

Anthony Lu

New York, NY | luanthony523@gmail.com | (+1) 737-484-6609 | [LinkedIn](#) | [GitHub](#) | [luanthony.xyz](#) | U.S. Citizen

Education

New York University

New York, NY

B.A. Computer Science and Data Science, Minor in Mathematics - **GPA 3.76/4.00**

Expected December 2026

- **Coursework:** Deep Learning & LLMs, Machine Learning, Programming Languages, Parallel Computing, Computer Systems, Algorithms, Data Structures, Linear Algebra, Probability & Statistics, Calculus III, Differential Equations.

Experience

NYU SAI Lab - Prof. Sai Qian Zhang, Prof. Yuzhang Lin

New York, NY

Undergraduate Researcher

September 2026 - Present

- Investigating **LLMs** and **reinforcement learning** for optimizing power-grid restoration and resource allocation post-disaster.
- Researching post-training methods and communication protocols for multi-agent systems in simulated environments.

InterSystems

Boston, MA

Software Engineer Intern

May 2026 - August 2026

- Developed an agentic AI assistant (**ObjectScript**, **Python**) for generating data transformations within InterSystems Interoperability, accelerating development time by **16x** over manual authoring.
- Built validation tools (**Rust**) and context framework, improving **accuracy by 15%+** and **token efficiency by 35%+**.
- Engineered an automated synthetic data generation and **LLM evaluation** harness across multiple client use cases, evaluating performance metrics and surfacing domain-specific model failure modes.

DermImpact

New York, NY

Founding Engineer

February 2026 - May 2026

- Built and deployed a full-stack platform (**Next.js**, **TypeScript**) for AI-assisted dermatology notes and image analysis, validated in a live pilot with a **30+ physician dermatology clinic**.
- Architected privacy-focused data pipelines with RLS, encrypted image storage, and PostHog usage analytics.

NYU Biokind Analytics

New York, NY

Data Engineer

January 2026 - May 2026

- Built FastAPI-based **data pipelines** processing **35k+** nonprofit donor records, automating database maintenance and reducing manual processing time by **60%** while increasing platform engagement by **20%**.

Pulse Foundry

New York, NY

Software Engineer Intern

December 2025 - February 2026

- Built and deployed an automated content recommendation pipeline (**FastAPI**, **OpenAI**) with **concurrent ingestion**, **semantic ranking**, and **scheduled workers**, serving **3,000+** subscribers in production.

Projects

hLLM | C++, Python, CUDA, MLX, gRPC

August 2026 - Present

- Developed a heterogeneous **LLM inference** runtime that partitions models across Apple Silicon and NVIDIA GPUs, executing **distributed** prefill and autoregressive decoding over gRPC.
- Implemented native **CUDA** and **MLX** backends with Safetensors weight loading and cross-device activation streaming.
- Built hardware-aware profiling and placement to optimize model partitioning across memory, compute, and network costs.

HyperZero | Python, PyTorch, FastAPI, React, Three.js

May 2026 - June 2026

- Built a **reinforcement learning agent** combining Monte Carlo Tree Search, self-play, and policy-value neural networks.
- Achieved **87% mean win rate** against heuristic and tree-search baselines across multiple game configurations.
- Developed a training pipeline optimized for high-throughput with **batched inference** and **parallelized simulations**.

Neural-ViT | Python, PyTorch, Vertex AI, BigQuery, Google Cloud Storage

December 2025 - January 2026

- Implemented a **3D Vision Transformer** for genotype classification on multi-trial neural spectrograms, achieving **0.96 AUC** (+9% over baseline models), trained on NVIDIA A100s via Google Cloud Vertex AI.
- Built an end-to-end ML pipeline processing **77k+** trials with preprocessing, experiment tracking, and reproducible training.

Skills

Languages:	Python, C++, C, JavaScript, TypeScript, Java, SQL, Rust, Go, HTML, CSS
ML/Data:	PyTorch, scikit-learn, Weights & Biases, Pandas, NumPy, Hugging Face Transformers
Frameworks:	React, Angular, Node.js, Next.js, FastAPI, Express.js, PostgreSQL, SQLite
Infrastructure:	GCP, AWS, Docker, Git, Linux, CUDA