MAHESH GODIKE

+91 8008538015 | maheshgodike17@gmail.com | Hyderabad, India
linkedin.com/in/mahesh-godike-647a12214 | github.com/godikemahesh | maheshgodike.netlify.app

PROFESSIONAL SUMMARY

AI Engineer with 1+ years of experience embedding AI systems into customer workflows and deploying intelligent solutions on NVIDIA Jetson AGX Orin. Specializes in end-to-end deep learning pipelines, from model training and fine-tuning to edge and client-side deployment. Hands-on across Computer Vision (YOLOv8/v9/v10, SegFormer, DeepLabV3), Large Language Models (RAG, LoRA/PEFT fine-tuning, Hugging Face Transformers), Agentic AI (LangChain, Agent Harness, Speckit, Graphify, OpenSpec), inference optimization (TensorRT, ONNX), and IoT/Embedded AI. Delivered measurable results in automotive inspection, enterprise knowledge retrieval, crowd analytics, workshop analytics, and workforce intelligence. Seeking Forward Deployed Engineer roles where applied AI engineering creates measurable business impact at the customer frontline.

PROFESSIONAL EXPERIENCE

AlumnX AI Labs **Aug 17, 2026 – Present**
Forward Deployed Engineer Intern (Full Time) — Hyderabad, India

Valkontek Embedded & IoT Services Pvt. Ltd. **May 2025 – Jun 2026**
AI & Computer Vision Engineer (Full Time) — Hyderabad, India

- Achieved **<50 ms inference latency** and **90%+ defect detection accuracy** in production by engineering a multi-model automated vehicle inspection pipeline (YOLOv8 + DeepLabV3 + SegFormer) covering 7 defect classes (scratches, dents, corrosion, oil leaks, tire faults, seat wear, number plates), deployed on NVIDIA Jetson AGX Orin.
- Delivered **24/7 surveillance across 8 camera feeds** by designing and deploying a Person Re-Identification pipeline (ByteTrack + StrongSORT) with a FastAPI microservice for live data streaming and identity continuity through frame drops and occlusions.
- Improved inference throughput by **~35%** over the baseline PyTorch deployment through quantization and batching optimizations (TensorRT/ONNX) on Jetson AGX Orin.
- Built a gender classification system with **85% accuracy**, including loitering and danger-zone monitoring.
- Promoted from Intern to AI Engineer within 2 months based on end-to-end project delivery.

TECHNICAL SKILLS

Languages: Python, Bash/Shell Scripting
AI / ML / DL: Machine Learning, Deep Learning, Computer Vision, Object Detection, Image Segmentation, Transfer Learning, Model Training & Optimization, Model Evaluation, Data Augmentation
CV Models: YOLOv8/v9/v10, SegFormer, DeepLabV3, UNet, ByteTrack, StrongSORT, DeepSORT, Re-Identification (Re-ID)
LLMs & GenAI: RAG Pipelines, Agentic AI, Prompt Engineering, Fine-tuning (LoRA/PEFT), LangChain Agents, CrewAI, Multi-modal Models, Gemini
Frameworks: PyTorch, OpenCV, LangChain, FastAPI, React, Streamlit, Playwright, Unsloth, Hugging Face Transformers, SentenceTransformers
Inference & Deploy: TensorRT, ONNX, Edge AI Deployment, Inference Optimization, Docker, REST API, Model Deployment
Databases: PostgreSQL, pgvector, ChromaDB, FAISS, Pinecone
Cloud, Data & DevOps: Git & GitHub, Linux, CI/CD, AWS (EC2, Bedrock, SageMaker), NVIDIA Jetson AGX Orin, Raspberry Pi
Data Libraries: NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, Annotation Pipelines, Dataset Engineering, ETL, EDA
FDE / Agentic Tooling: Speckit, Agent Harness, Graphify, OpenSpec

PROJECTS

Automatic Vehicle Inspection System | Edge AI, Computer Vision, PyTorch, YOLOv8, SegFormer, TensorRT, NVIDIA Jetson
- Built a modular detection pipeline (YOLOv8 + DeepLabV3 + SegFormer) with separate inference heads per defect category; optimized with TensorRT and ONNX for Jetson AGX Orin edge deployment.
- Achieved **<50 ms end-to-end latency** and **90%+ mAP** on the validation set across 7 defect categories in a single unified inference pass.

VeriSpec — Published Spec-Driven Quality Engineering Framework | Python, Pytest, Playwright, k6
- Developed and published an open-source framework that transforms software specifications into structured test strategies, traceable test cases, and executable tests for greenfield and existing codebases.
- Implemented workflows for test generation and execution, failure analysis, requirement-to-test traceability, Git change-impact analysis, and targeted regression testing.
- Integrated testing workflows with Pytest, Playwright, and k6, with structured JSON, Markdown, and HTML reporting to support evidence-driven quality engineering.

Signal AI — Email Executive Agent | Python, FastAPI, PostgreSQL, pgvector, React, LLMs, SentenceTransformers

- Designed an AI email intelligence system that prioritizes important messages, identifies actionable commitments, and reduces inbox noise using rule-based filtering and LLM-assisted analysis.
- Architected a tiered processing pipeline combining deterministic rules, heuristic classification, and LLM reasoning for complex emails, with structured signal extraction and contextual search.
- Designed an architecture that links related messages, tracks follow-ups, and enables natural-language queries across email history.

DealerOps Intelligence — AI-Powered Workshop Analytics | Python, FastAPI, React, PostgreSQL, Gemini AI

- Designed an AI-driven workshop intelligence platform consolidating job cards, technician activity, parts transactions, appointments, and service timelines into operational insights.
- Developed the data model and workflow architecture for tracking repair progress, identifying service delays, and analyzing technician productivity, bay occupancy, and workshop efficiency.
- Developed a role-based dealership management interface delivering operational insights to managers, with automated daily, weekly, and monthly reports to support performance monitoring and data-driven decision-making.

WorkTracker — Agentic AI Productivity System | Python, LangChain Agents, Git, Linux, Automation

- Built an agentic productivity monitoring system that processes Linux filesystem events, Git activity, and application usage to generate structured developer activity reports.
- Implemented tool-calling workflows to collect activity signals and automate report generation, reducing reliance on manual status reporting.
- Deployed internally at Valkontek for a team of **10+ remote developers**.

NutriBot — Domain Fine-Tuned LLM | PyTorch, Hugging Face Transformers, LoRA/PEFT, Unsloth

- Fine-tuned the IBM Granite 2B model with LoRA (PEFT) and Unsloth on 52,000 nutrition-domain conversations using Hugging Face Transformers.
- Achieved a **40% improvement** in domain-specific answer accuracy over the base model; deployed with conversational memory for multi-turn dietary coaching.

Enterprise RAG Knowledge System | Python, FastAPI, FAISS, LangChain

- Built a Retrieval-Augmented Generation system that answers questions from uploaded PDF and document collections using FAISS vector search.
- Generated document embeddings, retrieved relevant context, and used an LLM to produce accurate, context-aware responses through a FastAPI backend.

Autonomous Document Generator Agent | Python, FastAPI, LLMs, Agentic AI, python-docx

- Built an autonomous agent that interprets natural-language requests, generates an execution plan, performs multi-step tasks, and produces formatted Word (.docx) documents.
- Implemented planning, task decomposition, sequential execution, reflection/self-validation, and tool orchestration with separate planner, executor, validation, and document-generation modules for proposals, project plans, meeting minutes, SOPs, and technical reports.

EDUCATION

Siddartha Institute of Engineering and Technology **Aug 2025 – Apr 2028**

B.Tech, Artificial Intelligence & Machine Learning — Hyderabad, India. Admitted via ECET lateral entry; studying via correspondence alongside full-time industry work.

Government Polytechnic College, Masab Tank **Jun 2022 – May 2025**

Diploma, Electronics & Communication Engineering (ECE) — Hyderabad, India

CERTIFICATIONS & TRAINING

- Python Complete Course — Udemy (2023)
- AI/ML Internship Program — Pregrad (2023)
- Cloud Computing Internship — IBM Skills, Virtual (2023)
- NCC B & C Certificate — Government of India (2022–2025)

ACHIEVEMENTS

- Promoted from Intern to AI Engineer at Valkontek within 2 months based on end-to-end project delivery.
- One of few early-career engineers in India with hands-on production deployment on NVIDIA Jetson AGX Orin for automotive-grade AI workloads.
- Delivered 8 end-to-end projects spanning Computer Vision, LLM Fine-tuning, RAG, Agentic AI, Edge AI, Enterprise Analytics, and Quality Engineering, solo and in teams.