Personal Statement

Changye Li
Yuanpei College, Peking University
2200017853@stu.pku.edu.cn

Introduction

I am a junior undergraduate at Peking University Yuanpei College (Class of '26), majoring in artificial intelligence (General Arithmetic Intelligence Honors Class), currently focused on Reinforcement Learning (RL). My research interests also cover AI alignment, Interpretability and Autonomous Driving..

Research Interests

My research is driven by the following questions:

How to close the gap between artificial intelligence and human-level intelligence by designing reliable AI systems and efficient learning algorithms?

Concretely, while current large language models (LLMs) and multimodal models have achieved certain successes, particularly with LLMs almost reaching human-level performance in the nature language process (NLP) domain, the gap between artificial intelligence and human-level intelligence remains substantial. This gap manifests itself in two main aspects: 1. The disparity between AI and humans in simulating human activities, such as complex reasoning, image generation and autonomous driving. [1, 2, 3]. 2. The gap between AI and humans in receiving human instructions and human-AI collaboration, evident in the widespread instruction following and human preference learning challenges across various generative tasks [4, 5].

To address the first challenge, we need an AI system with efficient learning capacity. Although various subtasks in NLP and computer vision (CV) can be completed independently by specialized models, we inevitably need to build an AI system capable of efficient learning. Here, "system" implies a potential combination of multiple models, both in encoding and decoding aspects, corresponding to humans' ability to acquire information from multiple channels in their environment.

Regarding the second problem, two key points are making AI reliable from a human perspective and enabling AI to understand human instructions and preferences. This essentially covers both interpretability and alignment work. These two aspects are two sides of the same coin, because without understanding how AI works (whether through mechanical interpretability or more general interpretability), we can hardly design algorithms for truly efficient alignment. Current alignment methods, such as reinforcement learning from human feedback (RLHF) [6] and direct preference optimazation (DPO) [7], completely separate training and inference, and the post-training process requires substantial resources. This does not align with our true expectations for alignment. Taking assisted autonomous driving as an example, I believe that ideal alignment would involve initially obtaining a self-driving car with ordinary training, but after dozens of hours of driving, the AI system should understand human preferences and actively adapt to the driver's behavior. This kind of minimal, efficient, and generalizable training approach is far beyond what current learning algorithms can achieve.

In summary, considering data, models, and algorithms, we need more realistic and abundant data, more versatile and interpretable foundation models, and more efficient learning algorithms to close the gap between artificial intelligence and human-level intelligence within the next decade.

Research experience

Since the winter of 2023, I have been an intern in PKU Alignment and Interaction Research Lab (PAIR Lab) supervised by Professor Yang Yaodong. My previous research has focused on model alignment and interpretability, dedicated to aligning models with human preference and improving model performance.

In the spring of 2024, I participated in the paper, *Language Models Resist Alignment: Evidence From Data Compression* [8], which aimed to reveal the internal resistance mechanisms of language model alignment. My primary contributions included experiment design and model training. Through this work, I mastered the post-training pipeline for LLMs and gained practical experience with LLM finetuning [9], RLHF and other methods, developing a comprehensive understanding of the alignment field. This work is currently under submission to ACL 2025.

From March to August 2024, I contributed to the research, *Towards Efficient Collaboration via Graph Modeling in Reinforcement Learning* [10], This work implemented f-MAT, an efficient collaborative architecture for multi-agent RL. I independently established the environmental framework for RL training algorithms and applied this architecture to traffic light control scenarios for experiment, gaining proficiency in RL methods and practical implementation. This work has been accepted as a poster presentation at AAAI 2025.

Between August 2024 and January 2025, my research shifted toward multimodal models. Driven by curiosity about the internal mechanisms behind the failure cases of current multimodal models [11], I focused on acquiring higher-quality data for multimodal alignment. I led the paper, *SAE-V: Interpreting Multimodal Models for Enhanced Alignment* [12], which aimed to reveal cross-modal mechanisms from an interpretability perspective and improve multimodal model performance using alignment algorithms. This work represented my initial attempt to combine interpretability with multimodal alignment. I developed a Sparse autoencoder for vision (SAE-V) based on a Sparse autoencoder (SAE) [13], allowing to conduct sparse encoding of hidden states in both text and visual modalities. Using this tool, I designed a Cosine Similarity Filtering algorithm that utilizes cross-modal information extracted by SAE-V for data filtering, resulting in high-quality training data. We conducted experiments using both finetuning and DPO alignment methods. Ultimately, SAE-V-based data filtering methods achieved more than 110% performance with less than 50% data. Our results highlight SAE-V's ability to enhance interpretability and alignment in mulitmodal large langauge models, providing insights into their internal mechanisms. This work has been accepted as a poster presentation at ICML 2025.

Moveover, from March to May 2025, I interned in the Tsinghua MARS Lab supervised by Professor Zhao Hang, participating in the work, *Finetuning Generative Trajectory Model with Reinforcement Learning from Human Feedback* [14], My primary contributions involved reinforcement learning and diffusion model post-training. In this work, I conducted a deeper exploration of the multimodal domain and attempted to optimize models using RL methods. I designed a reward model for trajectory generation, based on which I developed the corresponding RL algorithms, achieving preference alignment trajectory generation with aggressive and conservative two styles. This work is scheduled to be submitted to NeurIPS 2025.

Finally, from June 2025, I joined the Ant Technology Research Institute and, under the guidance of Professor Yi Wu, took part in work related to the AReaL project [15].

Research Plan

Based on my previous research experience and current research interests, my future scientific research direction will focus on further in-depth research in the reinforcement learning and multimodal domain, as well as designing RL algorithm to improve training and data utilizing efficiency. Based on the current AReaL training and reasoning architecture, I plan to refine and extend the AReaL framework over the next six months so that it can be efficiently applied to multi-modal, multi-architecture models. Afterwards, leveraging AReaL's high efficiency training and inference capabilities, I will continue to develop RL algorithms - through algorithmic innovation and full data utilization - to achieve a more advanced Large Reasoning Model.

Concluding Remark

As I prepare for the next stage of my academic journey, I am deeply committed to advancing the fields of multimodal learning and embodied intelligence. My research experiences across AI alignment, reinforcement learning, and multimodal interpretability have equipped me with the technical foundation and perspective needed to address fundamental AI challenges. I am particularly excited about the opportunity to join the Institute for Interdisciplinary Information Sciences (IIIS) at Tsinghua University , where the exceptional faculty, abundant academic resources, outstanding collaborators, and productive research atmosphere provide an ideal environment for cutting-edge research. At IIIS, I can contribute to pioneering work while pursuing my goal of closing the gap between artificial and human-level intelligence through more efficient learning algorithms and interpretable AI systems. I believe that my technical skills, research vision, and passion for solving complex problems make me well-suited for your graduate program, and I look forward to the possibility of contributing to your vibrant research community.

References

[1] Sumukh K Aithal, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation, 2024. URL https://arxiv.org/abs/2406.09358.

[2] Mohamed Abdel-Aty and Shengxuan Ding. A matched case-control analysis of autonomous vs human-driven vehicle accidents. *Nature Communications*, 15, 06 2024. doi: 10.1038/s41467-024-48526-4.

[3] Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, Ranjay Krishna, Dustin Schwenk, Eli VanderBilt, and Aniruddha Kembhavi. Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world, 2024. URL https://arxiv.org/abs/2312.02976.

[4] Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley Ren, Udhay Nallasamy, Andy Miller, and Jaya Narain. Do llms "know" internally when they follow instructions?, 2025. URL https://arxiv.org/abs/2410.14516.

[5] Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity, 2024. URL https://arxiv.org/abs/2401.01967.

[6] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

[7] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

[8] Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, and Yaodong Yang. Language models resist alignment: Evidence from data compression, 2024. URL https://arxiv.org/abs/2406.06144.

[9] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

[10] Wenzhe Fan, Zishun Yu, Chengdong Ma, Changye Li, Yaodong Yang, and Xinhua Zhang. Towards efficient collaboration via graph modeling in reinforcement learning, 2024. URL https://arxiv.org/abs/2410.15841.

[11] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

[12] Hantao Lou, Changye Li, Jiaming Ji, and Yaodong Yang. Sae-v: Interpreting multimodal models for enhanced alignment, 2025. URL https://arxiv.org/abs/2502.17514.

[13] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

[14] Derun Li, Jianwei Ren, Yue Wang, Xin Wen, Pengxiang Li, Leimeng Xu, Kun Zhan, Zhongpu Xia, Peng Jia, Xianpeng Lang, Ningyi Xu, and Hang Zhao. Finetuning generative trajectory model with reinforcement learning from human feedback, 2025. URL https://arxiv.org/abs/2503.10434.

[15] Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. Areal: A large-scale asynchronous reinforcement learning system for language reasoning, 2025. URL https://arxiv.org/abs/2505.24298.

[16] Peide Huang, Xilun Zhang, Ziang Cao, Shiqi Liu, Mengdi Xu, Wenhao Ding, Jonathan Francis, Bingqing Chen, and Ding Zhao. What went wrong? closing the sim-to-real gap via differentiable causal discovery, 2023. URL https://arxiv.org/abs/2306.15864.

[17] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11441–11450, 2020. doi: 10.1109/CVPR42600.2020.01146.

[18] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation, 2023. URL https://arxiv.org/abs/2210.02697.