Siddharth Prabhu

Bengaluru, Karnataka | +91-8618839179 | siddharthprabhu22@gmail.com
LinkedIn | GitHub | Blog|Portfolio

Professional Summary

Data Science and AI Engineer with 1.6+ years of experience architecting production-grade ML/LLM systems. Specialist in optimizing computer vision pipelines (OpenVINO, DeepStream) and building scalable RAG architectures (Milvus, LangChain). Proven track record of reducing inference latency by 40% and deploying containerized microservices handling high-throughput real-time data streams.

Technical Skills

Core AI & LLM: PyTorch, LangChain, RAG Architecture, Hybrid Retrieval, Prompt Engineering, Gemini/OpenAI APIs
Vision & Edge: OpenVINO, OpenCV, DeepStream, Model Quantization (INT8/FP16), YOLOv8
Databases & Ops: Milvus, MongoDB, FAISS, Docker, Kubernetes (Basic), FastAPI, n8n Automations
Languages & Tools: Python, SQL, Git, NumPy, Pandas

Professional Experience

Nhance.ai Bengaluru, India
Data Science Engineer *Sept 2024 – Present*

- Engineered a multi-stream video analytics system processing 5 concurrent RTSP streams at 30 FPS; migrated inference to OpenVINO, reducing latency from 150ms to 90ms on Intel edge hardware.
- Architected a multimodal RAG service for technical documentation; optimized vector search using Milvus to achieve p95 retrieval latency of <250ms for a 10,000+ page corpus.
- Containerized ML microservices using Docker and implemented instance right-sizing that reduced cloud compute costs by roughly 15%.
- Developed automated ETL scripts using Python and n8n to preprocess large-scale datasets, reducing manual data handling time by 30%.

M.Tech Solutions Bengaluru, India
Information Security Intern *Aug 2024 – Sept 2024*

- Developed Python scripts to analyze 40+ security datasets, automating threat pattern identification and improving risk assessment turnaround time by 25%.
- Assisted in the implementation of Role-Based Access Control (RBAC) protocols to secure sensitive internal data lakes.

Key Projects

Enterprise-Grade RAG Knowledge Base | *Python, LangChain, Milvus, Docker*

- Designed a fallback-enabled RAG architecture switching between OpenAI and Gemini APIs to ensure high availability.
- Implemented Hybrid Search (Dense Vector + Sparse Keyword) to improve retrieval precision by 30% over baseline methods.
- Deployed as a scalable API handling 1.2M+ tokens of ingested technical documentation.

Traffic Surveillance & Anomaly Detection | *Computer Vision, CNN, IEEE Publication*

- Authored: "Image Processing Based Traffic Surveillance" (IEEE ESCI 2024).
- Validated a custom CNN architecture on 5,000+ annotated frames, achieving an F1-score of 92% for vehicle detection.

High-Throughput Document Processing API | *FastAPI, OCR, Celery*

- Built an asynchronous document ingestion service to process 100+ PDFs/hour without blocking the main event loop.
- Implemented multimodal capabilities to summarize PDFs and perform Visual Question Answering (VQA) on images.

Education

Siddaganga Institute of Technology Tumakuru, Karnataka
B.E. in Electronics & Communication Engineering *2020 – 2024*