Shivang Raval

+1-857-264-8278 | shivangraval50@gmail.com | linkedin.com/in/shivang-raval

Education

Northeastern University, Khoury College of Computer Sciences — Boston, MA
Master of Science in Computer Science — GPA: 4.0/4.0 — *Expected 2025 – 2027*

Pandit Deendayal Energy University — Gandhinagar, India
Bachelor of Technology in Computer Science — GPA: 3.5/4.0 — *2021 – 2025*

Technical Skills

Languages: Python, C++, Rust, OCaml, Java, SQL
Machine Learning: PyTorch, TensorFlow, XGBoost, LightGBM, JAX, scikit-learn, Time Series Forecasting, Model Optimization
ML Infrastructure: GPU Training (CUDA, Multi-GPU), Distributed Training, MLflow, Ray, Kubernetes, Docker, Model Serving
Quantitative Finance: Statistical Arbitrage, Market Microstructure, Execution Algorithms, Kalman Filters, Backtesting Frameworks
Systems Engineering: Low-Latency C++, Multi-threading, Concurrency, Microsecond Scheduling, Performance Optimization
Data Systems: FAISS, PostgreSQL, Kafka, ClickHouse, Real-time Pipelines, High-Frequency Data Processing

Experience

AI Department Manager — Jan 2025 – Jun 2025
Webearl AI — *Remote*

- Led team of 20 ML engineers building production inference systems, collaborating cross-functionally with product and infrastructure teams to ship 8+ models serving 100K+ daily requests
- Architected FastAPI-based model serving platform with LangChain integration achieving sub-500ms p95 latency through inference optimization, smart caching strategies, and A/B testing frameworks
- Scaled team output 50% in 6 months by implementing code review standards, mentoring junior engineers on ML best practices, and establishing rapid experimentation workflows

Data Science Intern — Jun 2024 – Jul 2024
Tor.ai — *Remote*

- Built Random Forest time series forecasting model for electricity demand prediction achieving MAE of 0.12 on production data, improving baseline accuracy by 20% through feature engineering on noisy sensor data
- Deployed automated retraining pipeline with drift detection enabling weekly model updates, collaborating with engineering team to optimize preprocessing for real-time inference

Projects

ML-Driven Statistical Arbitrage Strategy | *Python, PyTorch, LightGBM, Kalman Filters* — 2024

- Developed ensemble trading strategy combining gradient-boosted trees and Kalman filters on 3 years of high-frequency tick data, achieving Sharpe ratio of 1.4–1.6 in historical backtests across multiple market regimes
- Implemented GPU-accelerated feature engineering pipeline processing 50M+ ticks/day with 200+ technical indicators, reducing training time from 12 hours to 45 minutes through batched computation and distributed preprocessing
- Built automated stress testing framework simulating 1–5 sigma market shocks, iteratively improving model robustness and reducing maximum drawdown by 30% through adversarial validation techniques

Low-Latency Market Simulator & Execution Engine | *C++, Python, Multi-threading* — 2024

- Engineered multi-venue trading simulator with microsecond-granularity event scheduling and realistic order book dynamics, improving slippage prediction accuracy by 40% for execution algorithm evaluation
- Designed Python replay API enabling parallel backtesting of 6+ execution strategies across historical data, reducing strategy iteration cycles from 5 days to under 24 hours through efficient data structures
- Optimized C++ core for lock-free concurrency and cache-efficient memory layout, achieving 500K+ events/second throughput while maintaining deterministic replay for ML model training

Distributed ML Training Platform | *PyTorch, Ray, Kubernetes, MLflow* 2024

- Built distributed training infrastructure supporting multi-GPU and multi-node experiments, scaling model training from 8 hours (single GPU) to 45 minutes (16 GPUs) through data parallelism and gradient accumulation
- Implemented automated hyperparameter tuning with Ray Tune managing 100+ concurrent experiments, enabling rapid exploration of model architectures and training configurations for noisy tabular data
- Deployed MLflow tracking system with experiment versioning and model registry, collaborating with 5-person team to establish reproducible ML workflows and continuous model improvement cycles

Smart Order Router with Latency Optimization | *C++, Rust, Lock-Free Data Structures* 2023

- Designed Smart Order Router tracking real-time venue latency and fill probability across 8 exchanges, improving simulated fill rates by 12% through adaptive routing decisions and latency-aware scheduling
- Developed comprehensive fuzzing test suite catching 7 race conditions and concurrency bugs before production deployment, ensuring correctness under high-frequency trading loads

Production RAG System with Active Learning | *Python, FAISS, LangChain, FastAPI* 2023

- Implemented retrieval-augmented generation system with FAISS vector search and semantic reranking, improving retrieval accuracy (MRR) by 35% through embedding fine-tuning and negative sampling
- Launched production chatbot with sub-300ms response time serving real users, establishing weekly RLHF feedback loops that reduced factual errors by 18% through iterative model improvements

Research & Publications

Socio-Economic Benefits of ML Deployment Platforms in Business 2024

- Published in International Journal of Innovative Science and Research Technology (IJISRT)
- Conducted comparative analysis of modern ML deployment platforms (Baseten, RunwayML) versus conventional methods, examining efficiency gains and business implementation strategies
- Analyzed socio-economic impacts of production ML systems, exploring infrastructure trade-offs and deployment patterns relevant to enterprise ML operations

Certifications & Additional Information

Certifications: Google Cloud Certified Machine Learning Engineer
Interests: Algorithmic trading, reinforcement learning, high-performance computing, competitive programming