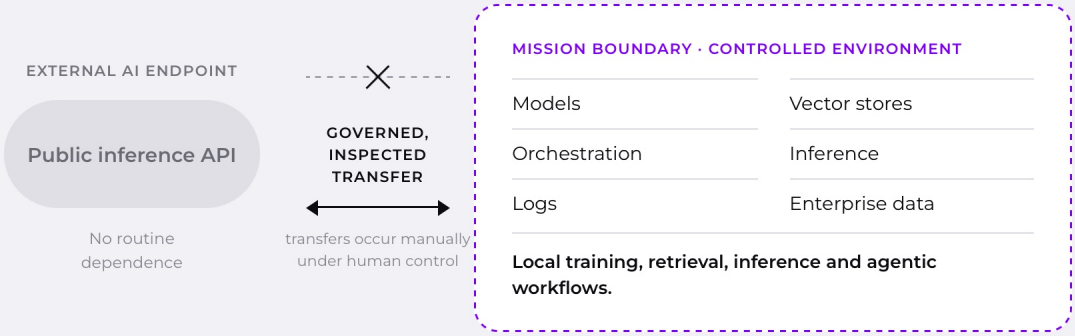
Air-gapped AI



BRING THE AI STACK TO THE DATA, NOT THE DATA TO THE AI STACK

Mission intelligence stays inside the mission boundary.

For defense, intelligence, financial services, healthcare and other high-consequence environments, air-gapped AI enables local training, retrieval, inference and agentic workflows without routine dependence on an external AI endpoint.

What "air-gapped" should mean.

Physical + logical separation

NIST defines an air gap as an interface where systems are not physically connected and any logical connection is not automated; transfers occur manually under human control.[1]

Local AI operations

Models, vector stores, orchestration, inference, logs and enterprise data remain within the controlled environment. Updates enter through governed, inspected transfer procedures, not an always-on service dependency.

Air gap is an architecture, not a magic shield.

Secure transfer stations, removable-media controls, signed artifacts, privileged-access management, patch governance, logging, backup and red-team testing remain essential.

Why high-security operators choose it.

No routine data egress

Prompts, source records, embeddings, model outputs and operational telemetry do not need to traverse a third-party inference API.

Reduced remote attack surface

Removing internet-facing inference paths eliminates a class of remote access and exfiltration routes. Physical media and insider risk still require controls.

Mission continuity

Local AI can continue in disconnected, intermittent and bandwidth-limited environments, supporting tactical and remote operations.

Sovereign control

The operator controls model versions, data location, retention, access policy, audit evidence and refresh timing.

Low-latency decisions

Inference occurs near sensors, analysts, clinicians or transaction systems, avoiding cloud round trips.

Predictable governance

A bounded system makes data-flow mapping, authorization boundaries and evidence collection easier to define and test.

Sector-specific value.

Defense & intelligence

Run C5ISR, ISR exploitation, maintenance analytics, mission planning and analyst copilots in classified or DDIL environments. Hub-to-edge packaging supports centralized governance with disconnected execution.

Financial services

Keep customer, trading, fraud, credit and investigation data inside the institution's controlled boundary. Local models avoid sending privileged records and prompts to a public inference endpoint.

Healthcare & life sciences

Process PHI, clinical notes, imaging metadata and research data locally. Air-gapping can reduce third-party data-flow exposure, but HIPAA compliance still depends on the full administrative, physical and technical safeguard program.

Critical & secretive industries

Support energy, manufacturing, legal, R&D and sovereign-government workloads where intellectual property, operational telemetry or national-security data cannot leave the site.

Zeblok security functionality.

Identity, secrets & cryptography

- Role-based access for developer, operations and product roles
- JWT authentication, bcrypt password hashing and two-factor authentication
- Break-glass escalation with audit logging
- LUKS full-disk encryption; AES-256 / TLS 1.2+ listed for protected data paths
- Vault-based dynamic secrets and automated key rotation

Workload & supply-chain controls

- Deny-all network policies by default
- Namespace and pod-level isolation for containerized workloads
- Docker image scanning and signed-model controls
- CORS, CSRF and DDoS rate-limit protections
- Kubernetes hardening aligned in Zeblok collateral to NSA/CISA guidance and CIS Benchmarks

Security that also changes the economics.

Own the operational boundary. Reduce recurring exposure to token, transfer and overprovisioning costs.

–60%

GPUs in Zeblok's published 70B workload scenario

2.5×

published capacity per GPU in the same scenario

~$6.5M

published five-year modeled savings

Where Zeblok can lower end-user cost of operation.

Local ownership

Replaces per-token external API charges with capacity the operator owns and schedules. Economics shift toward utilization, power, support and refresh planning.

Intelligent context offload

Zeblok states that context can be tiered from GPU memory to system memory and high-performance storage, allowing larger context and more concurrent work without scaling GPU count linearly.

Right-sized, portable stack

Pre-integrated software and hardware options reduce bespoke stack assembly. Hardware-agnostic deployment and model choice are positioned to reduce lock-in and improve price/performance selection.

A practical TCO decision model.

Compare five-year cost

Cloud/API: tokens + retrieval + data transfer + reserved capacity + vendor support + compliance overhead.

Private/air-gapped: hardware + software + power/cooling + facilities + operations + secure update process + refresh.

Measure with your workload

Benchmark model size, quantization, context length, concurrency, latency target, duty cycle and data volume. Include redundancy and mission-availability requirements. Validate every saving claim in a proof-of-value using representative workloads.

THE OUTCOME

Private AI at the point of decision, with a smaller external data path and economics tied to owned capacity rather than every prompt.

Sources & claim notes: [1] NIST CSRC, "air gap" glossary (CNSSI 4009-2022 / RFC 4949). Zeblok security features are based on 2026 internal product collateral and should be validated against the contracted release and deployment design.