ANOMNY

Sovereign AI Infrastructure Platform

Data Residency · GPU Compute · Secure by Design · Air-Gapped Ready

AI/ML Training & Inference | LLM Hosting & RAG | MLOps & Model Governance

ANOMNY Sovereign AI Infrastructure Platform enables enterprises, government departments and regulated industries to build, train, fine-tune and deploy AI/ML workloads entirely within national or organisational boundaries. It brings GPU-accelerated compute, secure data governance and full-stack MLOps tooling into the data centre — giving customers complete ownership of their data, models and infrastructure, without dependence on foreign hyperscale cloud providers.

100%	GPU Compute	Air-Gapped	CERT-IN	Made in
Data Residency	Accelerated Fabric	Deployment Ready	Compliant	India

A Sovereign Compute Infrastructure

GPU Cluster Orchestration

- High-density GPU clusters (NVIDIA H100/A100-class) with dynamic workload scheduling
- Multi-tenant GPU partitioning (MIG) for isolated workloads
- Auto-scaling compute pools for training and inference workloads
- Liquid and rear-door cooling support for high-TDP AI racks

Storage & Networking

- High-throughput NVMe storage tiers for large training datasets
- RDMA / InfiniBand and 400G Ethernet fabric for distributed training
- Dedicated storage and network isolation per tenant / workload
- Tiered hot/warm/cold storage for datasets, checkpoints and model artefacts

B Data Sovereignty & Governance

Data Residency Controls

- All data, models and inference logs remain within the designated national / organisational boundary
- Zero cross-border data transfer by architecture, not policy alone
- Configurable data residency zones for multi-site deployments

Governance & Ownership

- Full customer ownership of trained models and weights - no vendor lock-in
- Data classification and access-control policies enforced at ingestion
- Immutable audit trail of data access, training runs and model versions

C AI / ML Platform Layer

Model Development & MLOps

- End-to-end MLOps pipeline: data prep, training, fine-tuning, evaluation, deployment
- Support for open-source and custom LLMs, with on-prem fine-tuning
- Model registry with versioning, rollback and approval workflow

Inference & Serving

- Low-latency inference serving with auto-scaling endpoints
- Retrieval-Augmented Generation (RAG) pipeline with private vector database
- API gateway for internal application integration (REST / gRPC)

D Security & Compliance

Infrastructure Security

- Air-gapped and hybrid (connected) deployment modes

Regulatory Alignment

- Aligned with national data protection and localisation mandates

- AES-256 encryption at rest and in transit for all AI data and models
- Hardware-rooted trust (TPM / secure boot) on all compute nodes
- CERT-IN compliant, cyber-hardened stack
- Role-based access control (RBAC) with SSO (AD / LDAP / OAuth2)

E Deployment Architecture & Operations

Flexible Deployment

- On-premises, private cloud, or hybrid deployment within customer / national DC
- Modular scale-out design - start with a single GPU rack, scale to multi-site clusters
- Integrates with existing DCIM, BMS and VMS for unified facility + compute monitoring

Operations & Monitoring

- Centralised control room dashboard for compute, power and thermal telemetry
- Predictive maintenance and capacity planning via DCIM-linked analytics
- 24x7 NOC-ready monitoring, alerting and escalation

Why ANOMNY Sovereign AI?

- Complete data and model sovereignty - nothing leaves your jurisdiction
- No dependency on foreign hyperscalers for critical AI workloads
- Full-stack platform: compute, storage, MLOps and governance in one solution
- Open architecture - no vendor lock-in, integrates with existing data centre infrastructure
- Made in India - local R&D, support and on-site deployment
- Government and enterprise-grade security, CERT-IN compliant

Get in Touch

ANOMNY Technologies Pvt. Ltd. | www.anomny.com
For Sovereign AI demo, PoC and deployment enquiries: info@anomny.com

© 2025 ANOMNY Technologies Pvt. Ltd. | www.anomny.com | All Rights Reserved