Aaron (Jiaxun) Li

(+1) 510-520-5351 jiaxun_li@g.harvard.edu Cambridge, MA
Google Scholar Github LinkedIn

Education

Harvard University September 2023 - May 2025 (Expected)
M.E. Computational Science and Engineering (Thesis Track)
Cross-Registered at MIT EECS
GPA: 3.91/4.00

University of California, Berkeley August 2019 - May 2023
B.A. Computer Science (EECS Honors), GPA: 3.92/4.00
B.A. Psychology, GPA: 3.90/4.00

Research Interests

Explainable AI, Mechanistic Interpretability, Trustworthy AI, Large Language Models

Research Experience

AI4LIFE Research Group, Harvard University Sep. 2023 - Present
Graduate Student Researcher, advised by Prof. Himabindu Lakkaraju

- **RLHF's Impact on Language Model Trustworthiness**
  Conducted the first systematic evaluation of RLHF's impact on trustworthiness, revealing conflicts between alignment goals and dataset limitations; introduced a novel influence function-based data attribution method for RLHF, which enables downstream data-level mitigation.
- **Unified Evaluation for Robustness of Sparse Autoencoders (In Progress)**
  Explored the limitations of sparse autoencoders by evaluating the robustness of their generated concept-level interpretations of pretrained LLMs; working on efficient input-level attacks that manipulate the neuron activation patterns in the sparse latent representations.
- **Chain-of-Thought (CoT) in Weak-to-Strong Generalization (In Progress)**
  Evaluated the change in LLM performance with CoT prompting as the result of weak-to-strong generalization; working on explaining the changes in reasoning coherence and CoT faithfulness.
- **Certified LLM Defense**
  Provided certified robustness guarantees for empirical defense procedures against adversarial prompting targeting LLMs. Developed efficient variants of certifiable safety-checking algorithms.

Yu Group, UC Berkeley Aug. 2022 - Aug. 2023
Undergraduate Researcher, advised by Prof. Bin Yu

- **Efficient Concept-level Debugging for Prototype-based Neural Networks**
  Improved model interpretability of widely used prototype-based CNNs by aligning generated visual explanations with collected human preferences. Proposed the Reward-Reweighing, Reselecting, and Retraining (R3) debugging framework, which uses reward models trained with human feedback to perform corrective updates, improving both predictive performance and interpretability.

Extended Course Project, Harvard University Oct. 2023 - May. 2024
Advised by Prof. Finale Doshi-Velez

- **Interpretable Inverse Reinforcement Learning via Reward Decomposition**
  Designed an interpretable inverse reinforcement learning framework with reward decomposition, enabling transparent decision-making explanations and allowing users to evaluate and critique the trustworthiness of model outputs in high-stakes scenarios.

Shanghai AI Lab Jun. 2023 - Sep. 2023
Research Intern @ Speech Group

- **Post-hoc Evaluation of Content and Speaker Information**
  Used post-hoc explainability methods such as LIME and Shapley Values to analyze state-of-the-art text-to-speech and voice conversion frameworks, proposing an empirical gradient-based evaluation metric to quantitatively measure the disentanglement of content and speaker information.

Ponce Lab, Harvard University Jun. 2022 - Dec. 2022
Undergraduate Researcher, advised by Prof. Carlos R. Ponce

- **Online Input-level Neuron Control**
  Extended existing online neuron control algorithms from the continuous space to the discrete image space of fixed datasets. Proposed and implemented a GAN inversion method that leverages local geometric properties in the latent feature space, allowing for the adaptation of continuous methods to a discrete setting.

Publications

[1] On the Inherent Instability of Sparse Autoencoders
Aaron J. Li, Suraj Srinivas, Himabindu Lakkaraju
Paper in preparation, planned submission to ICML 2025

[2] More RLHF, More Trust? On the Impact of Preference Alignment on Trustworthiness
Aaron J. Li, Satyapriya Krishna, Himabindu Lakkaraju
Under review at ICLR 2025, Top 3% average score

[3] Improving Prototypical Visual Explanations with Reward Reweighing, Reselection, and Retraining
Aaron J. Li, Robin Netzorg, Zhihan Cheng, Zhuoqin Zhang, Bin Yu
ICML 2024

[4] Certifying LLM Safety Against Adversarial Prompting
Aounon Kumar, Chirag Agarwal, Suraj Srinivas, **Aaron J. Li**, Soheil Feizi, Himabindu Lakkaraju
COLM 2024

Teaching Experience

Course Staff @ UC Berkeley EECS Department
CS 170: Efficient Algorithms and Intractable Problems (Fall 2021)
CS 188: Introduction to Artificial Intelligence (Summer 2021)
CS 70: Discrete Mathematics and Probability Theory (Summer 2020)

Skills

Programming Languages: Python, Java, C++, C, MATLAB, R
Frameworks: PyTorch, CUDA, TensorFlow, Keras, Gym, Ray, etc.
Tools & Utilities: Git, Slurm, Conda, Bash, Jupyter, tmux, SQL, etc.

Coursework

Undergraduate: Machine Learning, Deep Learning, Computer Vision, Reinforcement Learning, Probability and Random Processes, Convex Optimization, Signal Processing, Efficient Algorithms, Human Neuroanatomy, Neuroimaging, Computational Models of Cognition
Graduate: Inverse Reinforcement Learning, Sensorimotor Learning, Spoken Language Processing, Geometric Machine Learning, Efficient Machine Learning