Marvin Li
marvinfli.github.io | marvinli@college.harvard.edu | LinkedIn | Google Scholar | Github

EDUCATION

Harvard College, *BA in Computer Science and Math, summa cum laude, GPA: 4.0* 09/21 – 05/25

Selected courses: Math 55a (Algebra & Group Theory) and b (Topology & Complex Analysis); *Math & Engineering Principles for Training Foundation Models*; *Foundations of Deep Learning*; *Deep Learning (MIT)*; *High-Dimensional Probability*; *Statistical Inference I*; *Algorithms for Data Science*; *Computational Learning Theory*; *Algorithmic Statistics (MIT)*; Systems Programming; Machine Learning; *Optimized Democracy*; *Economic Analysis as a Frontier of Theoretical CS*; *Fairness and Validity*; Functional Analysis. *Italics denotes graduate-level.*
Teaching Fellow: Data Structures & Algorithms; *Algorithms for Data Science.*

TECHNICAL SKILLS

python, bash, R, java, C++, SQL; pytorch, transformers, trl, diffusers, wandb, pandas, numpy, scikit-learn

RESEARCH EXPERIENCE

Harvard SEAS, *Research Assistant for Prof. Sitan Chen* 10/2023 – NOW

- Pioneered theory to explain *critical windows*, the sudden formation of features like answer correctness, reward hacking, or toxicity in generative models, with probability theory (**ICML 2024** [1], **ICML 2025 oral** [2]).
- Created new methods to identify LLM jailbreaks and reasoning failures based on the presence of critical windows.
- Invented a new zero-shot method to detect individual samples from training data for CIFAR-10 diffusion models by using critical windows around training points in `diffusers` library.

Harvard Secure and Fair Machine Learning Lab, *Research Assistant for Prof. Seth Neel* 02/2023 – NOW

- Developed MOPE, a novel zero-shot privacy attack against LLMs that identifies members of pre-training data by exploiting gradient information (**EMNLP 2023** [3]).
- Developed FLoRa, a loss ratio attack which extracts $> 90\%$ of LLAMA-7B's finetuning dataset.
- Contributed to a comprehensive codebase and library to evaluate different privacy attacks against LLMs.

Harvard SEAS, *Research Assistant for Prof. Cynthia Dwork* 02/2024 – NOW

- Analyzed how geometric bias in prompt embeddings leads to representational bias of race in text-to-image diffusion models using Lipschitz-based conditions on generative models in **ICML 2024 workshop** [4].

WORK EXPERIENCE

Hudson River Trading, *Quantitative Researcher Intern* 05/2024 – 08/2024

- Designed research pipeline to quickly test and iterate on research ideas for alpha discovery and modeling.
- Discovered new medium-frequency signals for equities by crafting features from alternative data sources.

Citadel Securities, *Quantitative Researcher Intern* 05/2023 – 08/2023

Two Sigma, *Software Engineering Intern* 05/2022 – 08/2022

- Created an internal workflow API in Java/SQL to liquidate out-of-universe securities.

HONORS AND AWARDS

Academic: Captain Jonathan Fay Prize (given to the three best theses of Harvard College's graduating class); Hoopes Prize; Detur Book Prize (top 6%, '23); John A. Harvard Scholar (top 5%, '22-'24).
Research: Regeneron Science Talent Search Finalist ('21), Research Science Institute ('20).

PUBLICATIONS (* denotes equal contribution.)

[1] Marvin Li and Sitan Chen. "Critical windows: non-asymptotic theory for feature emergence in diffusion models". In: *International Conference on Machine Learning*. 2024.

[2] Marvin Li, Aayush Karan, and Sitan Chen. "Blink of an eye: a simple theory for feature localization in generative models". In: *International Conference on Machine Learning*. **Oral, top 1%**. 2025.

[3] Marvin Li*, Jason Wang*, Jeffrey Wang*, and Seth Neel. "MOPE: Model Perturbation based Privacy Attacks on Language Models"". In: *Large Language Models and the Future of NLP: Main Conference of EMNLP 2023*. Association for Computational Linguistics, Dec. 2023.

[4] Sahil Kuchlous*, Marvin Li*, and Jeffrey G. Wang*. "Bias Begets Bias: the Impact of Biased Embeddings on Diffusion Models". In: *ICML Trustworthy Multi-modal Foundation Models and AI Agents workshop*. 2024.

[5] Marvin Li, Patricia Glibert, and Vyacheslav Lyubchich. "Machine Learning Classification Algorithms for Predicting Karenia brevis Blooms on the West Florida Shelf". In: *J.Mar.Sci.Eng.* 9.9 (2021). ISSN: 2077-1312.

[6] Jeffrey Wang*, Jason Wang*, Marvin Li*, and Seth Neel. "Firm Foundations for Membership Inference Attacks Against Large Language Models". In: *International Conference on Machine Learning*. Workshop on Data in Generative Models. 2025.