

Amogh Dongre

+91-96636 88182 | amoghdongre23@gmail.com | github.com/amogh-dongre | linkedin.com/in/amogh-dongre

EDUCATION

VIT-AP University

B.Tech in Computer Science (AI & ML Specialization) – CGPA: 9.12

Amaravathi, India

Jun 2026

- Relevant Coursework: Deep Learning, NLP, Computer Vision, Probabilistic Modeling, Generative Models

EXPERIENCE

Independent Consultant

ML Infrastructure Engineer

Remote

Jul 2026 – Present

- Deployed **vLLM on Kubernetes** for an enterprise running LLMs entirely on-premise, implementing GPU-aware scheduling and **zero-downtime model rollouts** that preserve in-flight requests.
- Instrumented the serving stack with Prometheus, Grafana and Langfuse for per-request tracing, making latency regressions attributable to a cause.

GalaxEye Space

Systems Software Engineer Intern

India

Jan 2026 – Jun 2026

- Accelerated PulseRiver E2E real-time SAR image-processing performance by **12x**, reducing a 4 GB dataset benchmark from ~75s to ~6s using GPU parallel processing, multithreading, and data-pipeline parallelism.
- Integrated RocksDB-backed storage and optimized stage-level data flow to preserve the ~6s runtime while extending the pipeline with Global Back Projection, DEM support, and a GBP kernel.
- Cut Chitra GitHub Actions CI runtime by **84%**, from ~19 min to ~3 min, by restructuring workflow execution and optimizing multi-layer Docker builds; added CI for PulseRiver with 20 static tests as an automated quality gate.
- Slimmed production Docker images by **67%**, from ~1.5 GB to ~500 MB, using layered builds and dependency cleanup across frontend/backend services.
- Migrated the production database from a Docker Compose-managed container to **AWS RDS** and supported EC2 provisioning/autoscaling, increasing concurrent-user scale **7.5x** (~20 to ~150); added Alembic migrations to make schema changes safe across environments.
- Built the first production observability layer using the **LGTM stack**, tracing logs across 8 services, monitoring 20+ metrics, and surfacing Grafana dashboards for service health, logs, and traces.

AWIROS

Computer Vision Intern

Bengaluru, India

Jun 2025 – Jul 2025

- Benchmarked Vision-Language Models (VLMs) for visual reasoning and query-based scene retrieval, and built vision-language alignment pipelines integrating textual queries with pixel-level understanding for automated annotation.
- Optimized OCR and crowd-estimation models using self-training and pseudo-labeling on unlabeled datasets, and deployed them to resource-constrained Jetson-class hardware with TensorRT.

AWIROS

Computer Vision Intern

Gurgaon, India

Jun 2024 – Jul 2024

- Trained and fine-tuned detection and tracking models for real-time perception, and built detection and alerting modules for live video streams under a sub-100ms latency budget.

PROJECTS

Hardware Certification Assistant | FastAPI, LlamaIndex, Ollama, PostgreSQL/pgvector

2026

- Built a retrieval system over the Indian hardware compliance corpus (BIS, WPC, TEC, MeitY), attaching the source clause to every answer so applicability claims can be verified.

Heterogeneous GPU Inference Cluster | vLLM, Ray, TensorRT-LLM, 200GbE RDMA

2026

- Built and benchmarked a 5-node arm64/x86 cluster (3x DGX Spark, 2x RTX PRO 5000 Blackwell) on a 200GbE ConnectX-7 fabric, measuring where tensor parallelism stops outperforming pipeline parallelism by model size.

Synthetic Data Generation Pipeline | PyTorch, Diffusion Models

2025

- Trained Latent Diffusion Models with CLIP-guided conditioning and contrastive objectives for high-resolution synthetic object-detection data, achieving a **15%** improvement in downstream detection on limited real data.

Flight Tracker | ADS-B, FastAPI, Socket.IO, PostGIS, AWS

Dec 2024 – Feb 2025

- Built a live flight-tracking system on raw ADS-B signals with custom real-time processing, serving **500+** concurrent users at >95% data accuracy and sub-second update latency.

TECHNICAL SKILLS

Languages: Python, C++, Java, SQL, Rust (learning); FastAPI, REST APIs, Alembic, Testcontainers, RocksDB

Cloud & DevOps: AWS (EC2, RDS, S3, SQS, Autoscaling), Docker, Kubernetes, GitHub Actions, CI/CD, Linux

Inference & Observability: vLLM, Ray, TensorRT-LLM, TensorRT, ONNX Runtime, llama.cpp, LlamaIndex, pgvector; Grafana, Loki, Tempo, Prometheus, Langfuse

ML/CV & Systems: PyTorch, PyTorch Lightning, CUDA, VLMs, Diffusion Models, SAR image processing, DEM, Global Back Projection