Ryan Rochmanofenna

rochmanofenna@gmail.com · 301–814–9421 · github.com/rochmanofenna · linkedin.com/in/ryanroch

Education

New York University — New York, NY
B.A. in Computer Science & Mathematics, Minor in Philosophy — *Expected May 2026*

- **Relevant Coursework:** Linear Algebra, Cryptography (Number Theory), Algorithms, Data Structures, Operating & Distributed Systems, Discrete Mathematics, Natural Language Processing, Combinatorics
- **Honors:** Dean's List; $50,000 Annual Merit Scholarship

Technical Skills

Math: Linear Algebra, Probability, Optimization, Numerical Methods, Stochastic Differential Equations,

Programming: Python, C/C++ (CUDA/low-latency), Rust, SQL, JavaScript, Bash, Go

ML: PyTorch, JAX, TensorFlow, scikit-learn, Optuna; GNNs, Transformers, Neural ODEs

Systems: CUDA/Triton kernels, Docker, Kubernetes, Linux, Git, Ray; AWS (EC2, S3, Lambda); HPC (NYU Greene)

Experience

Stealth Buy-Side Research Stack — New York, NY
Systems Engineer — ML, Low-Latency Infrastructure, and Alpha/Strategy Research — *Jun 2024 – Present*

- Architected an end-to-end research and execution stack unifying order book reconstruction, microstructure features, and venue-level analytics with sub-20 ms p99 latency.
- Designed cointegration- and mean-reversion strategies with statistically rigorous cross-validation; integrated Markowitz-style portfolio optimization and transaction cost models to control variance without eroding returns.
- Engineered custom CUDA kernels for Monte Carlo simulation and order book transforms, achieving 10× speedups over NumPy baselines through numerical methods and parallel stochastic integration.
- Delivered a live monitoring suite for P&L, VaR, drawdowns, and slippage; secured $45K in non-dilutive R&D funding to extend the platform's research capacity.

Sending Labs — Remote
Machine Learning Engineer (Contract) — *Jun 2025 – Aug 2025*

- Designed ManimGL pipeline generating protocol SDK visualizations and developer documentation from structured prompts.
- Integrated distributed GPU rendering with Modal/Fly.io and NVENC acceleration, boosting throughput ∼5× and cutting costs 35%.
- Contributed to open platform specifications and core infrastructure tooling under Linx Technologies.

Video Tutor AI — Remote
Machine Learning Engineer (Contract) — *Apr 2025 – Jun 2025*

- Architected an AI educational content pipeline using GPT-4o, TTS, and ManimGL for CFA/K-12 lessons.
- Deployed Redis-backed task queue for 500+ concurrent render jobs; cut compute spend by ∼40% via caching and batching.
- Optimized GPU worker containers for deterministic frame-level outputs across workloads.

Olo — New York, NY
Software Engineering Intern — *May 2022 – Aug 2022; Jun 2023*

- Built ARIMA and Prophet forecasting pipelines to anticipate kiosk order surges; improved RMSE by ∼28%.
- Applied $M/M/c$ queueing models to capacity allocation to mitigate throttling and revenue risk.

Research & Projects

EEG 2025: Contradiction-Aware Neural Pipeline — *NYU Greene HPC*

- Adapted BICEP→ENN→Fusion Alpha framework from quant trading to EEG IV-2a (129 channels).
- Combined stochastic path simulation, temporal encoders, and graph fusion with contradiction operators for cross-subject alignment.
- Achieved ∼60% faster training on 8×V100 cluster via custom CUDA kernels and 1000+ Optuna trials.

GPU Monte Carlo Engine for Derivative Market Price Forecasting (BICEP)

- Implemented CUDA-accelerated Euler–Maruyama path simulation with Sobol sequences and variance reduction, extending the stack's numerical methods layer for pricing Asian and Barrier options.
- Verified convergence rates and profiled kernel efficiency, demonstrating faster error decay than naive sampling under identical budgets; results fed into downstream portfolio risk and alpha research workflows.
- Extended randomness layer with cryptographic PRGs (AES-CTR/ChaCha20) to ensure i.i.d. Gaussian increments, eliminating correlation artifacts in large GPU ensembles.

Custom Neural Network (ENN) for Lightweight Sequence Modeling

- Developed a compact recurrent cell with PSD-constrained entanglement matrix and collapse head, enabling parameter-efficient sequence learning with stable gradients.
- Implemented in C++/Eigen with full backpropagation-through-time, spectral regularization, and AdamW optimization; profiled for GPU/CPU training efficiency.
- Integrated into the buy-side research stack to model order book dynamics and contradiction-aware sentiment features with low-latency inference.
- Validated ENN's cross-domain viability by deploying in a real-time gesture recognition system, achieving 98% accuracy at 25 FPS with <1 MB weights — $8\times$ smaller and $5\times$ faster than baseline CNN-LSTM models.

Leadership & Activities

- **NYU Hyperloop** (2021–2022): Control Systems.
- **NYU Tandon Made Challenge** (2020–2021): Winner and 2x finalist; awarded $5k pre-seed for biomedical/computer vision venture.
- **Additional:** NYU Web Publishing Consultant; Greek Life (Nu Alpha Phi)