TECHNICAL BRIEF · v1.0

Sanctum

Sovereign AI Appliance

Architecture, controls, operations, and service levels for the AI appliance deployed inside the customer perimeter. Prepared for architecture review boards, security officers, and platform teams conducting a formal technical evaluation.

PRODUCT	CLASSIFICATION	ISSUED	OWNER
Sanctum	Confidential	July 2026	Sciens Technologies

SECTION

Contents

The structure of this brief.

“Sovereign AI. Deployed in your perimeter.”

SECTION 01

Executive summary

The appliance model, in one page.

Regulated data stays in the client perimeter. AI is brought to it.

Sanctum is a rack-mounted appliance that runs foundation and fine-tuned models on-premises or in a client-controlled colocation. It is delivered pre-configured, commissioned by Sciens engineers, and operated under a documented service-level agreement. No inference traffic, prompts, embeddings, or model weights leave the customer network without written authorization.

The appliance is engineered for institutions where data cannot be exported: law firms with privileged material, hospital systems handling PHI, banks with covered transaction data, and public-sector programs with residency mandates. Sanctum reproduces the developer experience of a frontier-model API while keeping the data plane, control plane, and telemetry under the customer's administrative authority.

DESIGN PRINCIPLES

- **Sovereignty by default**
  Every byte of prompt, retrieval context, model weight, and audit record is stored and processed inside the customer boundary.
- **Deterministic operations**
  A fixed model catalog, versioned prompts, signed builds, and reproducible evaluations replace opaque hosted-API behavior.
- **Defense in depth**
  Attested boot, signed firmware, network egress controls, HSM-backed keys, and immutable audit are layered by default, not add-ons.
- **Operable by the client**
  Standard rack, standard identity providers, standard observability exporters. No proprietary console is required to run it.
- **Contracted service**
  99.9% availability, response-time commitments, quarterly compliance attestations, and a written exit path.

SECTION 02

Reference architecture

Three layers, one boundary.

Signed builds. Isolated runtime. Nothing crosses the boundary without an approved rule.

CLIENT NETWORK	The customer VLAN or VPC. Hosts identity, SIEM, and ticketing systems. Consumes Sanctum through the OpenAI-compatible API and admin console.
SANCTUM APPLIANCE	Hardened operating system with attested boot, signed firmware, HSM-backed keys, and controlled egress. The administrative plane runs here.
MODEL RUNTIME	GPU-backed inference, vector store, and immutable audit log. No third-party inference. Full prompt and response trail is retained on-device.

PLANES OF CONTROL

- **Data plane**
  Inference, retrieval, embeddings, and fine-tuning workloads. Runs entirely on appliance GPUs; never traverses a Sciens-controlled network.

- **Control plane**
  Model catalog, prompt registry, role bindings, and policy. Managed by client administrators through SAML/OIDC-backed console and API.

- **Telemetry plane**
  Metrics, logs, and traces exported to the customer SIEM and APM stack. Sciens support receives only what the client explicitly forwards.

SECTION 03

Threat model & trust boundaries

STRIDE-aligned. Assumptions stated explicitly.

IN SCOPE

- **Exfiltration of prompts or context**
  Blocked by egress controls, DNS pinning, and network policy enforced at the appliance NIC.
- **Model weight theft**
  Weights are encrypted at rest with HSM-held keys; runtime memory is isolated per tenant workload.
- **Supply-chain tampering**
  Signed firmware, signed model artifacts, and attested boot with measured launch verified before serving.
- **Privileged insider misuse**
  Four-eyes admin actions, immutable audit, separation of duties between platform and model administrators.
- **Prompt injection & data poisoning**
  Retrieval sources are versioned and signed; injection filters and output policies applied per workspace.

OUT OF SCOPE · CUSTOMER-OWNED

- **Physical security of the rack**
  Customer datacenter controls apply. Chassis intrusion detection is available as an option.
- **Endpoint compromise of API consumers**
  Application-layer authentication and DLP remain the customer's responsibility.
- **Model correctness for un-evaluated tasks**
  Only tasks with a signed evaluation set are covered by the service level.

SECTION 04

Data flow & residency

Where the bytes go, and where they do not.

INFERENCE REQUEST LIFECYCLE

- **STEP 01**
  Client application authenticates against the customer identity provider and receives a scoped token.
- **STEP 02**
  Request enters the appliance through the OpenAI-compatible API endpoint on the internal VLAN.
- **STEP 03**
  Policy engine evaluates workspace, model, and data-class rules; unauthorized combinations are rejected.
- **STEP 04**
  Retrieval layer pulls context from approved sources (S3, SharePoint, Snowflake, Postgres) over the internal network.
- **STEP 05**
  Model runtime executes inference on local GPUs; response is streamed back to the client.
- **STEP 06**
  Prompt, retrieved context, response, and policy decision are written to the immutable audit log.

RESIDENCY GUARANTEES

PROMPTS & RESPONSES	Never leave the appliance. Not written to any Sciens-controlled storage.
EMBEDDINGS & INDEXES	Materialized on appliance storage. Encrypted at rest.
FINE-TUNED WEIGHTS	Customer-owned. Bound to the appliance serial; export requires signed release.
SUPPORT TELEMETRY	Only what the customer forwards. Default profile excludes prompt bodies and identifiers.

SECTION 05

Deployment topologies

Three shapes. One security posture.

- **On-premises (dedicated rack)**
  Deployed in the customer datacenter. Full airgap or restricted egress. Preferred for classified, PHI, and privileged legal work.

- **Client-controlled colocation**
  Deployed in a colo cage under the customer's contract. Sciens performs commissioning; ongoing operations are shared or customer-run.

- **Sovereign cloud (customer tenant)**
  Deployed in the customer's VPC on a certified region. Same runtime; virtualized hardware envelope. Used when procurement mandates cloud-only.

SIZING TIERS

DEPARTMENTAL	1 node · 4× H100 · up to 40 concurrent users · single workspace
BUSINESS UNIT	2 nodes · 8× H100 · up to 200 concurrent users · multi-workspace
ENTERPRISE	HA pair, N+1 · 16–32× H200 · 1,000+ users · federated workspaces

SECTION 06

Model catalog & runtime

A curated shelf, versioned like software.

Every model has an owner, a version, a license record, and an evaluation set.

FRONTIER & OPEN-WEIGHT MODELS

GENERAL REASONING	Llama 3.3 70B, Llama 3.1 405B, Mistral Large 2, Qwen 2.5 72B, DeepSeek V3
CODE	DeepSeek-Coder V2, Qwen 2.5 Coder, Codestral
MULTILINGUAL	Aya 23, Qwen 2.5, Gemma 2 (customer choice)
VISION & DOCUMENT	Llama 3.2 Vision, InternVL 2, Qwen 2.5-VL
SPEECH	Whisper Large v3, Distil-Whisper (on request)

RUNTIME & SERVING

SERVING STACK	vLLM, TensorRT-LLM, SGLang. Batched, streaming, speculative decoding.
QUANTIZATION	FP16, BF16, FP8, AWQ, GPTQ. Selected per model with evaluated quality delta.
CONTEXT WINDOWS	Up to 128k tokens standard; 1M tokens on qualified models.
CONCURRENCY	Continuous batching. Per-workspace queues and priority classes.
COLD-START	Warm pools; median first-token latency under 400 ms on tier-1 models.

SECTION 07

Hardware envelope

Standard rack. Standard datacenter.

COMPUTE & MEMORY	
GPU	NVIDIA H100 SXM / H200; 4× or 8× per node
HBM	Up to 1.1 TB HBM3e across an 8-GPU node
CPU	Dual Intel Xeon 5th gen or AMD EPYC 9004; 96–128 cores
SYSTEM RAM	1–2 TB DDR5 ECC
STORAGE & NETWORK	
NVME	30–120 TB local, RAID-10, AES-256 self-encrypting drives
BACKING	Optional NAS/S3 for cold artifacts; encrypted at rest and in transit
NETWORKING	Dual 100 GbE data; 25 GbE management; airgap or allow-listed egress
FABRIC	NVLink / NVSwitch intra-node; RoCE v2 inter-node for HA pairs
FACILITY	
FORM FACTOR	4U or 8U rack chassis; ships pre-cabled
POWER	Redundant 3 kW PSUs; 208V single or three-phase
COOLING	Air-cooled to 35 °C inlet; liquid-cooled option for H200 8-way
WEIGHT & RACK	Fits standard 42U rack; installation kit included

SECTION 08

APIs & integration surface

OpenAI-compatible, on your network.

Existing SDKs and orchestration code work by changing only the base URL and key.

PROTOCOLS & SDKS	
REST	OpenAI-compatible: chat/completions, embeddings, files, fine_tuning, batches
GRPC	Streaming inference and bulk embedding for high-throughput pipelines
SDKS	Python, TypeScript, Java, .NET; drop-in for OpenAI, LangChain, LlamaIndex
ASYNC	Batch job API for long-running document workloads; webhook callbacks

IDENTITY & ACCESS	
FEDERATION	SAML 2.0, OIDC; SCIM 2.0 for user and group provisioning
AUTHORIZATION	Workspace, model, and data-class RBAC; ABAC policy hooks
KEYS	Per-user and per-service tokens; rotation, scoping, expiry enforced
PAM	Break-glass admin flow with dual approval and time-boxed elevation

DATA CONNECTORS (CUSTOMER-SIDE)	
OBJECT STORAGE	S3, Azure Blob, GCS, MinIO
FILES & MAIL	SharePoint, OneDrive, Google Drive, Exchange (read-only by default)
WAREHOUSES	Snowflake, BigQuery, Databricks, Redshift
DATABASES	Postgres, SQL Server, Oracle, MySQL via JDBC/ODBC

SECTION 09

Retrieval, memory & fine-tuning

Grounding without exporting.

RETRIEVAL

- **Ingestion**
  Scheduled or event-driven pulls from approved sources; PII detection and redaction hooks before indexing.
- **Chunking**
  Semantic and layout-aware; document structure preserved for citations.
- **Embeddings**
  BGE-M3, E5-mistral, domain-tuned encoders; multi-vector supported.
- **Vector store**
  PGVector or Milvus on the appliance; per-workspace isolation.
- **Reranking**
  Cross-encoder rerankers; hybrid BM25 + dense; per-query policy filters.

FINE-TUNING

- **Methods**
  LoRA, QLoRA, DPO, full-parameter SFT. Customer-owned weights and datasets.
- **Evaluation**
  Held-out sets, task-specific rubrics, regression suites. Signed evaluation report per release.
- **Governance**
  Training-data lineage recorded; opt-in retention. Weights bound to the appliance identity.

SECTION 10

MLOps & release lifecycle

Versioned like production software.

- **Prompt registry**
  Every production prompt is versioned, reviewed, and signed. Rollbacks are one command.
- **Model registry**
  Base model, adapter, quantization, and evaluation report are pinned together as a single release.
- **Environments**
  Dev, staging, and production workspaces on the same appliance with promotion gates.
- **Change control**
  CAB-friendly release notes, diffable configuration, signed by two roles.
- **Rollout**
  Canary, shadow, and A/B on live traffic with automatic rollback on regression.
- **Evaluations**
  Golden sets, LLM-as-judge, human review; scores stored alongside the release.

SECTION 11

Observability & audit

One record of every request.

TELEMETRY	
METRICS	Prometheus / OpenMetrics: latency, tokens, GPU util, queue depth, error class
TRACES	OpenTelemetry across API, policy, retrieval, and model steps
LOGS	Structured JSON; forwarded to Splunk, Datadog, Elastic, or the customer SIEM
DASHBOARDS	Pre-built Grafana boards; exportable JSON for customer standards

AUDIT	
COVERAGE	Every prompt, response, retrieval, admin action, and policy decision
STORAGE	Append-only, hash-chained, WORM-compatible. Retention configurable.
ACCESS	Read-only role for auditors; export to customer eDiscovery on request.
REDACTION	Field-level redaction for PII in dashboards while preserving the full audit.

SECTION 12

Security controls

Layered, attested, reviewed.

PLATFORM HARDENING	
BOOT	UEFI Secure Boot, measured launch, TPM 2.0 attestation before serving
OS	Minimal hardened Linux; CIS-benchmarked; read-only root; auto-patched
CONTAINER	Rootless; seccomp, AppArmor, and mandatory image signing (cosign)
NETWORK	Egress default-deny; explicit allow-list; DNS pinning; mTLS internal
CRYPTOGRAPHY	
IN TRANSIT	TLS 1.3 with modern cipher suites; internal service mesh mTLS
AT REST	AES-256-GCM; keys held in on-appliance HSM or customer KMS
WEIGHTS	Encrypted at rest and during load; runtime memory isolated per workload
ROTATION	Automatic key rotation with dual-control override
GOVERNANCE	
SEPARATION OF DUTIES	Platform admin, model admin, and auditor roles are non-overlapping
FOUR-EYES ACTIONS	Model publish, policy change, and egress rule change require dual approval
VULNERABILITY MGMT	Weekly CVE scan; SLA-bound patching; signed patch bundles
PEN TESTING	Annual third-party test; findings and remediation shared with customer

SECTION 13

Compliance mapping

Controls mapped to the frameworks your auditors already use.

SOC 2 TYPE II	Security, Availability, Confidentiality trust criteria
ISO 27001	Full ISMS scope covering appliance software and delivery
HIPAA	Technical, administrative, and physical safeguards; BAA available
GDPR / UK GDPR	Data minimization, residency, right-to-erasure workflows
EU AI ACT	High-risk system technical documentation and logging obligations
NIST AI RMF	Govern, Map, Measure, Manage functions documented per deployment
PCI DSS	Deployable within a compliant environment; scope-reduction guidance
FEDRAMP PATH	Controls aligned to Moderate baseline for public-sector programs

SECTION 14

Service levels & support

Contracted, measured, reported.

AVAILABILITY & RESPONSE	
AVAILABILITY	99.9% monthly at the appliance interface (HA pair: 99.95%)
P1 RESPONSE	Acknowledged in 30 minutes; 24×7 on-call engineering
P2 RESPONSE	Acknowledged in 2 business hours
RESTORE	Documented RTO / RPO per deployment; disaster-recovery runbook

LIFECYCLE	
UPDATES	Signed model, firmware, and platform bundles; staged rollouts
MODEL REFRESH	Quarterly catalog review; new models qualified with evaluation reports
ATTESTATIONS	Quarterly compliance report; annual penetration-test summary
EXIT TERMS	Documented data export, weight release, and decommissioning procedure

SECTION 15

Acceptance & pilot plan

Four weeks. Written criteria. See it in your environment.

- **Week 1 · Discovery**
  Use cases, data sources, identity, network, and success metrics documented and signed.
- **Week 2 · Install**
  Appliance racked or provisioned; integrations wired; baseline evaluations run.
- **Week 3 · Workloads**
  Two to three production-shaped workloads deployed with signed prompts and retrieval sources.
- **Week 4 · Review**
  Metrics, audit sample, security review, and go/no-go with the architecture review board.

ACCEPTANCE CRITERIA · TEMPLATE

- All prompt, response, and audit records remain on the appliance during the pilot window.
- Latency, quality, and cost targets meet the numbers agreed in Week 1.
- Security review completed with no unresolved high-severity findings.
- Operational runbooks handed over and rehearsed with the customer team.

SECTION A

Appendix · Software suite

Everything shipped inside the appliance.

Sanctum is not a bare model runtime. It is a complete, opinionated software stack designed for a single job: let regulated professionals use modern AI on their own documents, inside their own perimeter, without relinquishing control of the data, the weights, or the audit trail.

STACK AT A GLANCE

CONTROL PLANE	Sanctum OS. Web console for chat, workspaces, model selection, user management, telemetry, and system health. Served locally over TLS 1.3.
SECURITY LAYER	Guard. RBAC with attribute-level permissions, policy engine, tamper-evident write-once audit log, MFA, SSO via SAML 2.0 and LDAP.
APPLICATIONS	Search, Summarise, Translate, Chat, Draft assist, Q&A over corpus. All grounded on customer documents. All inference local.
DATA PLANE	Connect (DMS connectors). On-device ingestion, chunking, embedding, vector index, versioning, and deletion workflows.
RUNTIME	Model server (vLLM / TGI class). Catalog: Llama 3.3 70B, Llama 3.1 405B, Qwen3 72B / 235B, Mistral Large 2, DeepSeek V3.
OPS	Studio admin console. Package registry client. Signed release channel. Health telemetry stays on-device by default.

SECTION A.1

Sanctum OS

The control plane your users see.

Sanctum OS is the web interface every user opens. It runs as a service on the appliance and is reachable only on the firm's LAN or private VLAN. There is no Sanctum-hosted tenant, no shared URL, no external assets.

CHAT & WORKSPACES	Multi-turn chat with context isolation per user and matter. Workspaces group documents, conversations, and prompts under access-controlled scopes.
MODEL SELECTOR	Users pick from models the administrator has enabled. Default routing is policy-driven; sensitive workspaces can be pinned to a specific model or quantization.
PROMPT LIBRARY	Curated, versioned prompt templates for legal drafting, medical summarisation, financial memo preparation, and generic knowledge work.
CITATIONS & TRACE	Every generated answer links back to the source chunks it drew from. Users can open the original document at the exact page, verify, and cite.
LOCAL FILE DROP	Users attach documents to a session for one-off analysis. Attachments are indexed into the session scope only and purged on session close per policy.

SECTION A.2

Application suite

Packaged applications, out of the box.

Sanctum ships with a set of packaged applications tuned for regulated knowledge work. They share the same identity, policy, and audit layer, and operate against the firm's indexed corpus with zero outbound calls during inference.

SEARCH

On-device vector search across the firm document corpus. Hybrid retrieval combining dense embeddings and lexical BM25. Sub-second recall across millions of chunks with citation-grade source pinning.

SUMMARISE

Structured summarisation of documents, threads, and matter files. Configurable templates: executive brief, legal memo, clinical handover, deal one-pager.

TRANSLATE

On-premise translation for documents and correspondence across the major business language pairs. Glossaries and firm terminology can be pinned per practice group.

DRAFT ASSIST

In-line drafting help for memos, letters, clinical notes, and deal documents, grounded on the firm precedent library. Suggestions are attributable to the source precedents.

Q&A OVER CORPUS

Ask a natural-language question against a scoped set of documents, matters, or clinical records and receive a cited answer. Retrieval is scoped by ACL.

EXTRACT & STRUCTURE

Templated extraction of clauses, obligations, medications, dosages, financial line items, and structured fields. Emits JSON that downstream systems consume on the LAN.

SECTION A.3

Guard, Connect, Studio

Security, integrations, and operations.

GUARD	Policy and audit. RBAC with attribute-level permissions on documents, workspaces, prompts, and models. Central policy engine enforces retention, redaction, PII masking, and export restrictions. Write-once, append-only, cryptographically-signed audit log. SIEM export in CEF and JSON.
CONNECT	Document management bridges. Bidirectional, versioned connectors: iManage v10+, NetDocuments, Clio available today. Worldox, OpenText, and SharePoint DMS on the roadmap. Retrieval is on-device; ACLs mirrored from the source system.
STUDIO	Administrator surface for user and group management, model enablement and pinning, index management, retention and deletion workflows, package updates, health metrics, and audit review.
FINE-TUNE WORKBENCH	Optional LoRA and QLoRA fine-tuning on firm data using the packaged trainer. Weights produced by the workbench are stored on-device, versioned, and pinnable to specific workspaces.

SECTION A.4

Runtime, APIs and extensibility

What developers get to build on.

Sanctum exposes a stable set of local APIs so the firm can build its own tooling on top of the appliance. Every endpoint is authenticated, audited, and reachable only on the LAN.

INFERENCE API	OpenAI-compatible /v1/chat/completions, /v1/completions, /v1/embeddings. Existing SDKs work by pointing at the appliance base URL. Streaming, function calling, and structured JSON supported.
RETRIEVAL API	Scoped search and rerank endpoints over the on-device vector index. Filters by workspace, matter, document type, and ACL. Returns citations with page and chunk offsets.
INGESTION API	Programmatic document upload, batch ingestion, deletion, and reindex. Emits webhooks on completion.
ADMIN & AUDIT	User, group, model, policy, and package management under Guard permissions. Full audit-log export in CEF and JSON for SIEM integration.
RUNTIME	High-throughput model server with continuous batching and paged attention. AWQ, GPTQ, and FP8 quantization supported.
EXTENSIBILITY	Signed plugin surface for firm-built extensions. Plugins run inside the Guard sandbox with explicit permissions.

SECTION 16

Glossary & references

Terms and sources used in this document.

GLOSSARY

SANCTUM	The Sciens sovereign AI appliance.
WORKSPACE	A logical isolation boundary with its own models, prompts, retrieval sources, and users.
MODEL REGISTRY	Versioned catalog of base models, adapters, and quantizations.
PROMPT REGISTRY	Versioned catalog of production prompts with owners and evaluations.
ATTESTED BOOT	Hardware-rooted verification that firmware and OS have not been tampered with.
EGRESS DEFAULT-DENY	No outbound network flow is permitted unless explicitly allow-listed.

REFERENCES

- NIST AI Risk Management Framework 1.0 (2023).
- ISO/IEC 27001:2022 and ISO/IEC 23894:2023 (AI risk management).
- EU Regulation 2024/1689 (Artificial Intelligence Act).
- HIPAA Security Rule, 45 CFR Part 164 subpart C.
- OWASP Top 10 for LLM Applications.

NEXT STEP

Book a scoped pilot.

See Sanctum running in your environment. Four weeks. Based on scope.

sciens.app/pilot · contact@scienstechnologies.com