Deepak Inugala

LLM & MLOps Engineer | AI Platform Engineering | GPU Inference Infrastructure | Agentic AI | International Delivery

Email: deepak.1990@hotmail.com
Phone: +971 50 494 5921
Location: Abu Dhabi, UAE
LinkedIn: linkedin.com/in/deepak-inugala

LLM Serving	End-to-End	Agentic AI	Air-Gapped	6 Countries	7B – 72B
vLLM / Triton / LiteLLM	MLOps Pipelines	RAG / LangGraph / MCP	Sovereign LLM Ops	International Delivery	Model Scale Deployed

PROFESSIONAL SUMMARY

LLM & MLOps Engineer with 10+ years of infrastructure and platform engineering experience, specializing in the complete lifecycle of Large Language Model and AI/ML workloads in production — from model sourcing, quantization, and fine-tuning infrastructure through production serving, MLOps pipelines, agentic AI systems, and operational monitoring. Deep hands-on expertise deploying vLLM, NVIDIA Triton, and LiteLLM on Kubernetes with NVIDIA H100/H200 GPU clusters; building end-to-end MLOps pipelines with Kubeflow, Argo Workflows, MLflow, and Seldon Core; and deploying autonomous AI agent platforms (OpenClaw, NemoClaw, LangGraph, n8n). Proven experience building sovereign and air-gapped LLM infrastructure across GOVINT, OSINT, and Smart Nation verticals at G42 with international delivery across 6 countries.

TECHNICAL SKILLS

LLM Serving & Inference	vLLM (PagedAttention, continuous batching, tensor parallelism, prefix caching, AWQ/GPTQ), NVIDIA Triton Inference Server (TensorRT, ONNX, dynamic batching, ensemble), LiteLLM (OpenAI-compatible gateway, load balancing, fallbacks, cost tracking), Ollama, Ray Serve, BentoML
LLM Frameworks & Tools	OpenAI-compatible APIs, HuggingFace Transformers & Hub (offline mode, snapshot_download), ONNX Runtime (CUDAExecutionProvider), LangChain, LangGraph, AutoGen, Open WebUI, Mattermost LLM integration
Quantization & Opt.	AWQ 4-bit (AutoAWQ), GPTQ 4/8-bit, GGUF (llama.cpp), FP8 inference; tensor & pipeline parallelism; TensorRT model compilation; model benchmarking (MMLU, HumanEval, MT-Bench, perplexity)
Agentic AI & RAG	LangGraph (stateful agents, supervisor topology), n8n workflow automation, OpenClaw, NemoClaw; Qdrant vector store (StatefulSet, semantic search), Weaviate; MCP (Model Context Protocol) servers; tool-use pipelines; human-in-the-loop gates; audit logging
MLOps Toolchain	MLflow (experiment tracking, model registry, stage promotion, MinIO artifact store), Kubeflow Pipelines (GPU-scheduled pipeline SDK), Argo Workflows (DAG orchestration), DVC (data versioning, pipeline reproducibility), Weights & Biases, Seldon Core (canary / shadow / A/B deployments)
GPU & HPC	NVIDIA H100, H200, A100; MIG Configuration (GPU Operator MIG Manager); DCGM Exporter; KubeRay (RayCluster); CUDA/cuDNN stack management; InfiniBand / NVLink; NCCL tuning
MLOps Data & Features	Feature Stores (Feast — offline/online store, Redis online serving); MinIO / S3 artifact storage; DVC data versioning; HuggingFace Datasets; model lineage & governance (MLflow Registry + Kyverno admission policy)
Kubernetes & Cloud	AKS, EKS, RKE2, Rancher; Helm, ArgoCD (Argo Rollouts + Prometheus analysis gates), Kustomize; Docker, Containerd, Harbor; Azure (AKS, Key Vault, Defender), AWS (EKS, S3), G42 Cloud
Air-Gapped LLM Ops	HuggingFace Hub mirroring (HF_HUB_OFFLINE=1, TRANSFORMERS_OFFLINE=1, snapshot_download); Harbor registry mirrors; offline model weight packaging; AWQ quantization for size reduction; internal MinIO model registry
Observability & Sec	Prometheus (vLLM metrics, DCGM, custom recording rules), Grafana (30+ dashboards, golden signals), Loki + Fluent Bit; Falco, Trivy, OPA Gatekeeper, Kyverno; Keycloak OIDC; Zero-Trust NetworkPolicies
IaC & Scripting	Terraform, Ansible, GitLab CI/CD, GitOps, Python (HuggingFace SDK, MLflow SDK, Kubeflow SDK, LangChain/LangGraph agent development), Bash

WORK EXPERIENCE

LLM & MLOps Engineer | Senior SRE | AI Platform | Group 42 (G42)

07/2019 – Present | Abu Dhabi, UAE | On-Site: Kazakhstan, Angola, Bahrain | Remote: Maldives, Ethiopia

International AI Platform Delivery — On-Site & Remote (6 Countries)

- Deployed and commissioned G42 AI and HPC platforms at client sites across Kazakhstan, Angola, and Bahrain — sole technical authority for GPU infrastructure, networking, storage, and security; additionally supported Maldives and Ethiopia remotely via secure VPN. Managed all pre-deployment readiness: data center vendor coordination, GPU server rack-and-stack, high-speed networking commissioning, storage validation, and client UAT — travelling approximately 50% of each month; led knowledge-transfer workshops and formally handed over fully operationalized systems with runbooks and SLA documentation.
- Delivered HPC and AI infrastructure concurrently across GOVINT (classified air-gapped GPU clusters for government intelligence), OSINT (GPU-accelerated NLP inference and large-scale data processing), and Smart Nation projects (real-time AI inference and city-scale analytics pipelines).

LLM Serving Infrastructure — Production at Scale

- Built and operated production vLLM serving infrastructure on NVIDIA H100/H200 GPU clusters — deployed LLMs from 7B to 72B parameters with tensor parallelism, AWQ 4-bit quantization (72B: 144GB → 36GB), continuous batching, and prefix caching; achieved P99 TTFT under 200ms at 1,000+ concurrent requests; configured Kubernetes HPA on DCGM GPU utilization metrics for auto-scaling, and liveness/readiness probes on the /health endpoint.
- Deployed LiteLLM as an OpenAI-compatible API gateway in front of multiple vLLM instances — configured load balancing, model-level rate limiting, API key management, cost tracking per team, and automatic fallback routing; enabled Open WebUI, n8n, LangChain agents, and Mattermost bot to use identical API calls regardless of backend model.
- Deployed NVIDIA Triton Inference Server for embedding and encoder models — ONNX Runtime backend with dynamic batching, TensorRT-optimized models for 2-3x speedup, and ensemble pipelines for pre/post-processing; exposed gRPC endpoints for sub-50ms RAG retrieval in the SRE knowledge-base pipeline. Managed AWQ and GPTQ quantization pipelines with quality validation against MMLU and HumanEval benchmarks before production promotion.

MLOps Pipeline Design & Implementation

- Architected end-to-end MLOps pipelines using Kubeflow Pipelines and Argo Workflows — defined GPU-scheduled component DAGs for data ingestion, fine-tuning (KubeRay distributed training), evaluation (accuracy, perplexity, latency P99 benchmark gates), MLflow registration, and ArgoCD-triggered vLLM rollout; reduced model release cycles from weeks to hours.
- Established MLflow as the organizational model lifecycle platform — set up MLflow Tracking Server (PostgreSQL backend, MinIO artifact store), defined experiment conventions across teams, implemented stage promotion workflow (None → Staging → Production → Archived) with approval gates; deployed DVC for training data versioning with MinIO remote eliminating the 'which dataset?' audit gap; implemented Kyverno admission policy rejecting vLLM Deployments referencing unregistered model paths.
- Implemented canary and blue-green model rollout strategies via Argo Rollouts and Seldon Core — defined Prometheus AnalysisRuns (error rate, P99 latency) as automated rollout gates; deployed Seldon shadow mode to test new model versions against 100% of production traffic with zero user impact before any traffic shifting.

Agentic AI Systems & RAG Pipelines

- Deployed OpenClaw and NemoClaw autonomous AI agent platforms on Kubernetes — connected to internal vLLM endpoints via NetworkPolicy-restricted Services; built production RAG pipelines using Qdrant (Kubernetes StatefulSet, Longhorn PVCs) indexing SRE runbooks and incident post-mortems, achieving semantic search with under 50ms retrieval latency at production query volumes.
- Designed LangGraph-based stateful multi-agent SRE automation workflows — supervisor agent routes Prometheus alerts to specialized sub-agents: retrieval agent (Qdrant RAG), execution agent (kubectl / Ansible via RBAC-scoped ServiceAccount), and validation agent (Prometheus query verification); human-in-the-loop approval gates for destructive actions with full audit logging; reduced overnight on-call interventions by 40%.
- Implemented MCP (Model Context Protocol) servers as Kubernetes Deployments exposing internal APIs (Kubernetes API, GitLab, Prometheus, Jira) as structured tool endpoints for LLM agents; built n8n workflow automation (StatefulSet, OAuth2-proxy sidecar) integrating LLM agents with CI/CD, PagerDuty, and Mattermost for automated incident triage and runbook suggestion — reducing operational toil by 50%.

Air-Gapped LLM Ops & Observability

- Designed and operated complete air-gapped LLM deployment pipeline for sovereign client environments — downloaded, quantized, and packaged models on internet-connected bastion with HF_HUB_OFFLINE=1 and all tokenizer/config files included; transferred through secure data diode; loaded into air-gapped Harbor registry (nvcr.io, docker.io, ghcr.io mirrors); deployed on isolated Kubernetes clusters with internal MinIO model registry for versioned artifact management.
- Built comprehensive LLM observability stack — vLLM Prometheus metrics (e2e_request_latency_seconds, num_requests_running, gpu_cache_usage_perc, generation_tokens_total) combined with DCGM GPU metrics into

30+ Grafana dashboards; configured Alertmanager rules for P99 TTFT degradation, KV-cache exhaustion, and GPU hang detection; implemented distributed tracing for multi-agent workflows.

Cloud Engineer | First Abu Dhabi Bank (FAB)

02/2018 – 07/2019 | Abu Dhabi, UAE

- Cloud Infrastructure & Big Data Administration: completed FGB-NBAD bank merger IT integration with zero major incidents; designed and managed AWS architectures (EC2, EKS, RDS, S3, Lambda); installed and administered Cloudera CDH clusters (HDFS NameNode HA, YARN, Hive, HBase, Spark) with Kerberos authentication and Ranger authorization; deployed multi-node Elasticsearch clusters with ILM policies and Kibana dashboards.
- Vulnerability Management & Monitoring: owned full patching lifecycle — CVE triage, Ansible-based OS patching, post-patch scan validation; implemented centralized monitoring via CloudWatch, Prometheus, and Grafana; led incident response reducing repeat incidents by 35%.

Linux Engineer | HCL InfoSystems Limited

04/2015 – 01/2018 | Abu Dhabi, UAE

- Large-Scale Linux & Data Center Operations: managed 3,000+ physical Linux servers (RHEL, CentOS) — OS provisioning, Ansible configuration management, hardware fault diagnosis (iDRAC/iLO), and firmware lifecycle management across Dell, HP, and IBM server platforms in enterprise data centers.
- Network, Firewall & Migration: configured Cisco/IBM switches (VLAN management, STP/RSTP, LACP); administered firewall rule sets, DMZ segmentation, and NAT rules; led on-premises-to-cloud migration of Linux infrastructure and Cloudera Hadoop clusters with under 2 hours downtime and 0% data loss.

CERTIFICATIONS

Microsoft Certified: Azure Administrator Associate (AZ-104)
Valid: 02/2026 - 07/2027

Microsoft Certified: Azure Solutions Architect Expert (AZ-305)
Valid: 07/2025 - 07/2027

AWS Certified Solutions Architect - Associate
Valid: 07/2022 - 07/2025 | ID: PFR7HTQ25EFQQS9T

Certified Kubernetes Administrator (CKA)
Valid: 09/2021 - 09/2023 | Renewal in Progress

G42 Cloud Certified Engineer
Valid: 10/2021 - 12/2024 | ID: G42C/SVD/CRT/0475

Red Hat Certified Engineer (RHCE)
Valid: 02/2015 - 02/2018 | ID: 150-012-904

EDUCATION

Master of Business Administration (MBA)
Jawaharlal Nehru Technology University | 2012 - 2015

Bachelor of Technology (B.Tech)
Jawaharlal Nehru Technology University | 2008 - 2012

LANGUAGES & SOFT SKILLS

Languages
English & Hindi (Professional) | Telugu (Native) | German (Elementary)

Soft Skills
Cross-functional Collaboration | Stakeholder Communication | Problem Solving | Rapid Learning | Adaptability