

Mickel Liu

✉ mickelliu7@gmail.com | 🏠 mickel-liu.github.io | 🎓 Mickel Liu

Education

W Paul G. Allen School, University of Washington

PH.D. STUDENT IN COMPUTER SCIENCE, CO-ADVISORS: [TIM ALTHOFF](#), [NATASHA JAQUES](#)

Seattle, WA

Sep. 2024 - 2029

- **Research Interest:** AI Alignment, LLM post-training (RLHF), Multi-agent Learning

School of Computer Science, Peking University

MASTER'S IN COMPUTER SCIENCE, ADVISORS: [YIZHOU WANG](#), [YAODONG YANG](#) (Co.)

Beijing, PRC

Sep. 2020 - Jan. 2024

- **Research Interest:** AI Alignment, Mutli-Agent Learning, Reinforcement Learning, Large Language Models

Faculty of Applied Science and Engineering, University of Toronto

B.A.S. IN APPLIED CHEMISTRY AND CHEMICAL ENGINEERING (MINOR IN AI ENGINEERING)

Toronto, ON

Sep. 2015 - Aug. 2020

Internship Experience

AllenNLP @ Allen Institute for AI (AI2)

STUDENT COLLABORATOR, HOST: NATHAN LAMBERT

Seattle, WA

May. 2024 - Jul. 2024 (2 months)

H2Lab @ University of Washington

VISITING RESEARCH STUDENT, HOST: PROF. HANNANEH HAJISHIRZI

Seattle, WA

Jan. 2024 - Jul. 2024 (6 months)

Baichuan Intelligent Technology

INTERN, LARGE LANGUAGE MODEL (LLM) - ALIGNMENT RESEARCH

Beijing, PRC

Jun. 2023 - Sep. 2023 (3 months)

Beijing Institute of General Artificial Intelligence (BIGAI)

RESEARCH SCIENTIST INTERN, MULTI-AGENT LEARNING (MAL) LAB

Beijing, PRC

Mar. 2022 - Jun. 2023 (15 months)

Cenovus Energy Inc.

DATA SCIENTIST INTERN (CO-OP), PROCESS AND AUTOMATION TEAM (SUNRISE)

Calgary, AB

Sep. 2018 - Aug. 2019 (12 months)

Publications

(* DENOTES EQUAL CONTRIBUTION)

BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset ([Link](#))

Jiaming Ji*, **Mickel Liu***, Juntao Dai*, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, Yaodong Yang

NeurIPS 2023 Poster

Safe-RLHF: Safe Reinforcement Learning From Human Feedback ([Link](#))

Juntao Dai*, Xuehai Pan*, Ruiyang Sun*, Jiaming Ji*, Xinbo Xu, **Mickel Liu**, Yizhou Wang, Yaodong Yang

ICLR 2024 Spotlight

Baichuan 2: Open Large-scale Language Models ([Link](#))

Mickel Liu, with 54 other authors (alphabetical ordering)

Technical report in public archive

Proactive Multi-Camera Collaboration For 3D Human Pose Estimation ([Link](#))

Hai Ci*, **Mickel Liu***, Xuehai Pan*, Fangwei Zhong, Yizhou Wang

ICLR 2023 Poster

MATE: Benchmarking multi-agent reinforcement learning in distributed target coverage control ([Link](#))

Xuehai Pan, **Mickel Liu**, Fangwei Zhong, Yaodong Yang, Song-Chun Zhu, Yizhou Wang

NeurIPS 2022 Poster

Skills & Background

Frameworks & Tools	PyTorch, Ray & RLlib, DeepSpeed, 🧠 Transformer & 🧠 Diffuser
Engineering	Python, Conda, bash, Slurm, Linux, Docker
Domain Knowledge	Natural Language Processing (NLP), Deep Reinforcement Learning (RL), Multi-Agent Learning, Human Pose Estimation, Convex Optimization, Statistical Learning
Technical Expertise	LLM Post-Training, LLM Training Optimization, Distributed Training, Prompting Engineering, Data Annotation Project Management
Languages	English (Native), Mandarin (Native)

Service

Reviewer	NeurIPS 2022-24, ICLR 2024-25, ICML 2023, AAAI 2024
----------	---

Awards

2024-25	Paul G. Allen School CSE Fellowship (given to selected first-year Ph.D. students)
2023	NeurIPS Scholar Award
2021-23	Chinese Scholarship Council (CSC) Scholarship