

>_ TECHNICAL WHITEPAPER

Sovereign AI Security Whitepaper

Security architecture, controls and verification for air-gapped and on-premises enterprise AI.

VERSION	1.2
DATE	24 September 2026
CLASSIFICATION	Public
PUBLISHER	reinvented.digital
WEB	https://reinvented.digital

CONTENTS

What is in this paper

01	Executive summary
02	Audience, scope and status of this document
03	Threat model
04	Reference architecture
05	Control domains
06	Verification: evidence, not assertion
07	Governance and regulatory alignment
08	How we handle your information during an engagement
A	Appendix A — Questions to put to any AI vendor
B	Appendix B — Security posture of this website
C	Appendix C — Limitations and revision history

01

Executive summary

Generative AI is now useful enough that regulated organisations want it inside their most sensitive workflows. Sending prompts, documents and embeddings to a multi-tenant cloud API creates a third-party data path, a vendor dependency and gaps in auditability that many organisations cannot accept.

Running open-weights models inside a perimeter the organisation controls removes that third-party path. It does not, by itself, make the system secure. An AI deployment introduces new attack surface, including prompt injection, poisoned retrieval data, tampered model files and over-privileged agents, and it must be designed, built and verified as a security system rather than as a GPU cluster with a chat interface.

This paper describes the approach reinvented.digital takes to designing and verifying such deployments. In short:

- **Deny first, then verify.** No data path leaves the client-controlled perimeter by default. Egress is blocked at every boundary and then proven blocked by test.
- **Controlled intake.** Model weights, container images and dependencies enter only through a one-way, human-approved transfer process that verifies integrity and provenance.
- **Access enforced where the data is.** Entitlements are checked at retrieval time, per user and per document, so the AI cannot return what the user could not open in the source system.

- **Least privilege for agents.** Autonomous workflows run with scoped identities, allow-listed tools and human approval for consequential actions.
- **Evidence by design.** Material events are logged locally in tamper-evident form to support incident response and regulatory examination.
- **Honest claims.** Where we refer to frameworks such as ISO/IEC 27001 or the NIST AI RMF, we mean design references, not certifications.

02

Audience, scope and status of this document

Audience. CISOs, CIOs, CTOs, risk and compliance leads, enterprise architects and board advisers evaluating private AI deployments.

Scope. Deployments of large language models on client-controlled infrastructure: on-premises bare metal or an isolated cloud network with no internet gateway. This includes retrieval-augmented generation (RAG), fine-tuning and agentic workflows. It does not cover consumer AI tools or a client's use of public AI services.

Status. This paper describes our design approach and default control set. The controls that apply to a specific engagement are defined in the engagement documents and depend on the client's environment, data classification and risk appetite.

IMPORTANT

This document is not a certification, audit report or attestation of compliance with any standard or law, and it is not legal advice. References to ISO/IEC 27001, ISO/IEC 42001, SOC 2, the NIST AI RMF or the EU AI Act indicate frameworks we use as design references and to structure our work. Whether a given system meets a given obligation is a determination for the client, its auditors and its legal counsel.

Threat model

We threat-model each deployment before design begins. The table lists the categories we treat as standard. It draws on the OWASP Top 10 for LLM Applications and MITRE ATLAS, and is extended per engagement.

THREAT	WHAT IT LOOKS LIKE	PRINCIPAL CONTROLS
Data exfiltration to third parties	Prompts, documents or embeddings sent to external APIs; hidden software telemetry; crash and update reporting.	Air-gapped or isolated design; default-deny egress; no external SaaS in the inference path; telemetry disabled and verified by capture.
Prompt injection (direct and indirect)	Malicious instructions hidden in a document, email or web page steer the model or its tools.	All model input treated as untrusted; least-privilege tools; output validation; human approval for consequential actions.
Sensitive information disclosure	The model or retrieval layer returns content the requesting user is not entitled to see.	Retrieval-time access control per document; collection isolation; output filtering as a secondary control; audit logging.
Model and software supply chain	Tampered weights, code-executing serialised files, compromised dependencies or images.	Approved-source policy; hash verification; safe weight formats; pinned, mirrored, scanned and signed artefacts; SBOMs.
Excessive agency	An agent with broad credentials takes a damaging or irreversible action.	Scoped service identities; allow-listed tools; approval gates; step and spend limits; kill switch.
Insider misuse and privilege abuse	An administrator reads confidential prompts or outputs; access accumulates beyond need.	Role-based access; separation of duties; privileged-access logging; time-boxed elevation; documented break-glass.
Data and model poisoning	Corrupted fine-tuning or index data degrades or steers the system.	Provenance for training and index data; reviewed ingestion pipelines; evaluation gates before promotion.
Availability and resource exhaustion	Runaway or hostile requests saturate GPU capacity.	Quotas and rate limits; queue isolation; capacity headroom; monitoring and alerting.
Infrastructure compromise	Exploited firmware, management controller, hypervisor or driver vulnerability.	Segmented management network; hardened baseline; patching via the transfer process; hardware inventory.

Reference architecture

Our reference design separates the system into zones with explicit, allow-listed flows between them. Nothing inside the perimeter has a route to the internet. The single way material enters is the transfer station.

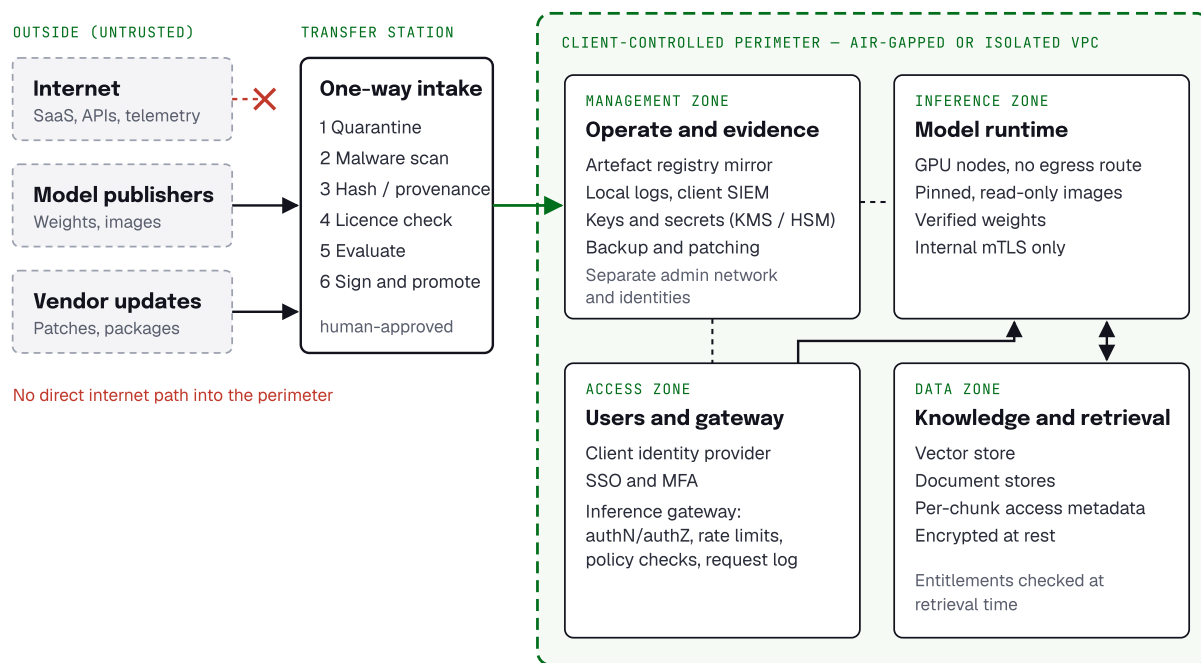


Figure 1. Reference zones. Solid arrows are allow-listed data flows; dashed lines are management and logging flows. The transfer station is the only way material enters the perimeter.

Access zone

Users authenticate through the client's identity provider with multi-factor authentication. An inference gateway performs authorisation, rate limiting, policy checks and request logging before any prompt reaches a model.

Inference zone

Model runtime on GPU nodes with no outbound route. Images are pinned and read-only; weights are verified at load; internal traffic uses mutually authenticated TLS.

Data zone

Vector and document stores. Access-control metadata is attached to each chunk at ingestion and enforced at query time. Data is encrypted at rest with client-managed keys.

Management zone

Registry mirror, logging, key management, backup and patching on a separate administrative network with separate identities.

ISOLATED VPC VARIANT

Where a physical air gap is not required, the same zones are built in an isolated cloud network with no internet gateway, private endpoints only and explicit egress-deny policy. The residual risk in that variant is the cloud provider's control plane and staff. We assess it explicitly and do not describe it as equivalent to an air gap.

Control domains

The following domains make up our default control set. Each is tailored to the client's environment and risk appetite.

5.1 Network isolation and egress control

- Default-deny at every zone boundary. Inference and data zones have no default route to the internet.
- Internal-only DNS and time services, with no external forwarders.
- Software telemetry, update checks and crash reporting are disabled in every component, and we *verify* this by traffic capture rather than trusting configuration.
- Management interfaces (BMC/IPMI and similar) sit on an isolated management network with unused interfaces disabled.
- Continuous egress monitoring at the perimeter, with alerts on any attempted external connection.

5.2 Model and software supply chain

- Artefacts come only from an approved-sources list and enter through the transfer station: quarantine, malware scan, hash and provenance verification, licence check and evaluation.
- We prefer weight formats that cannot execute code when loaded (for example safetensors or GGUF). Pickle-based formats are avoided or converted in a sandbox.
- Dependencies and container images are pinned, mirrored internally, scanned, and accompanied by a software bill of materials (SBOM). Images are signed, and only signed images are admitted.
- The hash of every model file is recorded at import, verified at load and periodically re-verified.
- Fine-tuned weights and adapters are promoted only after evaluation gates (safety, regression and leakage tests) and remain the client's property.

5.3 Identity, access and secrets

- Integration with the client's identity provider (SSO, MFA) and role-based access with defined roles, for example user, power user, model operator, platform administrator and auditor.
- Separation of duties: platform administrators are not able to silently read business content.
- Service-to-service authentication by mTLS and short-lived credentials. Secrets live in the client's vault or HSM and never in images, prompts or source control.
- Privileged access is time-boxed and logged. A documented break-glass procedure covers emergencies.

5.4 Data protection and retrieval security

- Data is classified before it is indexed, and the client decides what belongs in the knowledge base.
- Access-control metadata is stored per chunk and the caller's entitlements are applied at retrieval time. A user cannot retrieve what they could not open in the source system.
- Encryption in transit (TLS, mTLS) and at rest (volume and database), with keys managed by the client.

- Retention of prompts and outputs is configurable and defaults to the minimum that audit requires. Redaction of personal data in logs is available.
- Backups are encrypted, and restores are tested.

5.5 Agentic workflows

- Retrieved text and tool output are treated as untrusted input that can carry injected instructions.
- Agents act only through allow-listed tools with defined schemas and validated parameters.
- Each agent has its own least-privilege identity, read-only by default. Write actions are bounded or require approval.
- Human approval is required for high-impact actions such as payments, external communications and deletions. Step, time and spend limits apply, and a kill switch is available.
- Input and output guardrail filters are used as defence in depth and are never the only control.
- Each run is traced: prompt, tools called, arguments, results and approvals.

5.6 Logging, monitoring and audit

- Logged events include authentication, administrative actions, model and version used, request metadata, retrieved document identifiers, tool calls, approvals and policy blocks. Logging of full content is a per-client decision.
- Logs are stored locally in append-only or integrity-protected form (for example hash chaining or write-once storage) and forwarded to the client's SIEM with synchronised time.
- Alerting covers egress attempts, unusual volumes, repeated policy blocks and privilege changes.
- Evaluation and prompt-injection test suites are re-run on a schedule and on every model or configuration change to detect drift.

5.7 Operations and resilience

- Patching on a regular cadence with an emergency path, delivered through the transfer station, tested in staging first and reversible.
- Documented backup, recovery and change-management procedures with approvals and records.
- An incident-response plan that covers AI-specific cases: rolling back a model, rebuilding an index, revoking tokens and preserving evidence.

Verification: evidence, not assertion

A security property that has not been tested is a hypothesis. For each deployment we agree a verification plan with the client and record the method and result, instead of quoting a generic figure.

PROPERTY	HOW IT IS TESTED	EVIDENCE DELIVERED
No unapproved egress	Boundary packet capture during a representative workload, plus canary connection attempts from each zone that must fail.	Capture method, duration, result and the approved-destination list.
Hardened configuration	Review against an agreed hardening baseline for hosts, containers and network devices.	Baseline, deviations and their disposition.
Access control on retrieval	Queries by test identities with different entitlements, checked against source-system permissions.	Test matrix and results.
Supply-chain integrity	Recorded hashes compared with publisher values; SBOM and image-signature checks; diff on each update.	Import register and SBOM.
Resistance to misuse	Red-team exercises covering prompt injection, jailbreaks, data leakage and tool abuse against the deployed system.	Findings, severity, remediation and retest.
Auditability	Trace reconstruction: pick an interaction and rebuild who asked what, which model answered, and which documents were used.	Sample traces and log-integrity check.

Findings are tracked to closure and retested. Independent testing by a third party is encouraged and supported.

Governance and regulatory alignment

We structure our work around recognised frameworks. The table states how each is used, and none of these uses is a claim of certification or compliance.

FRAMEWORK	HOW WE USE IT	NOTE
ISO/IEC 27001	Design reference for information-security controls and the evidence that supports them.	Not a certification.
ISO/IEC 42001	Structure for AI governance: roles, risk assessment, lifecycle controls and continual improvement.	Not a certification.
NIST AI RMF	The Govern, Map, Measure and Manage functions organise risk work and reporting.	Voluntary framework.
SOC 2 Trust Services Criteria	Design reference for security, availability and confidentiality controls.	No attestation is implied.
OWASP LLM Top 10, MITRE ATLAS	Threat modelling and test-case selection.	Updated as the guidance evolves.
EU AI Act	Helping the client establish its role (provider or deployer), likely risk category, transparency duties, record-keeping, logging and human-oversight documentation.	Depends on the system, its use and the client's role. Legal advice is required.
POPIA and GDPR	Privacy by design, data minimisation and support for impact assessments. Keeping processing local reduces cross-border transfer exposure.	Controller duties remain with the client.

WHAT AIR-GAPPING DOES AND DOES NOT DO

It does remove the third-party data path, reduce transfer and vendor-dependency risk, and give the client direct control over logs and evidence.

It does not by itself change how a system is classified or regulated, remove the client's obligations under laws such as the EU AI Act, POPIA or GDPR, or protect against insider misuse, prompt injection or poor operational practice. Those need the controls in this paper and the client's own governance.

How we handle your information during an engagement

Our default practice, set out in the engagement agreement, is as follows:

- **NDA first.** A mutual non-disclosure agreement is executed before any topology, network map or data catalogue is shared.
- **Work in your environment.** Work is performed on client-controlled hardware or environments wherever possible, using client-issued accounts.
- **Your data stays in your environment.** The systems we build keep client data in the client's environment, not ours, by default. If a client chooses to share data with us for analysis, we remove it on terms agreed with the client.
- **Least privilege, time-boxed access.** Access is limited to what the task needs and revoked when the engagement or milestone ends, with an access log available to the client.
- **Client ownership.** Code, adapters, weights and configuration produced for the client are assigned to the client under the contract.
- **Client channels.** Sensitive documents are exchanged through the client's own channels.
- **Incident notification.** A security incident affecting client information is notified without undue delay, on the timeline set in the contract.

APPENDIX A

Questions to put to any AI vendor

These questions are useful whether you are evaluating us, a cloud provider or anyone else. A good vendor will answer each with evidence.

- ☐ Where do prompts, documents, embeddings and outputs go, and which third parties can access them?
- ☐ Is any telemetry, analytics or crash reporting sent out of the environment, and how do you demonstrate it is not?
- ☐ How do model weights and dependencies enter the environment? What is verified, and what is recorded?
- ☐ Can the model return documents the requesting user is not entitled to see? Where is access enforced?
- ☐ What can an AI agent do without a human approving it? What are its credentials and limits?
- ☐ What do you log, where is it stored, who can alter it, and can I reconstruct a single interaction end to end?
- ☐ How is prompt injection tested, how often, and what were the last findings?
- ☐ Who administers the platform, and can they read my business content?
- ☐ Which certifications or attestations do you hold, issued by whom and covering what scope? (Frameworks used as references do not count.)
- ☐ How are vulnerabilities in models and dependencies patched in a disconnected environment?
- ☐ What happens to my data, weights and adapters at the end of the contract?
- ☐ What is your incident-notification commitment, and has it been tested?

APPENDIX B

Security posture of this website

As a matter of consistency, this is how reinvented.digital protects its own public website as of the date of this paper:

- Served over HTTPS with HTTP Strict Transport Security, and a strict Content Security Policy that does not permit inline scripts.
- Fonts, icons and scripts are hosted on our own domain rather than loaded from third-party services.
- The enquiry form is protected by Cloudflare Turnstile bot verification, per-address rate limiting, input validation and a request-size limit.
- The site sets no cookies of its own and uses privacy-oriented Cloudflare Web Analytics.
- Enquiry details are used only to respond to the enquiry and are retained for up to 24 months. See the Privacy Policy at reinvented.digital/privacy.
- Submitting the form sends an automatic confirmation to the visitor's address through the email provider Brevo. It repeats only the two options the visitor selected, never free-text they typed, so the form cannot be used to send arbitrary text to a third party's address.
- The optional site assistant sends a visitor's question to Cloudflare Workers AI, which is why visitors are asked not to include confidential details.

The website is not the deployment described in this paper. It is a public marketing site and holds no client data.

APPENDIX C

Limitations and revision history

Limitations. No system is completely secure. This paper describes a design approach and default controls, not a guarantee of outcomes, and it should be read together with the terms of any engagement. Technology, threats and regulation change, so the controls described here are reviewed and updated. Provided for information only, without warranty. It is not legal, regulatory or compliance advice.

VERSION	DATE	CHANGE
1.0	24 September 2026	Initial public release.
1.1	24 September 2026	Appendix B: added the Brevo confirmation email.
1.2	24 September 2026	Section 8: clarified how client data is handled.

CONTACT

To discuss a deployment, request an executive architecture briefing at reinvented.digital/#scheduler. For privacy matters, write to privacy@reinvented.digital.

© 2026 reinvented.digital. All rights reserved.