

Aaron (Jiaxun) Li

(+1) 510-520-5351 aaronjli@berkeley.edu Berkeley, CA
Google Scholar Github LinkedIn

Education

University of California, Berkeley

August 2025 - Present

Ph.D. in Computer Science

Advisors: Prof. Bin Yu and Prof. Ion Stoica

GPA: 4.00/4.00

Harvard University

September 2023 - May 2025

M.E. in 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. in Computer Science (EECS Honors), GPA: 3.92/4.00

B.A. in Psychology, GPA: 3.90/4.00

Research Interests

LLM Evaluation, AI Safety, Agentic AI, Trustworthy Machine Learning, Interpretability

Publications

- [1] Evaluating Adversarial Robustness of Concept Representations in Sparse Autoencoders
Aaron J. Li, Suraj Srinivas, Usha Bhalla, Himabindu Lakkaraju
EACL 2026
- [2] More RLHF, More Trust? On the Impact of Preference Alignment on Trustworthiness
Aaron J. Li, Satyapriya Krishna, Himabindu Lakkaraju
ICLR 2025 (Oral Presentation), Top 1.8%
- [3] Improving Prototypical Visual Explanations with Reward Reweighting, 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
- [5] Speaker Disentanglement of Speech Pre-trained Model Based on Interpretability
Xiaoxu Zhu, Junhua Li, **Aaron J. Li**, Yiming Ren, Baoxiang Li
Preprint, 2025

Research Experience

UC Berkeley

Aug. 2025 - Present

Affiliated with Sky Computing Lab and BAIR; advised by Prof. Ion Stoica and Prof. Bin Yu

- **Mutual Evaluation Between Tasks and LLMs**

Proposed a unified framework that jointly estimates task difficulties and LLM capabilities on a shared latent scale, which supports both static benchmarks and open-ended LM Arena logs; explored the potential of using decisions given by LLM judges to replace human feedback.

- **Evaluation of LLMs for Realistic Medical Diagnosis**

Proposed a scalable LLM evaluation framework for realistic medical diagnosis tasks when a single definite ground truth does not exist as in traditional benchmarks; conducted case studies on how common user biases might influence evaluation results, and provided practical user-centric guidelines.

- **Rethinking LLM Sycophancy**

Empirically characterized LLM sycophancy across domains and interaction settings, and examined its connections to model reliability, uncertainty, and latent knowledge boundaries.

Harvard University

Sep. 2023 - May 2025

Advised by Prof. Hima 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**

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 space.

- **Certified LLM Defense**

Provided certified robustness guarantees for empirical defense procedures against adversarial prompting targeting LLMs; developed efficient variants of certifiable safety-checking algorithms.

UC Berkeley (Undergraduate Researcher)

Aug. 2022 - Aug. 2023

Advised by Prof. Bin Yu

- **Efficient Concept-level Debugging for Prototype-based Neural Networks**

Improved the model interpretability of widely used prototype-based CNNs by aligning generated visual explanations with collected human preferences; proposed the Reward-Reweighting, Reselecting, and Retraining (R3) debugging framework, which uses reward models trained with human feedback to perform corrective updates, improving both predictive performance and interpretability.

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