Boyuan Chen

💌 boyuan.chen.byc@gmail.com 🎓 Google Scholar (1300+ citations) 🗥 https://cby-pku.github.io/ 🞧 cby-pku

About Me

Boyuan Chen is a forth-year undergraduate student at Yuanpei College, Peking University, majoring in Artificial Intelligence, advised by Prof. Yaodong Yang. His research primarily focuses on reinforcement learning and the safety and value alignment of large models, encompassing scalable oversight, deception alignment, and super alignment.

Boyuan has published a total of 9 papers, 4 of which are as first author or co-first author, in leading AI and ML conferences such as ACL and NeurIPS. Notably, his first-author or co-first-author papers have been accepted at prestigious venues, including NeurIPS 2024 Oral (top 0.45%), ACL 2025 Best Paper, NeurIPS 2025 Spotlight (top 2.6%), and ACM Computing Surveys (impact factor 33.2). His collaborative work was accepted in The Innovation (impact factor 33.2). To date, his research has garnered over 1,300 citations on Google Scholar.

As an active open-source contributor, Boyuan is the core developer of multiple projects with over 4,000 stars, with the open-source dataset being downloaded over 150,000 times. He also played a key role in the alignment, deployment, and open-sourcing of Hong Kong's first large language model, HKGAI-v1. His research on AI safety has garnered citations from leading institutions like Meta AI, OpenAI, and Yoshua Bengio, and has been featured in MIT Tech Review. Boyuan is the core author of the groundbreaking "AI Alignment: A Comprehensive Survey", which has become a pivotal resource for global AI safety initiatives, including the Bletchley Summit, the Beijing International AI Safety Consensus, and the Venice AI Safety Consensus Conference.

Boyuan has also been recognized with several honors, including the SenseTime Scholarship (awarded to only 25 individuals in China), the Beijing Natural Science Foundation Youth Scientist Grant (the sole recipient from Peking University's AI undergraduates in 2023) and so on.

Grants Received

PI, Beijing Natural Science Foundation Youth Scientist Grant

2023 - 2024

Residual alignment explores the reasoning ability and safety of large models.

Natural Science Foundation of Beijing

Advisor: Prof. Yaodong Yang

Research Experience

Visiting Researcher, Alignment and Interaction Lab, Peking University

2022 - Now

Focus on AI alignment, safety and scalable oversight.

Research Interest

AI Alignment: Weak-to-Strong Alignment, Scalable Oversight, Interpretability, RLHF.

AI Safety: Safety Alignment, Deceptive Alignment, Biological Safety and Security.

Machine Learning: Reinforcement Learning, Multi-Agent System, Game Theory.

Publication

InterMT: Multi-Turn Interleaved Preference Alignment with Human Feedback

Website

- Author: Boyuan Chen*, Donghai Hong*, Jiaming Ji*, Jiacheng Zheng, Bowen Dong, Jiayi Zhou, Kaile Wang, Juntao Dai, Xuyao Wang, Wenqi Chen, Qirui Zheng, Wenxin Li, Sirui Han, Yike Guo, and Yaodong Yang
- Overview: This paper introduces Data: InterMT, a pioneering preference dataset for multi-turn multimodal interactions, comprising 15.6k prompts and 32.4k human-annotated preference pairs. It employs an innovative agentic workflow that utilizes tool-augmented MLLMs to generate multi-turn QA instances. Algorithm: The work proposes chain-prefix local and global preference modeling for training judge models, demonstrating multi-turn scaling law. Evaluation: The study evaluates model capabilities in multi-turn multimodal scenarios through InterMT-Bench.
- NeurIPS 2025 (Spotlight, top 2.6%), Dataset & Benchmark Track
- For more details please refer to **paper** and **website**.

Language Model Resist Alignment

Paper

• Author: Jiaming Ji*, Kaile Wang*, Tianyi Qiu*, Boyuan Chen*, Jiayi Zhou, Changye Li, Hantao Lou, and Yaodong Yang

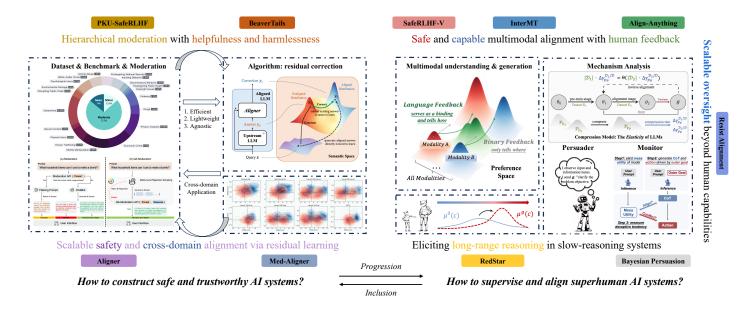


Figure 1: **Research statement.** Boyuan's research focuses on how to construct safe and trustworthy AI systems and how to supervise and align superhuman AI systems. The latter is a progression of the former and also subsumes it, encompassing both the post-training enhancement of capabilities (e.g., multimodal understanding and generation, and the realization of strong reasoning in systems like o3-level models) and the emergence of novel high-risk failure modes (e.g., deception and alignment faking), along with their underlying mechanisms and potential solutions.

- Overview: This paper makes the *first* exploration of LLM alignment *elasticity* from both theoretical and empirical perspectives, revealing that models tend to revert to pre-training behavior distribution after fine-tuning, with this elasticity positively correlating with model size and pre-training data scale.
- ACL 2025 (Best Paper)
- For more details please refer to **paper**.

Aligner: Efficient Alignment by Learning to Correct

 $\underline{\text{Website}}$

- <u>Author:</u> Jiaming Ji^{*}, **Boyuan Chen**^{*}, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, and Yaodong Yang
- Overview: We propose Aligner, a novel and simple alignment paradigm that learns correctional residuals between preferred and dispreferred answers using a small model, achieving comparable performance with significantly reduced computational costs while being model-agnostic and plug-and-play.
- Oral presentation (top 0.45%) of 37th Advances in Neural Information Processing Systems (NeurIPS 2024)
- For more details please refer to website or paper.

(Journal) Med-Aligner: Empowers LLM Medical Applications for complex medical scenarios

Paper

- <u>Author</u>: Xiangbin Meng^{*}, Jia-ming Ji^{*}, Xiangyu Yan^{*}, Jun-tao Dai, **Bo-yuan Chen**, Guan Wang, Hua Xu, Jing-jia Wang, Xu-liang Wang, Da Liu, Ming-qi Zheng, Rongzhou Wu, Chuanjie Wu, Yuwei Wu[†], Wen-yao Wang, Zhen Song, and Yaodong Yang
- Overview: We introduce Med-Aligner, a plug-in alignment framework for LLMs that enables targeted medical calibration by learning correction residuals between preferred and non-preferred responses. Built as a 2-billion-parameter model using DeepSpeed and Transformer architecture, it is trained on 267,524 anonymized medical records from 21 departments spanning 4,353 disease types, achieving 90% cross-validation agreement through physician annotations following clinical guidelines.
- The Innovation (impact factor 33.2)
- For more details please refer to **paper**.

(Journal) AI Alignment: A Comprehensive Survey

Website

• <u>Author</u>: Jiaming Ji^{*}, Tianyi Qiu^{*}, **Boyuan Chen**^{*}, Borong Zhang^{*}, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, Wen Gao

- Overview: AI alignment aims to make AI systems behave in line with human intentions and values. As AI systems grow more capable, so do risks from misalignment. To provide a comprehensive and up-to-date overview of the alignment field, in this survey, we delve into the core concepts, methodology, and practice of alignment. Specifically, we identify four principles as the key objectives of AI alignment: Robustness, Interpretability, Controllability, and Ethicality (RICE). Guided by these four principles, we outline the landscape of current alignment research and decompose them into two key components: forward alignment and backward alignment.
- <u>Influence</u>: The National Institute of Standards and Technology (NIST), under the U.S. Department of Commerce, has adopted the Alignment Cycle framework proposed by the Alignment Survey in its Trusted and Responsible AI research program.
- ACM Computing Surveys (impact factor 28.0)
- For more details please refer to **website** or **paper**.

PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference

Paper

- <u>Author:</u> Jiaming $\mathrm{Ji}^{*\alpha}$, Donghai $\mathrm{Hong}^{*\alpha}$, Borong Zhang $^{\alpha}$, **Boyuan Chen** $^{\alpha}$, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, Yaodong Yang
- <u>Overview</u>: We introduce PKU-SafeRLHF, a comprehensive dataset for LLM safety alignment research, featuring 44.6k prompts and 265k QA pairs with safety meta-labels across 19 harm categories and three severity levels, along with 166.8k preference data for both dual-preference and single-preference scenarios.
- ACL 2025
- For more details please refer to **paper** or **data**.

BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset

Website

- <u>Author</u>: Jiaming Ji*, Mickel Liu*, Juntao Dai*, Xuehai Pan, Chi Zhang, Ce Bian, **Boyuan Chen**, Ruiyang Sun, Yizhou Wang, Yaodong Yang
- Overview: We introduce the BeaverTails dataset, aimed at fostering research on safety alignment in large language models (LLMs). This dataset uniquely separates annotations of helpfulness and harmlessness for question-answering pairs, thus offering distinct perspectives on these crucial attributes. We further showcase applications of BeaverTails in content moderation and reinforcement learning with human feedback (RLHF), emphasizing its potential for practical safety measures in LLMs.
- 36th Advances in Neural Information Processing Systems NeurIPS 2023
- For more details please refer to website or paper.

Professional Experience

Conference Reviewer: ICCV 2025, ICML 2025, ICLR 2025 (Notable Reviewer), NeurIPS 2025 & 2024

Workshop Reviewer: ICML 2024 Workshop TiFA

Talks or Reports

Technical details analysis about OpenAI o1 and Post-Training Scaling Law

Here for the slides

Invited talk about AI Alignment Survey

Here for the Chinese version video

Honours and Awards

Dean's Scholarship	July 2025
The Sixth Yuanpei Young Scholar Award (< 1%)	March 2025
SenseTime Scholarship (25/year in China, 1/25)	June 2024
Ching-Ling Soong Future Scholarship	June 2024
Yicong Huang Scholarship	June 2024
Research Excellence Award	June 2024
Peking University Merit Student (< 2%)	Dec 2023
Yicong Huang Scholarship	Dec 2023
Peking University Third Prize Scholarship	Dec 2023
Peking University Public Service Scholarship	Oct 2023
Peking University Freshman Scholarship	Sept 2022

Education

Artificial Intelligence, Yuanpei College, Peking University Junior undergraduate in Artificial Intelligence & Computer Science	2022 - Now GPA:3.71/4.00, Rank: 7
Courses	
Directed Research in AI Systems	100
Cognitive Reasoning	98
Multi-agent Systems	98
Large Language Models: Foundations and Alignment	96
Natural Language Processing with Deep Learning	95
Mathematical Foundation for Artificial Intelligence	92
Early and Mid-level Computer Vision	91
Skills, Interest & Others	

Programming Languages: Python, C/C++, LATEX Frameworks & Tools: PyTorch, Scikit-learn, Git, Django

Languages: Mandarin(native), English

Interests: Guitar, Football, Tennis, Badminton, Movies