# AVIDAN (AVI) SHAH

65 Dorison Drive, Short Hills NJ 07078 | amcshah@berkeley.edu |shavidan123.github.io| 973-902-3032

## EDUCATION

**University of California, Berkeley |** May 2025 **Cumulative GPA:** 3.8

B.A. Computer Science & B.A. Applied Mathematics (Data Science Concentration)

**Awards & Honors**: UPE CS Honor Society, UCB Dean's List, National Merit Finalist, Nationally Certified EMT

## PROFESSIONAL / RESEARCH EXPERIENCE

**Millennium Management** June – August 2023, June – August 2024

*Data Science and Quantitative Research Intern* *New York, NY*

- Designed and implemented a generative adversarial network for unsupervised anomaly detection on market data
- Built an automated classifier to detect information that would result in disruption of data delivery or PM trading
- Used natural language processing techniques on textual market data to generate and evaluate trading signals
- Performed data ingestion and assisted with the maintenance of systematic data pipelines used daily by over 300 different investment teams

**Berkeley Artificial Intelligence Research (REDS Group)** September 2022 – Present

*Undergraduate Researcher* *Berkeley, CA*

- Conducting research as part of Professor David Wagner's group in both ML for security and security for ML, working with Julien Piet and Chawin Sitawarin
- Designing deep learning based frameworks for SSH anomaly detection under multiple threat models
- Developed automated jailbreaking algorithm for adversarial suffix generation on large language models

**MIT Data to AI Laboratory** May – August 2022

*Undergraduate Researcher* *Cambridge, MA*

- Researched and developed ML pipelines for more accurate unsupervised anomaly detection in public time series data using the lab's open-source Orion anomaly detection library
- Built a fully automated, end-to-end workflow for continuously updating public data acquisition, model driven anomaly detection, and visualization via GitHub pages

**Berkeley Department of EECS** January – May 2024

*Course Reader, ENG 125* *Berkeley, CA*

- Graded assignments, held office hours, and managed student questions as a course reader. Assisted professor and graduate student instructors with discussion and lecture.

## PROJECTS / PUBLICATIONS

**Stronger Universal and Transfer Attacks by Suppressing Refusals** February 12, 2025

NAACL 2025 Main Conference Poster

- Created IRIS attack algorithm for state of the art jailbreak success rate on GPT-4o and other proprietary LLMS via internal refusal representation inhibition
- Discovered the universal phenomenon of jailbreaking suffixes and that individual behavior optimization can be suboptimal, even when transferring to black-box models.
- Abridged version of paper also accepted to NeurIPS 2024 SafeGenAI Workshop

**Deep Learning for SSH Traffic Anomaly Detection** September 2022 – Present

- Developed multiple supervised and unsupervised learning models for time series data to detect network intruders using inter-keystroke timings and packet lengths in SSH connections in single and multi-site settings
- Paper will be submitted in early 2025, preprint will be available on arxiv

**Efficient Bus Bunching Mitigation through Automated Curriculum Learning** December 2023

CS285 (Deep Reinforcement Learning) Final Project, arxiv preprint

- Tested a novel approach to curriculum learning utilizing adversarial setter model to increase bus system efficiency
- Adversarial curriculum setter model performs well on custom bus environment without requiring extensive domain knowledge or training, paper available on arxiv.

## SKILLS, PERSONAL INTERESTS

**Skills:** Python, Java, SQL, PyTorch, Pandas, Deep Learning, LLM Agents, Spanish (Limited)

**Interests:** Strategy Games, Piano, Swimming, Fiction Writing, Emergency Medicine, Sigma Chi Fraternity