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, Biology — *Expected Grad: May 2027*

- **GPA: 3.9 / 4.0**; Dean's List
- **Awards:** Amazon Future Engineer Scholar (1 of 400 National), List College Fellow (1 of 3 at Columbia University)
- **Relevant Coursework:** Data Structures and Algorithms, Machine Learning for Functional Genomics, Advanced Systems Programming, Fundamentals of Computer Systems, Linear Algebra, Probability, Calculus III, OOP

Experience

NVIDIA — New York, NY
Machine Learning Systems Engineer Intern — *January 2026 – Current*

- Spearheading GPU-accelerated optimizations for Retrieval-Augmented Generation (RAG) pipelines

Google — Sunnyvale, CA
Software Engineer Intern — *May 2025 – August 2025*

- Leading cross-team development of an agentic investigation system for SREs on GCP's Agent Development Kit—projected **$150K annual savings** and 100+ developer hours saved.
- Fine-tuned Gemini via LoRA and knowledge distillation, achieved **22% improvement** in reasoning accuracy.
- Built distributed TPU training pipeline, processing 2,000+ bug tickets per weekly cycle.
- Optimized similarity search for RAG via vector approximation algorithms to achieve **4× faster query response**.

Nearly Human — Harrisburg, PA
Data Science Intern — *June 2023 – September 2024*

- Led development of client-specific LLM solutions for two private clients (>$10B AUM), including a banking RAG system projected to **save $1M+ annually** in onboarding and compliance workflows.
- Utilized LangChain to create an agentic system for training client-specific LLMs on 10,000+ documents.
- Implemented multi-threaded model inferencing to **increase API throughput 5×** during peak loads.
- 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 improving response quality.

Projects

Low-Latency Cloud Gaming Streaming Platform | *C++, WebRTC, NVENC, Windows API*

- Built a GeForce NOW-style game streaming system using NVIDIA NVENC H.264/H.265 hardware encoding to compress 1080p60 gameplay with sub-20ms latency.
- Implemented DirectX Desktop Duplication API for efficient screen capture and WebRTC data channels for bidirectional input transmission (keyboard/mouse) between client and host.
- Optimized network performance using adaptive bitrate streaming and TURN/STUN servers for NAT traversal, achieving playable latency (50ms RTT) over typical home networks.

Technical Skills & Interests

Spoken Languages: Fluent in English, Russian; Conversational in French, Hebrew
Programming Languages: Java, Python, JavaScript (Node.js), Go, SQL (MySQL, Postgres), HTML/CSS
Frameworks: Flask, PyTorch, Spark, React.js, FastAPI, Agile
Developer Tools: Git, AWS (EC2, S3), Azure, Google Cloud Platform, Slack, Jira, Docker, Linux, UNIX
Libraries: pandas, NumPy, Matplotlib, Selenium, PySpark
Interests: Classical Piano, Brazilian Samba, Men's Gymnastics, Ice Hockey, Pickleball, Golf, Linguistics