Li Ding

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

SUMMARY

My research focuses on aligning AI systems with human values and enable safe, self-directed agent learning in open-ended environments. I developed methods using reinforcement learning from human feedback (RLHF) to enhance the safety, creativity, and generalization of AI agents and large language models (LLMs).

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

- Contributed to various Gemini post-training efforts, including performance optimization, model compatibility, and quality/safety evaluation.
- Led the development of Gemini models in PyTorch, optimizing it for efficient inference and fine-tuning on various hardware accelerators including next-generation chipsets.
- Developed novel methods for fine-tuning multimodal LLMs, focusing on unifying parameter-efficient fine-tuning (PEFT), quantization-aware training (QAT), and multi-task learning.
- Designed evaluation pipeline for multimodal LLMs involving quality and safety tests.

Massachusetts Institute of Technology
Research Engineer 2017.9 - 2020.6

- Developed novel methods in deep learning for video scene segmentation, led the development of the MIT DriveSeg dataset for autonomous driving scene segmentation.
- Developed a multi-task computer vision framework for joint face landmark detection and cognitive state assessment, improving both performance and efficiency.

SELECTED PUBLICATIONS

GOOGLE SCHOLAR

- R. Boldi, L. Ding, L. Spector, and S. Niekum, "Pareto-optimal learning from preferences with hidden context," *preprint (under review), arXiv:2406.15599*, 2024
  - POPL learns Pareto-optimal policies/rewards in RLHF, catering diverse group preferences without needing group labels, thus offers safe and fair alignment of RL 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 exploration of novel solutions, enhancing task-solving of RL agents and creativity of generative models.

- L. Ding, E. Pantridge, and L. Spector, "Probabilistic lexicase selection," in ***GECCO***, 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 and L. Spector, "Evolutionary quantum architecture search for parametrized quantum circuits," in ***GECCO** Companion*, 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. Ding and C. Xu, "Weakly-supervised action segmentation with iterative soft boundary assignment," in ***CVPR***, 2018

Internships

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 vision transformers architectures and transfer learning methods for AR/VR.

Honors and Awards

Google Research Travel Scholarship (NeurIPS), *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

Services

Reviewer for ICLR, NeurIPS, ICML, 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 code, demo, and tutorial for QDHF.

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