

SHIVEN KHURDI

Boston, MA (Open to Relocate) | (+1)-617-259-6437 | shiven.khurdi@gmail.com | LinkedIn | Github | Website

Education

Northeastern University

Boston, MA

Master of Science in Artificial Intelligence (*GPA: 3.9*)

Jan 2025 - Present

- **Coursework:** Foundations of AI, Program Design Paradigm, Algorithms, Machine Learning, Cybersecurity, AI Capstone, Computer Vision & Pattern Recognition

Birla Institute of Technology and Science Pilani - Dubai Campus

Dubai, UAE

Bachelor of Engineering in Computer Science (*GPA: 8.22*)

Jan 2021 - Jan 2025

- **Electives:** Deep Learning, Data Mining, Optimization, AI, NLP, Foundations of Data Science

Skills

Languages: Python, SQL, Java, C++, JavaScript, Bash/Shell

ML & Deep Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, Hugging Face Transformers, LLM Fine-tuning, RAG, NLP, Generative AI, Multi-Agent Systems, Prompt Engineering, LLM Evaluation, Reinforcement Learning

Tools & Cloud: Docker, AWS (EC2, ECR, S3, Bedrock), FastAPI, Git, Linux, MLflow, GitHub Actions CI/CD

Libraries & Data: NumPy, Pandas, Matplotlib, OpenCV, LangChain, FAISS, PostgreSQL, MongoDB, React.js, Next.js

Experience

AI Engineering Intern

Jun 2026 - Aug 2026

Interlock Systems Inc. (Harvard Innovation Labs)

Boston, MA

- Building a computer-vision pipeline that flags placement, orientation, and sizing discrepancies between technical design files and their specs, using a fine-tuned object detector with a deterministic verification layer that abstains under uncertainty rather than emit a false verdict.
- Developing agent capabilities, prompts, and integrations for a multi-agent platform that extracts data from engineering documentation (drawings, CAD files, BOMs) and autonomously executes actions in customer ERP and MRP systems; building evaluation harnesses and benchmarks measuring agent performance across extraction, reasoning, and tool-use tasks.

Web Development Intern

Aug 2024 - Jan 2025

ODeX Global

Dubai, U.A.E

- Resolved 500+ frontend bugs across the logistics platform over 6 months, improving UI consistency and reducing QA backlog for the frontend team.
- Built 100+ dynamic form components and a comprehensive operations dashboard from scratch using React.js, streamlining internal workflows for the operations team.

Undergraduate Research Assistant

Feb 2024 - May 2024

BITS Pilani Dubai Campus

Dubai, U.A.E

- Researched efficient LLM fine-tuning under resource constraints (QLoRA, PEFT); fine-tuned Llama-2 7B with 4-bit quantization on a T4 GPU, achieving BLEU scores of 0.5-0.7 on conversational tasks.
- Benchmarked QLoRA against full fine-tuning; authored a technical report and deployed model to HuggingFace.

Projects

ModelRouter: Multi-Model Inference Cost Optimizer [Code]

Mar 2026

- Built an LLM routing system that classifies query difficulty via fine-tuned DistilBERT (96% accuracy, 30ms inference) and dispatches across Llama 8B/70B (Groq) and Claude Sonnet 4 (AWS Bedrock).
- Deployed containerized FastAPI service on AWS with Docker, achieving 58.6% cost savings over an all-Tier-3 baseline on 24 live routed queries with zero errors; built React monitoring dashboard and LLM-as-judge evaluation pipeline.

ReproAgent: ML Paper Reproducibility Analysis [Code]

Feb 2026

- Engineered a multi-agent system that parses arXiv papers, extracts training methodology with per-field confidence scoring, and generates self-contained PyTorch scripts using Llama 3.3 70B via Groq.
- Built keyword-density chunking to prioritize experiment-heavy sections, achieving 0.91 avg extraction confidence and reproducibility scores of 94.3 (ResNet) and 89.1 (Attention Is All You Need); shipped full-stack with FastAPI and Next.js.

CalibratedReturns: Stock Predictions with Coverage Guarantees [Code]

Jan 2026

- Built a conformal prediction system outputting 5-day return intervals with statistical guarantees; trained quantile regression on 20+ engineered features, achieving 92.3% empirical coverage across 3,755 backtested predictions.
- Deployed full-stack application (FastAPI + React) on Render and Vercel for live stock ticker analysis.