Nikhil Reddy

732-691-6442 | nikhilalwaysreddy@gmail.com | github.com/NickelReade

Education

University of Chicago (Transfer from NYU) Chicago, IL

Bachelor of Science in Computer Science, Mathematics, Economics, Combined GPA: 3.76 *Expected Graduation: May 2026*

- Paragon Global Investments - Quant Analyst, Maroon Capital - Quant Analyst
- Operating Systems/Systems Programming, Algorithms, Computer Architecture, Data Structures, Object-Oriented Programming, Discrete Mathematics, Machine Learning in Finance, Databases, Real Analysis, Fundamentals of Machine Learning, Differential Equations, Numerical Analysis, Markov Chains, Brownian Motion
- Machine Learning Club - President, Poker Club - Vice President, WorldQuant, IMC Prosperity, Akuna Options 201

Technical Skills

Languages: C, C++, C#, Perl, Python, Java, Javascript, Lisp, Ruby, React, Qiskit, SQL, Kotlin, Go
Tools/Frameworks/Libraries: Git, Docker, Azure, Django, TensorFlow, Pytorch, MATLAB, Kubernetes

Experience

AI/ML Engineer Intern June 2025 – Present

Amazon Web Services

- Collaborated with data scientists to productionize Jupyter-based notebooks into Athena-queryable pipelines, using AWS Step Functions and Glue jobs.
- Engineered Athena-compatible feature extraction layer using SQL + PySpark UDFs for downstream model training on Amazon SageMaker, reducing feature pipeline latency by 35%.

Quantitative Research Intern December 2024 – March 2025

Blockhouse

- Designed and implemented smart order routing algorithm using Kyle's model and eigenliquidity principles that led to a projected $160M in cost savings and later adopted by Tradeweb to optimize their trading operations.
- Used Proximal Policy Optimization and Temporal Fusion Transformers for time-series along with the Almgren-Chriss model to analyze over 1TB of market data for market impact optimization, resulting in a 22% reduction in average transaction costs and a 35% improvement in latency metrics.
- Integrated Monte Carlo simulations for order execution modeling and used LightGBM for real-time order placement, increasing order fill rates by 18% and reducing slippage by 12%. Researched jump-diffusion models for price dynamics and transient market impact models, improving execution efficiency and liquidity resilience by 27%.

Software Engineer II Intern (Generative AI Team) May 2024 – August 2024

Walmart Global Tech

- Reduced average response times by 20% on Walmart's 'Troubleshooter' Generative AI with caching and nearest neighbor algorithms using Java Spring Boot, redis, scikit-learn, and DBeaver
- Streamlined testing and deployment workflows by automating unit tests with JUnit and implementing CI/CD pipelines using Docker and Kubernetes, reducing downtime and increasing efficiency by 30% via Jira
- Developed an endpoint, processing logic, and Helm chart for the "Troubleshooter" generative AI system, utilizing REST APIs and Retrieval-Augmented Generation (RAG) processing to support over 2 million requests per day
- Created E2E tests using Ginko and Gomega to achieve a coverage of 97% on 5 codebases of over 10000 lines

Projects

Quantitative Trading Project | *Python, C++, MXNet, XGBoost, Reinforcement Learning* November 2023

- Utilized C++ and MXNet for PPO used in cryptocurrency markets, delivering a consistent annualized return of 15% and a maximum drawdown of 7% on personal account. Also optimized to a processing time of 75ns
- Achieved top 5% ranking in WorldQuant Brain by developing predictive financial models using Python, XGBoost and recurrent neural networks, resulting in a Sharpe ratio of 1.6 and an ROI of 12% on simulated trading strategies.

C++ Projects | *C++, Multithreading, Networks, Distributed Systems* October 2023

- Created distributed web crawler that searched 1000 pages each of 200M which was used on 20 machines
- Implemented highly-concurrent network file system in C++ and Boost with complex locking system
- Developed thread-safe virtual memory manager in C++ which serviced concurrent mmap requests
- Wrote threading library in C++ which included condition variables, scheduling, and mutexes

Art Pricing Project (AMD/Cloudera Hackathon) | *Kaggle, ImageAI, TensorFlow, PyTorch, Git, HTML/CSS* April 2023

- Developed an art pricing detection model using Kaggle, ImageAI, TensorFlow, and PyTorch. Estimated prices of given sample art pieces consistently within a 20% margin to achieve 2nd place in the AMD/Cloudera ML hackathon
- Created comprehensive UI and frontend interface using JavaScript, HTML, CSS, etc to present to judges.