SIDDHARTH PRABHU

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

PROFESSIONAL SUMMARY

Data Science and AI Engineer with 1.5+ 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 (Docker, FastAPI) handling high-throughput real-time data streams.

TECHNICAL SKILLS

- **Core AI & LLM:** PyTorch, LangChain, RAG Architecture, Hybrid Retrieval, Prompt Engineering, Gemini/OpenAI APIs.
- **Computer Vision & Edge:** OpenVINO, OpenCV, DeepStream, Model Quantization (INT8/FP16), YOLOv8.
- **Vector Databases:** Milvus, MongoDB, FAISS, Pinecone (Semantic Search Vector Embeddings).
- **Deployment & Ops:** Docker, Kubernetes (Basic), FastAPI, CI/CD Pipelines, n8n Automations.
- **Languages & Tools:** Python, SQL, Git, NumPy, Pandas.

PROFESSIONAL EXPERIENCE

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

- **Real-Time Vision Pipeline:** Engineered a multi-stream video analytics system capable of processing 5 concurrent RTSP streams at 30 FPS. Migrated inference from vanilla PyTorch to OpenVINO, achieving a 40% reduction in latency (150ms to 90ms) on Intel edge hardware.
- **Enterprise RAG System:** Architected a multimodal Retrieval-Augmented Generation (RAG) service for querying technical PDFs. Optimized vector search using Milvus, achieving a p95 retrieval latency of ¡250ms for a corpus of 10,000+ pages.
- **Cost Optimization:** Containerized all ML microservices using Docker. Implemented auto-scaling rules and instance right-sizing that reduced cloud compute costs by roughly 15%.
- **Data Pipeline Automation:** Developed automated ETL scripts using Python and n8n to preprocess and clean 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 the identification of threat patterns 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 that switches 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*

- **Published Author:** ”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 under varying lighting conditions.

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

- Built an smart asynchronous document ingestion service using FastAPI and background workers 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, extending support to diverse document formats.

EDUCATION

Siddaganga Institute of Technology Tumakuru, Karnataka
B.E. in Electronics & Communication Engineering *2020 – 2025*