OMKAR SATAPATHY

SENIOR AI/ML ENGINEER

Hyderabad, India | omkarsatapathy001@gmail.com | +91-7609937926 | LinkedIn | Git Hub

AI Engineer with 4+ years of experience in LLM training, fine-tuning, and production-grade deployment. Specialized in building multimodal RAG systems, agentic AI pipelines, and API-driven chatbots for the fashion industry using AWS SageMaker, Lambda, and n8n. Avid follower of frontier AI research — from MoE and MLA architectures to emerging reasoning paradigms — with a knack for rapidly productionizing cutting-edge techniques into scalable, high-efficiency inference solutions that drive business impact.

TECHNICAL SKILLS

LLM Fine-Tuning	AWS SageMaker	Langchain, Strands	LoRA, QLoRA, Finetuning
Multimodal RAG	AWS Lambda, API Gateway	Google ADK, Crew AI	Machine learning, Trainning
PyTorch, TensorFlow	GCP (CloudRun, Firestore)	Python, LLMs/OpenAI GPT	Multi modal RAG
Agentic AI	n8n, Hugging Face	Chatbot Development	Model deployment

PROFESSIONAL EXPERIENCE

Okkular (Independent) - Senior AI/ML Engineer **July 2025 - Present**

- Built Okkular Plus, an AI styling chatbot with RAG and agentic capabilities, enabling celebrity style lookups and personalized outfit recommendations for fashion retailers.
- Developed product description generation pipelines processing millions of fashion images with LLM vision models and brand guardrails for SEO-optimized content across e-commerce platforms.
- Led VLM fine-tuning to distill 70B+ models into efficient 2-3B models for tag generation, targeting 85-92% cost reduction and sub-500ms inference latency.
- Engineered scalable tag generation pipelines on AWS (Lambda, DynamoDB, S3) with image and text-based attribute extraction, improving client satisfaction from 80% toward 95%.

Innovant (Independent) - Data Scientist **Oct 2024 - July 2025**

- Developed pricing models, churn prediction, and customer segmentation using XGBoost and Neural Networks to drive data-driven sales and retention strategies.
- Built an AI-powered sales chatbot with RAG and agentic capabilities using LangChain, automating lead engagement and intelligent content generation.

AKT Global - Data scientist **April 2023 - Oct 2024**

- Built ML models using Scikit-Learn for predictive analytics, feature engineering, and data-driven decision-making with Power BI dashboards.
- Developed and optimized question-answering models on local databases, managing data preprocessing pipelines and query resolution systems.

JSW Steel - Assistant Manager - Analytics **April 2022 - April 2023**

- Performed time-series forecasting using ARIMA and SARIMA models with statistical analysis in Python to optimize pricing strategies and demand prediction for India's leading steel conglomerate.

PROJECTS

- **AetherAI — Perplexity-Style Multi-Agent AI Platform:** Engineered a production-grade multi-agent backend using Google ADK, FastAPI, and Firebase with hierarchical routing to 7+ specialized sub-agents (Email, Research, Shopping, Maps, Code Generation) and a dedicated code mode with aesthetic code rendering using Anthropic Opus.Integrated 5 LLM providers (Gemini, GPT-4o, Claude, Ollama) via LiteLLM with real-time SSE streaming, session-isolated RAG pipelines using LlamaIndex, OAuth Gmail integration, and multi-tier Redis/Firestore caching. Deployed on GCP Cloud Run with Vertex AI.
- Key features: Pay-as-you-go billing (usage-based, not monthly), live user tracking of consumption and limits, real-time feedback score visibility, and per-user cost analytics dashboard. [Live Demo]
- **VLM Fine-Tuning for Fashion Attribute Prediction:** Fine-tuned Qwen2.5-VL-7B using QLoRA (4-bit, LoRA rank=8, α=16, bf16 precision) on 50K fashion images across 2-3 epochs with gradient checkpointing and DDP on H100 GPUs, achieving sub-500ms inference and 85-92% cost reduction over 70B+ API models for production attribute prediction.
- **Multi-Modal RAG Image Retrieval System** — Developed a text-to-image retrieval pipeline using Qwen3-VL vision-language embeddings, ChromaDB vector search, and FastAPI, enabling semantic product image search for fashion e-commerce catalogs.

EDUCATION

Masters in Data Science and Gen AI (DLP) - IIIT Bangalore (3.4/4.0 CGPA) July 2024 - July 2025
MBA in Business Analytics - KIIT University (7.8 CGPA) May 2020 - May 2022
Bachelor of Technology in Mechanical Engineering - SOA University (7.9 CGPA) May 2016 - May 2020

CERTIFICATION

- AWS cloud practitioner essentials course
- Generative Deep Learning with TensorFlow
- Advanced Computer Vision with TensorFlow
- Sequences, Time Series and Prediction