Lawrence Feng

925-819-8690 lawrence.feng2017@gmail.com lawrencefeng17.github.io

Education

Carnegie Mellon University **Fall 2022 – Spring 2026**

Bachelor of Science in Statistics and Machine Learning *Pittsburgh, Pennsylvania*

- Additional major in Artificial Intelligence and minor in Mathematics
- GPA: 3.90

Relevant Coursework

- Introduction to Deep Learning (PhD)
- Introduction to Machine Learning (PhD)
- Probabilistic Graphical Models (PhD)
- Probability and Mathematical Statistics (PhD)
- Algorithm Design and Analysis
- Parallel and Sequential Data Structures
- Advanced Methods for Data Analysis
- Monte Carlo Methods and Applications

Experience

Inflection AI **June 2025 – Present**

Member of Technical Staff Intern

- Contributing to post-training infrastructure and alignment systems for large language models.

Carnegie Mellon Computer Science Department **Fall 2024 – Present**

Research assistant advised by Professor Wenting Zheng

- Investigating novel AI watermarking approaches by combining cryptographic theory with empirical ML methods
- Developed novel attack methods exposing vulnerabilities in existing work through systematic validation
- Synthesizing complex experimental results into clear visualizations and accessible summaries for research collaboration
- Discovered empirical parameters that cryptographic theory must guarantee for effective watermarking implementation
- Sole programmer managing all aspects: experimental design, implementation, analysis, and progress reporting

Robotics Institute at Carnegie Mellon **Spring 2024 – Fall 2024**

SURA Summer Research Intern

- Developed and automated experimental frameworks to evaluate multimodal, generative transformer-based robot trajectory planners, enabling efficient management of long-running experiments
- Gained expertise in navigating and contributing to complex research codebases with limited documentation

Carnegie AI Safety Initiative **Fall 2023 – Present**

Incoming President

- Leading weekly alignment reading groups and active member of the executive team driving campus engagement

Projects

Interpreting Vision Language Models | *Python, Hugging Face, TransformerLens* **Fall 2024**

- Led a project investigating vision language models using tools like Hugging Face and TransformerLens.
- Found that a language-only sparse autoencoder can provide insights into a multimodal model's black-box behavior
- Demonstrated the ability to alter model behavior predictably by intervening on intermediate activations

Solving Jigsaw Puzzles using Reinforcement Learning | *Python, PyTorch, Pandas, Git* **Fall 2023**

- Led a team of four in designing, implementing, and testing a deep reinforcement learning system inspired by AlphaGo
- Achieved $> 90\%$ on image reassembly task through innovative model architecture and tree search integration
- Managed project timeline, code integration, and technical direction while coordinating team efforts

MyTorch | *Python, NumPy* **Fall 2023**

- Neural network library implements MLPs, CNNs, RNNs, GRUs, and reverse automatic differentiation
- Demonstrated deep understanding of ML fundamentals through rigorous implementation of core components

Awards

Program on AI and Reasoning (PAIR) **Summer 2024**

- Accepted on a full scholarship to a competitive 2-week camp focusing on AI, cognition, and rationality

Technical Skills

Programming Languages: Python, R, C, C++, Java, SQL, SML, LaTeX
Developer Tools: Google Cloud Platform, Amazon AWS, Git, WandB, VSCode, Jupyter Notebook
Frameworks and Libraries: PyTorch, TensorFlow, Hugging Face, OpenCV, NumPy, Pandas, Matplotlib, Seaborn, ROS