Benjamin Shvartsman

717-379-9809 | b.shvartsman@columbia.edu | linkedin.com/in/bshvartsman/ | github.com/frogsterr

Education

Columbia University — New York City, NY
Bachelors in Computer Science, Jewish History, Computational Biology — *GPA: 3.9/4.0*

- **Amazon Future Engineer Scholar**: Highly competitive (1 of 400) national scholarship for exceptional leadership potential and academic excellence in computer science.
- **Y Combinator Startup School**: 1 of 2,000 hand-selected internationally for mentorship from tech industry titans including Musk, Nadella, and Altman
- **Columbia Blockchain Analyst**: Managing a portfolio of 40 ETH ($100,000) and conducting in-depth research on emerging blockchain projects and cryptocurrencies to inform investment

Experience

Google — Summer 2025
Software Engineer Intern — *Sunnyvale, CA*

- Led 3-team development of an agentic investigation system on GCP's Agent Development Kit—projected $100K annual savings and 500+ developer hours..
- Developed distributed TPU training pipeline with gRPC server integration, implementing self-training via knowledge distillation and RLHF to improve SRE effort-estimation accuracy by 33%.
- Implemented Cloud SQL embedding storage solution for RAG workflows, optimizing semantic similarity search through vector approximation algorithms to achieve 4× faster query response.

Nearly Human — Summer 2023, 2024
Data Scientist Intern — *Harrisburg, PA*

- Utilized Langchain to create synthetic QA model for training client-specific LLMs on 10,000+ documents.
- Migrated inference engine to Microsoft Azure AI's GPT, incorporating ETL pipelines to streamline data ingestion and preprocessing, slashing average inference time by 83% and significantly improving response quality.
- Implemented multi-threaded model inferencing to reduce API response times 5x during peak loads.

Projects

High-Performance Backtesting and Order Management System | *Python, Pandas, C++, Multi-threading*

- Implemented a high-speed order matching engine in C++ with a latency of ∼1 microsecond per trade match.
- Developed a synthetic market data generator using Python and Pandas, enabling realistic backtesting scenarios for high-frequency trading strategies.
- Enhancing system performance through parallelization and multi-threading techniques, optimizing the order management system for handling large volumes of trades in microsecond timeframes.

Real-Time ASL Translation Chrome Extension | *TensorRT, Pytorch, AWS, Docker*

- Developed optimized object classification pipeline achieving 60 FPS inference throughput on V100 GPU implementing custom CNN model.
- Leveraged TensorRT for model optimization and CUDA acceleration. Reduced latency by 4x compared to unoptimized PyTorch model through quantization, layer fusion and kernel auto-tuning.
- Partnered directly with deaf and sign language communities to research needs, evaluate and continuously collect feedback to enhance accuracy throughout development.

Technical Skills

Languages: Java, Python, Javascript, Go, SQL (MySQL, Postgres), HTML/CSS
Frameworks: Flask, PyTorch, React.js, Node.js, FastAPI, BootStrap
Developer Tools: Git, AWS (EC2, S3), Azure, Google Cloud Platform, Slack, Jira, Docker, Linux
Libraries: pandas, NumPy, Matplotlib, Selenium, PySpark