Haoyu Li

haoyuli5@illinois.edu | (424) 293-9235 |
412 E Healey St. Apt 306, Champaign, Illinois | Personal Website | Google Scholar

Education

University of Illinois, Urbana-Champaign Aug 2024 - May 2029 (Expected)

- PhD in Computer Science
- Advisor: Professor Huan Zhang

University of California, Los Angeles Sep 2020 – June 2024

- B.S. in Mathematics
- GPA: 3.93/4.0

Research Interests

I am broadly interested in machine learning, especially in building stronger and more reliable machine learning models. My recent research centers on the following two directions:

- Trustworthiness of LLMs, in particular for the current Large Reasoning Models.
- Verification of machine learning models in safety-critical scenarios, e.g. safety/stability/robustness guarantees for learning based autonomous systems.

Publications & Preprints

(* indicates equal contribution, for a full list see my Google Scholar)

- DecepChain: Inducing Deceptive Reasoning in Large Language Models — Under Review 2025
  Wei Shen*, Han Wang*, **Haoyu Li***, Huan Zhang. [PDF], [Project Page], [Code]
- On The Fragility of Benchmark Contamination Detection in Reasoning Models — Under Review 2025
  Han Wang*, **Haoyu Li***, Brian Ko*, Huan Zhang. [PDF], [Code]
- Learning to Learn a Zeroth-Order Optimizer for Fine-tuning LLMs — Under Review 2025
  Kairun Zhang*, **Haoyu Li***, Yanjun Zhao*, Yifan Sun, Huan Zhang. [PDF], [Code]
- Two-Stage Learning of Stabilizing Neural Controllers via Zubov Sampling and Iterative Domain Expansion — NeurIPS 2025 (Spotlight)
  Haoyu Li*, Xiangru Zhong*, Bin Hu, Huan Zhang. [PDF], [Code]
- Safe Domains of Attraction for Discrete-Time Nonlinear Systems: Characterization and Verifiable Neural Network Estimation — CDC 2025
  Mohamed Serry*, **Haoyu Li***, Ruikun Zhou*, Huan Zhang, Jun Liu. [PDF], [Code]
- Neural Contraction Metrics with Formal Guarantees for Discrete-Time Nonlinear Dynamical Systems — L4DC 2025
  Haoyu Li*, Xiangru Zhong*, Bin Hu, Huan Zhang. [PDF]
- Predicting and Interpreting Energy Barriers of Metallic Glasses with Graph Neural Networks — ICML 2024
  Haoyu Li*, Shichang Zhang*, Longwen Tang, Matheiu Bauchy, Yizhou Sun. [PDF], [Code]
- Interpretability through Training Samples: Data Attribution for Diffusion Models — TMLR 2024
  Tong Xie*, **Haoyu Li***, Andrew Bai, Cho-Jui Hsieh. [PDF], [Code]

Research Experience

Large Language Models

- Reasoning model contamination (co-first author): Showed that even brief GRPO can conceal contamination signals introduced during SFT contamination; proposed theoretical analysis that pin the effect on PPO-style importance-sampling/clipping. Paper under review.
- DecepChain (co-first author): Introduced a backdoor that makes CoT look benign while flipping the final answer, by exploiting LLM's own hallucination with SFT on self-generated wrong rollouts and GRPO with a flipped verifiable reward; achieves >95% attack success and non-differentiable human trust compared to the

benign case. Paper under review.

- ZO-Finetuner (co-first author): Proposed a compact learned zeroth-order optimizer that learns perturbation strategies once per LLM that can transfer across tasks; outperforms previous ZO baselines in 4 LLMs $\times$ 7 datasets across model sizes with minimal time/memory overhead. Paper under review.

Learning-Based Control & Verification

- Two-Stage Neural Controller (first author, NeurIPS'25 Spotlight): Proposed a novel two stage training pipeline for co-learning controllers and neural certificates, and strengthened α,β-CROWN pipeline for fast continuous-time verification; yields ROA volumes $5 - 1.5 * 10^5$ times larger than baselines and 40-10,000 times faster verification than the commonly used SMT solver dReal.
- Neural Lyapunov function for Discrete-Time Systems (co-first author, CDC'25): Extended Zubov theorem to the discrete-time setting; Implemented scalable verification pipeline with α,β-CROWN, demonstrated much greater scalability compared to dReal.
- Neural Contraction Metrics (first author, L4DC'25): Proposed a new Jacobian-/LMI-free sufficient condition for contraction in discrete-time systems, enabling scalable certification with non-smooth neural network controllers with ReLU activations.

Awards

- NeurIPS 2025 Scholar Award
- First place in the 6th International Verification of Neural Networks Competition (VNN-COMP 2025) for both the regular and extended tracks. Member of team alpha-beta-CROWN.
- L4DC 2025 Travel Grant

Services

- Reviewer for NeurIPS 2025, ICLR 2025-2026, L4DC 2025-2026
- Assistant instructor for UCLA Olga Radko Endowed Math Circle, 2022-2024