

Benjamin Shvartsman

717-379-9809 — b.shvartsman@columbia.edu — [linkedin.com/in/bshvartsman](https://www.linkedin.com/in/bshvartsman) — github.com/frogsterr

EDUCATION

Columbia University

B.A./M.S. in Computer Science, Jewish History, Biology

New York City, NY

Expected May 2028

- **GPA: 3.9/4.0**; Dean's List
- **Awards:** ICPC, Amazon Future Engineer Scholar (1 of 400 nationally), List College Fellow
- **Coursework:** Data Structures & Algorithms, ML for Functional Genomics, Advanced Systems Programming, Computer Architecture, Linear Algebra, Probability

EXPERIENCE

YouTube

Machine Learning Software Engineer Intern

Mountain View, CA

May 2026 – Present

- Co-architected a Transformer-XGBoost classification pipeline for copyright-spam detection with **99% precision**, orchestrating inference for **1M+ weekly claims** within a high-throughput C++ takedown aggregator.
- Refactored the TensorFlow feature-registration workflow within a C++ scheduler, **reducing code footprint by 40–60%** and optimizing downstream ingestion latency.
- Engineered an automatic prompt-optimization pipeline using **NSGA-II** and Pareto ranking for dynamic prompt-topology selection, improving safety and accuracy by **18%**.

Columbia Creative Machines Laboratory

Machine Learning Research Intern

New York, NY

May 2026 – Present

- Formulated a mechanistic-interpretability framework to track topological drift of safety guardrails during continual multilingual training of Transformers.
- Optimized distributed training for 1B+ parameter models with **GaLore** and **BF16**, writing custom CUDA and Triton kernels to improve GPU utilization and training throughput.

NVIDIA

ML Systems Engineer Intern, Research

New York, NY

January 2026 – April 2026

- Pioneered neuro-symbolic causal-graph reasoning for GPU-kernel crash diagnosis by modeling cross-subsystem dependencies and defect patterns; achieved **35% faster root-cause resolution** than embedding-based retrieval.
- Achieved up to **50× greater token efficiency** for tool-augmented code reasoning by architecting a gRPC LLVM/Clang semantic-index server, replacing $O(n)$ retrieval with graph-structured LSP queries and saving **20M+ tokens/month**.
- Delivered a **9× startup speedup** for internal ML tooling through lazy-import refactoring and OS-level caching, diagnosed with cProfile, Nsight, and Windows Performance Analyzer.

Google

Software Engineer Intern

Sunnyvale, CA

May 2025 – August 2025

- Led cross-team development of an agentic SRE investigation system on GCP's Agent Development Kit, projected to save **\$150K annually** and 100+ developer hours.
- Fine-tuned Gemini with LoRA and knowledge distillation, improving reasoning accuracy by **22%**; built a distributed TPU pipeline processing **2,000+ weekly bug tickets**.
- Optimized approximate-vector retrieval for RAG, delivering **4× faster query response**.

PROJECTS

ML-to-Production Autonomous Driving Behavior Mining | JAX, C++, XLA, Waymax, HDBSCAN

- Built a batched JAX/XLA pipeline in Waymax to replay **25K+ Waymo driving scenarios** and compare a behavior-cloned trajectory policy against a deterministic C++ planner at **8× simulation throughput**.
- Clustered **12K+ policy disagreements** across 14 safety, comfort, and route-progress features; surfaced six failure modes and ported gap-acceptance and yielding logic into C++, reducing collision/off-road events by **17%** with sub-millisecond overhead.

TECHNICAL SKILLS & INTERESTS

Programming Languages: C/C++, Python, CUDA, Triton, Java, Go, SQL, R, TypeScript

ML & Data: PyTorch, JAX, TensorFlow, DeepSpeed, scikit-learn, pandas, NumPy, Spark, OpenCV

Developer Tools: Git, Docker, Linux, Nsight, cProfile, GCP, AWS, Azure, HPC clusters

Interests: Ironman Triathlon, Watchmaking, Quantitative Finance, Classical Piano, Tennis