Tianyi Qiu

qiutianyi.qty@gmail.com | tianyiqiu.net | github.com/TianyiQ | Google Scholar (420 citations)

Research Experience

AI Safety Fellow (incoming), Anthropic 2025
Conducting frontier AI safety and alignment research (acceptance rate < 1%). *London, UK*

Research Intern, Center for Human-Compatible AI, UC Berkeley 2024
Led research on a game theory method for LLM scalable oversight, with Prof. Stuart Russell. *Berkeley, CA, USA*

Visiting Researcher, Alignment and Interaction Lab, Peking University 2023 – 2024
Led multiple research projects on LLM alignment, with Prof. Yaodong Yang. *Beijing, China*

Grants Received

PI, NSFC Youth Scientist Grant 2024 – 2026
LLM alignment from non-stationary feedback. *National Natural Science Foundation of China (**UKRI-endorsed**)*

co-PI, Lambda Research Grant 2025 – 2026
Characterizing the impact of LLM alignment strategies on user ideation. *Lambda Inc. (USA)*

Education

Peking University 2022 – 2026 (est.)
BSc in Computer Science, member of the Turing Experimental Class. *Beijing, China*

University of California (UCEAP Program) 2024
Selected as reciprocity exchange student to the UC system. *CA, USA*

Selected Awards

Best Paper Award (sole authorship), NeurIPS 2024 Pluralistic Alignment Workshop 2024

Rising Star in Computer Science Research, Peking University 2024

NeurIPS 2024 Scholar Award (merit-based financial aid grant), NeurIPS Program Committee 2024

Award for Scientific Research, Peking University 2023

John Hopcroft Scholarship, Center on Frontiers of Computing Studies, Peking University 2023

Gold Medal, Chinese National Olympiad in Informatics 2020

Grandmaster in Competitive Programming (top 0.1% globally), CodeForces 2019

Publications

[1] **T. Qiu**, Y. Zhang, X. Huang, J. X. Li, J. Ji, Y. Yang (2024). ProgressGym: Alignment with a Millennium of Moral Progress. *NeurIPS 2024 (**Spotlight**, Datasets & Benchmarks Track).*

[2] J. Ji, B. Chen, H. Lou, D. Hong, B. Zhang, X. Pan, **T. Qiu**, J. Dai, Y. Yang (2024). Achieving Efficient Alignment through Learned Correction. *NeurIPS 2024 (**Oral**).*

[3] **T. Qiu** (2024). Representative Social Choice: From Learning Theory to AI Alignment. *NeurIPS 2024 Pluralistic Alignment Workshop (**Best Paper Award**). Under review at the Journal of Artificial Intelligence Research.*

[4] J. Ji, **T. Qiu**, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, F. Zeng, K. Y. Ng, J. Dai, X. Pan, A. O'Gara, Y. Lei, H. Xu, B. Tse, J. Fu, S. McAleer, Y. Yang, Y. Wang, S. C. Zhu, Y. Guo, W. Gao (2023). AI Alignment: A Comprehensive Survey. *Under review at ACM Computing Surveys.*

[5] **T. Qiu**, Z. He, T. Chugh, M. Kleiman-Weiner (2025). The Lock-in Hypothesis: Stagnation by Algorithm. *Accepted to ICLR 2025 BiAlign Workshop. Under review at ICML 2025.*

[6] **T. Qiu**, M. Carroll, C. Allen (2025). Truthfulness Despite Weak Supervision: Evaluating and Training LLMs Using Peer Prediction. *Under review at ICML 2025.*

[7] Z. He, **T. Qiu**, T. Lin, M. Glickman, J. Wihbey, M. Kleiman-Weiner (2025). Position: AI Systematically Rewires the Flow of Ideas. *Accepted to ICLR 2025 BiAlign Workshop. Under review at ICML 2025.*

[8] **T. Qiu**, F. Zeng, J. Ji, D. Yan, K. Wang, J. Zhou, Y. Han, J. Dai, X. Pan, Y. Yang (2024). Reward Generalization in RLHF: A Topological Perspective. *Under review at ACL 2025.*

[9] J. Ji, K. Wang, **T. Qiu**, B. Chen, C. Li, H. Lou, J. Zhou, Y. Yang (2024). Language Models Resist Alignment: Evidence From Data Compression. *NeurIPS 2024 Socially Responsible Language Modelling Research Workshop.*

[10] J. Ji, H. Hong, B. Zhang, B. Chen, J. Dai, B. Zheng, **T. Qiu**, B. Li, Y. Yang (2024). PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference. *Under review at ACL 2025.*

[11] J. Ji, J. Zhou, H. Lou, B. Chen, D. Hong, X. Wang, W. Chen, K. Wang, R. Pan, J. Li, M. Wang, J. Dai, **T. Qiu**, H. Xu, D. Li, W. Chen, J. Song, B. Zheng, Y. Yang. Align Anything: Training All-Modality Models to Follow Instructions with Language Feedback. *Under review at ACL 2025.*

Open-Source Projects

ProgressGym Models & Datasets (paper, huggingface with 20,000+ downloads) 2024 – 2025
Project lead. LLM moral alignment with temporal data. Downloads summed over 8 months.

Align-Anything (paper, github with 3,000+ stars) 2024 – 2025
Library for post-training SOTA multi-modal foundation models at scale.

Professional Service

Reviewing:

Mar 2025 ICML 2025
Jun 2024 NeurIPS 2024
Sep 2024 NeurIPS 2024 Pluralistic Alignment Workshop
Sep 2024 DAI 2024

Contributed Talks:

Dec 2024 NeurIPS 2024 Pluralistic Alignment Workshop (link)
May 2024 VAISU 2024 (link)

Invited Talks:

Jun 2025 University of Washington (upcoming)
Jun 2024 PKU Turing Forum
Nov 2023 Stanford AI Alignment

Mentoring:

Feb 2024 – May 2024 PKU Alignment and Interaction Lab