Li Ding

liding256@gmail.com Mountain View, CA https://liding.info

Research Summary

The goal of my research is to align AI systems with human values and enable them to safely learn in open-ended environments. I develop algorithms to enhance the safety and creativity of generative models through reinforcement learning from human feedback (RLHF), alongside methods to improve the generalization of foundation models. My work spans multiple domains, including large language models (LLMs), image generators, and robotics.

Education

University of Massachusetts Amherst
Ph.D. in Computer Science 2020.9 - 2024.7
- Dissertation: "Optimization with intrinsic diversity: Towards generalizable, safe, and open-ended learning".
- Committee: Lee Spector (Chair), Scott Niekum, Subhransu Maji, Jeff Clune.

Massachusetts Institute of Technology
Graduate Study in EECS (non-degree) 2019.9 - 2020.1

University of Rochester
M.S. in Computational Science 2016.6 - 2017.5
- Advisor: Chenliang Xu

Work Experience

Google
Machine Learning Engineer 2024.7 - present
- Conducted research to enhance the performance and robustness of multimodal LLMs, focusing on parameter-efficient fine-tuning (PEFT) and quantization-aware training (QAT).
- Designed and implemented hardware-accelerated LLM agents, optimizing performance on edge devices for scalable and reliable deployment.

Massachusetts Institute of Technology
Research Engineer 2017.9 - 2020.6
- Performed research in deep learning for spatiotemporal perception and led the development of the MIT DriveSeg dataset for autonomous driving scene segmentation.
- Developed a multitask learning framework for joint face landmark detection and cognitive state assessment, improving both performance and efficiency.

Internship

Google
Research Intern 2023.6 - 2023.9
- Developed a JAX optimizer for efficient neural architecture search of foundation models through knowledge distillation and meta-learning.

CarperAI
Student Researcher 2023.2 - 2023.6
- Proposed novel algorithms to enhance the creativity of generative models through RLHF, leading to more diverse and open-ended model behavior.

Meta
Research Scientist Intern 2022.5 - 2022.8
- Developed model architectures and transfer learning methods to optimize vision transformers for AR/VR capabilities.

Selected Publications

Google Scholar

- R. Boldi, L. Ding, L. Spector, and S. Niekum, "Pareto-optimal learning from preferences with hidden context," *under review, preliminary version to appear at **NeurIPS**: Pluralistic Alignment Workshop,*, 2024
  - POPL learns Pareto-optimal policies or reward functions in RLHF, addressing diverse group preferences without needing group labels, and offering safe and fair alignment of agents and LLMs.
- L. Ding, J. Zhang, J. Clune, L. Spector, and J. Lehman, "Quality diversity through human feedback: Towards open-ended diversity-driven optimization," in ***ICML** & **NeurIPS**: ALOE Workshop **(Spotlight)***, 2024
  - QDHF learns diversity metrics from human feedback and optimizes the discovery of novel solutions, enhancing task-solving capabilities of RL agents and boosting creativity of generative models.
- L. Ding, E. Pantridge, and L. Spector, "Probabilistic lexicase selection," in ***GECCO***, 2023
- L. Ding, M. Zoghi, G. Tennenholtz, and M. Karimzadehgan, "Ever evolving evaluator: Towards flexible and reliable meta-optimization for knowledge distillation," in ***NeurIPS**: RealML Workshop*, 2023
- L. Ding, J. Terwilliger, and *et al.*, "CLERA: A unified model for joint cognitive load and eye region analysis in the wild," *ACM Trans. Computer-Human Interaction (**TOCHI**)*, 2023
- L. Ding and L. Spector, "Optimizing neural networks with gradient lexicase selection," in ***ICLR***, 2022
- L. Ding, J. Terwilliger, R. Sherony, B. Reimer, and L. Fridman, "Value of temporal dynamics information in driving scene segmentation," *IEEE Trans. Intelligent Vehicles (**T-IV**)*, 2021
- L. Fridman, L. Ding, B. Jenik, and B. Reimer, "Arguing machines: Human supervision of black box AI systems that make life-critical decisions," in ***CVPR** Workshops*, 2019
- L. Ding and C. Xu, "Weakly-supervised action segmentation with iterative soft boundary assignment," in ***CVPR***, 2018

Honors and Awards

Google Research Travel Scholarship, *Google*.	2023
SOAR (Supporting Open Access Research) Fund, *UMass Amherst*.	2023
4th Place (among 150 teams, top 3%), *MIT Miniplaces Challenge*.	2019
Graduate School Scholarship ($30,000), *University of Rochester*.	2016
Meritorious Winner (top 5% worldwide), *COMAP's Mathematical Contest In Modeling*.	2015

Teaching

TA for MIT 6.S094: Deep Learning for Self-Driving Cars.	2018 - 2019
TA for MIT 6.S099: Artificial General Intelligence.	2018
Co-instructor (w/ Tom Bertalan) for MIT Robocar Workshop.	2018

Community

Reviewer for ICLR, NeurIPS, JMLR, CVPR, ICCV, ECCV, *etc.*	2020 - present
Ph.D. Admissions Committee (UMass Amherst CICS).	2024

Open Source Projects

`pyribs`: an open-source library for diversity-driven optimization.

- Contributed the demo and tutorial for the QDHF work.

`mit-deep-learning`: MIT Deep Learning open courses (10k+ stars).

- Main contributor for tutorials and coding assignments.

`MIT AI Podcast`: now the *Lex Fridman Podcast*, 4M+ subscribers on Youtube.

- Helped find candidates and prepare interview materials in early episodes.

Skills

Python, C++, JavaScript, PyTorch, JAX, Tensorflow, Git, LLM workflows (instruction-tuning, RLHF, PEFT, quantization).