Darsh Vora

darshvora.mk@gmail.com | +1 (857) 707-9440 | linkedin.com/in/darsh-vora29 | github.com/Darsh29

PROFESSIONAL SUMMARY

ML/AI Engineer with 2+ years building production LLM systems, deep learning models, and scalable ML pipelines, delivering $229K+ verified business impact. Proficient in Python, PyTorch, and TensorFlow with hands-on experience in GenAI, RAG pipelines, and end-to-end MLOps deployment.

TECHNICAL SKILLS

Programming: Python, C# (.NET), C++, R, Go, React, TypeScript, SQL (PostgreSQL, MySQL), MongoDB, Git
ML Engineering: PyTorch, TensorFlow, Keras, scikit-learn, Ensemble Methods (XGBoost, LightGBM), CNN, LSTM, RNN, CUDA, SHAP, MLflow, OpenCV, spaCy, statsmodels
GenAI & NLP: LLMs, RAG, LangChain, LangGraph, Transformers, BERT, LoRA, QLoRA, PEFT, NLTK, VectorDBs
MLOps & Cloud: Docker, Kubernetes, AWS, Azure, CI/CD, FastAPI, Airflow, Kafka, Terraform, Version Control
Data & Visualization: Snowflake, ETL, Spark, PySpark, dbt, Tableau, Power BI, Streamlit, Plotly

WORK EXPERIENCE

Tatum Robotics (Tech-Telecommunications) — **Boston, MA**
AI Software Engineer — **Aug 2025 - Present**

- Deployed production ASR service processing **500+ daily utterances** with **95%+** accuracy and under **200ms** latency by architecting a WhisperASR pipeline with automated quality validation and **CI/CD** version control
- Enabled real-time ASL translation across **3,000+ phrases** by building a gesture mapping engine on a C# (.NET) backend, supporting **26 hand configurations** and diverse signing contexts
- Reduced ASL interpretation latency by **40%** by redesigning the gesture-to-phrase mapping pipeline, improving response consistency across varying input conditions
- Accelerated on-device inference **3x** with **70%** model compression via post-training quantization (FP32 to INT8), benchmarking GPU (CUDA) vs. CPU latency profiles with less than **1%** accuracy loss

Crewasis AI (Marketing Intelligence-Tech) — **New York, United States**
ML Engineer Intern — **Jan 2025 - Jun 2025**

- Powered marketing intelligence across **5K+** daily multimodal social media assets by fine-tuning **BLIP-2** with **LoRA** adapters and deploying a **RAG** system over audio, video, and text, containerized with **Docker**
- Scaled batch preprocessing **60x** (30min to 30sec), saving **$19K+ annually**, by deploying Python workers on **AWS Lambda** with **Airflow** triggers and automated data quality checks
- Constructed a search system across **1.6M+** records by integrating REST APIs (YouTube, Instagram, TikTok) with **FAISS** vector retrieval at sub-3s query latency, orchestrated with **Kubernetes** for reliable scaling
- Validated a **29% cost advantage** across **20+ A/B experiments** by evaluating multimodal pipeline variants with **MLflow** tracking and translating results into deployment decisions

Red Moments Pvt Ltd (Manufacturing/E-commerce) — **Mumbai, India**
Jr. Data Scientist — **Jun 2022 - May 2023**

- Improved production planning by **23%** by developing time-series forecasting models (Prophet, XGBoost) on **75K+** transactions with feature engineering in SQL, deployed as a scoring pipeline for business planning
- Generated **$100K annually** with **16%** inventory reduction by designing A/B testing frameworks translating business questions into structured recommendations for senior stakeholders
- Lifted margins by **9%** and produced **$80K** revenue by constructing ETL pipelines with **CI/CD** workflows enforcing schema consistency across all reporting layers
- Slashed reporting from **3 days to real-time**, saving **$30K annually**, by building Power BI dashboards surfacing KPI definitions for cross-functional stakeholders

ACADEMIC PROJECTS

Speech Emotion Recognition System | ***PyTorch, TensorFlow, Keras, Librosa, CNN, LSTM, HuggingFace***

- Achieved **90.5% accuracy** and **90.4 F1-score** across **8 emotion classes** from **15K+** audio samples by designing a **CNN-LSTM** architecture with multimodal audio feature extraction (MFCC, mel-spectrogram, chroma)
- Outperformed InceptionV3 baseline by **3%** while training **25% faster** by evaluating 5 model architectures and selecting optimal depth and feature representation

FinSight RAG: Financial Document Analysis | ***Python, LangChain, RAG, FAISS, ChromaDB, FinBERT***

- Attained **94% query success** and **4.25/5 relevance** across **200 SEC 10-K queries** by building a hybrid **RAG** pipeline with MiniLM embeddings, dense/sparse retrieval, and semantic reranking over **10 S&P 500 filings**
- Cut retrieval latency by **42%** and API costs by **40%** by designing an LLM-as-judge evaluation framework benchmarking 7 retrieval strategies, validated by financial domain experts

EDUCATION & CERTIFICATIONS

Master of Science in Data Analytics Engineering (GPA: 3.94/4.0) — **Sep 2023 - Dec 2025**
Northeastern University | Boston, MA

Bachelor of Technology in Electronics Engineering (GPA: 3.87/4.0) — **Sep 2019 - May 2023**
Mumbai University | Mumbai, India

AWS Certified Machine Learning Engineer - Associate — **Jan 2026 - Jan 2029**