Nicholas Bykhovsky

(415) 806-3192 | nico@bykhovsky.com | nicobykhovsky.com | github.com/Bykho

Summary

Masters Student at Columbia University studying Computer Science (ML track). Experience working with Apache Spark, PyTorch, HuggingFace, and founded Silo—a portfolio generator company widely used at Columbia. Experienced in ML research for healthcare data, GPU accelerated programming (CUDA), and LLM enabled product design.

Education

Columbia University, Fu Foundation School of Engineering and Applied Sciences New York, NY
MS in Computer Science, Machine Learning Track, GPA 4.0 *Aug 2024 – Dec 2025*
Courses: High Performance Machine Learning, Embedded Systems, Deep Learning Robot Manipulation, Applied ML

New York University, Courant Institute of Mathematics New York, NY
Double B.A. in Math, Economics; Minor in Computer Science, GPA 3.3 *Aug 2018 – Dec 2022*
Dean's List Fall 2021, Spring 2022. Courses: ML, NLP, Computer Vision, Linear Algebra, Real Analysis, Abstract Algebra

Professional Experience

Silo New York City, NY
Co-Founder and CTO *Nov 2023 – Present*

- Co-founded Silo (silorepo.com), a platform that auto-generates structured portfolios from raw code, papers, and presentations using LLMs. Grew to 1,000+ weekly active users across Columbia's engineering community.
- Built a retrieval system that extended basic RAG using Pinecone and metadata filtering, added structured chunking for long documents, and incorporated memory scaffolding with optional user input to guide multi-stage generation. Used OpenCV OCR to extract text from PDFs/slides, and orchestrated LLM inference across OpenAI, Llama, and Groq APIs to produce clean, structured portfolio entries from varied source formats.
- Led full-stack development across a React frontend, Flask backend, MongoDB database, and Pinecone vector store. Built CI/CD pipelines, implemented unit/integration testing, and integrated Mixpanel for product analytics and user behavior tracking.
- Drove product from idea to live deployment, managing architecture, user onboarding, and iterative feature rollouts.

CodaMetrix Boston, MA
ML Research Intern *Feb 2023 – Sept 2023*

- Designed and implemented an LLM-based NER extraction pipeline for noisy EHR data. Used LangChain and HuggingFace to run quantized models on AWS p3 EC2 instances.
- Improved rare disease code detection by +0.05 AUC and +0.03 F1 through a Bayesian smoothing post-processing layer atop SVM predictions, addressing extreme class imbalance.
- Improved rare code detection using open-source data cleaning libraries applied at scale to noisy EHR datasets via PySpark.

Zetta Venture Partners New York City, NY
Venture Analyst Intern *Jun 2021 – Aug 2022*

- Evaluated 30+ early-stage AI startups weekly, analyzing model architectures, sector trends, and scalability.

Selected Research & Projects

SpecAdapt: Efficient Transformer Inference Columbia University
High Performance ML Research *Jan 2024 – May 2024*

- Built a PyTorch and vLLM benchmarking harness for speculative decoding (Medusa, EAGLE, draft-and-verify), with GPU profiler integration to track SM utilization and memory transfer bottlenecks. Shipped optimizations that improved throughput $\sim 1.6\times$ under fixed compute by tuning draft/verify strategies and KV cache reuse.

Sonar on FPGA Hardware Columbia University
Embedded Systems / SystemVerilog *2023 – 2024*

- Designed and implemented sonar on a DE1-SoC FPGA using an integrated ultrasonic sonar sensing with real-time VGA visualization and servo actuation.
- Developed SystemVerilog modules for sensor interfacing, timing control, and display pipelines; debugged interrupts, signal synchronization, and hardware bring-up at the lab bench.

Interests & Languages

Interests: Surfing, History, Math, Lived in Buddhist Monastery

Skills: Python, Java, C++, Go, SQL, R, React, HTML/CSS, Git, Docker, Spark, OpenCV, PyTorch, SysVerilog

Frameworks/Tools: CUDA, PyTorch Profiler, Nsight Systems, PyVISA, SPICE, Sapien/ManiSkill3, HuggingFace, LangChain, Pinecone, MongoDB, OpenCV. FPGA design with SystemVerilog.