Nitin Jayavarapu

Pensacola, FL | nitinjayavarapu12@gmail.com | +1-448-219-7141
GitHub | LinkedIn

Professional Summary

Artificial Intelligence Engineer with 2+ years of experience building and operating production ML systems, focused on real-time computer vision and low-latency inference. Hands-on experience deploying containerized inference services with FastAPI and Docker on AWS, optimizing model performance, and iterating on data quality to improve reliability in live environments.

Technical Skills

Programming & Core ML: Python, SQL, PyTorch, Scikit-Learn, TensorFlow
ML Systems & Deployment: FastAPI, Docker, MLflow, CI/CD, AWS (EC2, S3, Lambda)
Computer Vision: Object Detection (YOLOv5, YOLOv8), OpenCV
LLM & NLP Systems: Hugging Face, Embeddings, RAG, FAISS, Prompt Engineering
Model Evaluation & Monitoring: Retrieval Precision@k, Grounding & Response Quality Evaluation, Data Drift Detection, Latency Analysis
Data & Storage: Pandas, NumPy, MySQL, SQL Server

Experience

AI Engineer January 2022 – November 2023
Black Box *Bangalore, India*

- Built and productionized a face recognition system used by 10,000+ daily users, working with imperfect camera feeds, inconsistent lighting, and partial occlusions across multiple real-world deployment environments.
- Improved real-time inference performance by iterating on model architectures, batching strategies, and quantization, reducing end-to-end latency by 40% after debugging GPU memory constraints and accuracy regressions introduced during optimization.
- Iterated on data-centric model improvements by analyzing production misclassifications, curating hard-example datasets from failure logs, and retraining models, leading to a 20–30% reduction in recurring recognition errors across the successive releases.
- Deployed and operated containerized inference APIs using FastAPI and Docker on AWS EC2, handling service restarts, logging gaps, and traffic spikes while maintaining 99.9% uptime and balancing latency, accuracy, and infrastructure cost under live usage.

Education

University of West Florida Pensacola, FL
Master of Science in Data Science, GPA: 3.51/4.0 *January 2024 – December 2025*

- Relevant Coursework: Statistical Modeling, Machine Learning, Deep Learning, Cloud Computing

Technical Projects

Real-Time Anomaly Detection & Monitoring System | *Python, Kafka, FastAPI, Scikit-learn, Docker*

- Designed and implemented an end-to-end computer vision inference service, exposing YOLOv8 through FastAPI for image and video-based object detection.
- Profiled system performance at the frame level by separating video decode time from model inference time, enabling clear identification of latency bottlenecks.
- Evaluated system stability under concurrent workloads (5–10 parallel inference requests), measuring tail latency (p50/p95) and verifying zero request failures on local hardware.
- Built a modular, production-style codebase with structured logging, performance analysis scripts, and optional annotated outputs to support debugging and iteration.

Production LLM System (Retrieval-Augmented Generation) | *Python, FastAPI, Transformers, Docker, MLflow*

- Built a RAG-based LLM service over 1k document chunks, delivering grounded responses with end-to-end latency of 1–2 seconds per query, including retrieval, reranking, and generation.
- Implemented document ingestion, chunking, and FAISS vector search, achieving 20–30% improvement in top-k retrieval precision compared to keyword-only baselines on evaluation queries.
- Developed an async FastAPI inference service with schema-validated structured outputs, reducing malformed or hallucinated responses by 35–40% during prompt and retrieval iterations.
- Created an LLM evaluation workflow measuring answer relevance, faithfulness to retrieved context, and quality regression across prompt and model iterations.