Research***, *Journal of Machine Learning Research*, vol. 25, no. 285, pp. 1–6, 2024. 🔗[Online]. Available: `http://jmlr.org/papers/v25/23-0681.html`.

Awards

Sep 2021 — Merit Student; PKU Scholarship From **Peking University**

Skill

Coding	Adept in **Python**, **JavaScript/TypeScript** and **Swift** Familiar with **C**, **C++**, **Shell**, **Matlab**, **React** Basic knowledge of **Java** Loving **Open Source**, contributor of **Hugging Face Transformers**
Language	**Mandarin Chinese** (Native), **English** (Fluent)
AI Framework	Pytorch, Tensorflow, Hugging Face, numpy, pandas, matplotlib, seaborn
Tools	Git, Docker, Linux, LaTeX, SSH, Tmux, Vim, ...