Changye Li

Peking, China antoine031106@gmail.com +86 13860472996 antoinegg1.github.io
antoinegg1

Summary

I am a second-year undergraduate at Peking University (Class of '26), currently focused on AI alignment, with a particular interest in alignment algorithms and mechanistic interpretability. I also have prior experience in Reinforcement Learning and AI for Medicine. My research is driven by the following questions:

- How can the gap between artificial intelligence and human-level intelligence be quantified through mechanistic interpretability methods, and how can it be bridged using learning algorithms?
- How can the fundamental nature of intelligent behavior be uncovered through the interpretation of different models?

Education

Peking University *Sept 2022 – May 2026*
B.S. Student in Artificial Intelligence

Fellowships & Awards

Peking University Freshman Scholarship (¥25000 RMB)(2022)

Research Experience

Visiting Student Researcher at PAIR Lab: PKU Alignment and Interaction Research Lab *Peking, China 2023 –*

Currently working on Alignment and Interpretability of Language Models under the guidance from Dr. Yaodong Yang.

Publications

Language Models Resist Alignment,*Accepted at Neurips 2024 SoLar Program* Oct 2024
Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, ***Changye Li***, Hantao Lou, Jiayi Zhou, Josef Dai, Yaodong Yang,

Towards efficient collaboration via graph modeling in reinforcement learning, *In Submission* August 2024
Wenzhe Fa, Zishun Yu, Chengdong Ma, ***Changye Li***, Yaodong Yang, Xinhua Zhang,