Krish Shah

2003kshah@gmail.com
linkedin.com/in/krish-n-shah | github.com/krish-shah 609-556-8516

Education

Georgia Institute of Technology Atlanta, GA
M.S. in Computer Science (OMSCS), Specialization in Computing Systems; GPA: 4.0/4.0 *Jan 2026 – Present*

Relevant Coursework: GPU Hardware and Software; Advanced Operating System; Computer Networks; ML for Trading

Research: "Contextual-Integrity Gatekeeper for Internal-Channel Privacy in Multi-Agent LLM Systems"

Boston University Boston, MA
B.S. in Computer Engineering, Concentration in Machine Learning *Sept 2022 – May 2025*

FIRST Robotics Presidential Scholar – top 0.5% of engineering admits; four-year half-tuition scholarship, funded research

Experience

Clearwater Analytics New York, NY
Software Engineer *June 2025 – Present*

- **Distributed Agent Orchestration & Execution Engine**: Built a distributed execution engine for 35+ LLM agents with streaming control flow, typed tool calls, and routing across Azure OpenAI and AWS Bedrock. Saved 16+ engineering hours per feature; runs commodity risk workflows for 100+ researchers.
- **LLM Infrastructure & Retrieval Systems (Async, Streaming)**: Built Tornado async APIs with WebSocket streaming, an OpenAI-compatible interface, and FAISS retrieval over financial knowledge bases. Also built a typed tool registry and JSON-RPC MCP server - schema validation, backpressure handling, retry logic - for production use.
- **AI-Driven Stress Testing for Commodity Risk**: Built an app where a meta-agent LLM converts plain-English descriptions into structured shocks across commodity forward curves, power shapes (5x16, 7x8), and vol surfaces, fed into Beacon's full-revaluation VaR engine with daily automated runs via Bob Scheduler — so risk managers can stress test books without writing code.
- **Agentic Copilot for Excel**: Added an AI layer to Beacon by CWAN's Excel Add-In that reads and writes cells mid-conversation, executes 40+ Excel operations via Office.js, and surfaces risk reports, VaR, scenario analysis, and market data through 18 domain skill packages — accessible via sidebar chat or native Excel formula calls.
- **Technical Demoware & Client Presentations**: One of three engineers on the NY team, working across London and Tokyo to build persona-driven demoware on Beacon's platform and present directly to 40+ institutional clients - risk managers, quants, and traders - throughout the sales cycle.

Kolochalama Laboratory Boston, MA
Machine Learning Research Intern *Sept 2024 – May 2025*

- **RAG Pipeline for Biomedical QA**: Built a RAG pipeline using Sentence-BERT embeddings and FAISS over PubMed literature. Cut hallucination rate by 40% vs. baseline; 98.5% citation accuracy on clinical QA.
- **Evaluation Framework & Benchmarking**: Built an evaluation framework measuring hallucination rate, citation accuracy, and BERTScore - made results reproducible and comparable across biomedical LLM runs.

Beacon Platform Inc (Pre-IPO, Series C) New York, NY
Software Engineer Intern *Sept 2023 – Sept 2024*

- **Equity Derivatives Accelerator**: Built an equity derivatives accelerator and scenario generation system; 70% client adoption, 50% faster analysis setup. Covered data ingestion, scenario generation, and analytics output end-to-end.
- **Data Quality Framework**: Built a configurable data quality framework on Great Expectations (chosen over dbt tests after direct comparison) with custom expectation libraries (aggregate, map, regex, query-based), YAML-driven config, and Slack alerting to catch trade and market data issues before they hit downstream risk calculations.

Projects

- **Quorum: Autonomous Multi-Agent Trading System**: Parallel Haiku analyst subagents debate across equities and Kalshi prediction markets; an Opus chairman scores and decides. Quant pre-screen (Altman Z, FCF yield, 12 hard vetoes) anchors LLM output; PreToolUse hooks and live circuit breakers cut in before bad trades clear.
- **Near Real-Time Plasma Diagnostics on HPC (Princeton Plasma Physics Laboratory)**: Built MPI-parallelized processor components in Python for Delta, an adaptive streaming framework bridging magnetic fusion experiments to NERSC supercomputers for sub-10-minute plasma pulse analysis. Scaled linear solvers across multi-node Cori configurations and extended kernels with GPU acceleration.

Technical Skills

- **Languages**: Python, C++, C, SQL, Rust, TypeScript
- **Systems & Infra**: Linux, perf, flamegraphs, CI/CD, Docker, Kubernetes, AWS, GCP, PostgreSQL, Snowflake
- **Frameworks**: React, Next.js, FastAPI, CUDA, OpenMP
- **ML & Data**: PyTorch, NumPy, Pandas, sklearn, FAISS, Hugging Face, Ray

US Citizen | Eagle Scout