SHAKUDO

The Sovereign AI Reference Architecture

A design and evaluation reference for running agentic AI inside infrastructure you control: your own cloud account today, your own facility later.

COST MODEL · SHARED RESPONSIBILITY · DATA FLOW · NETWORK TOPOLOGY
IDENTITY & ACCESS · SOC 2 AND HIPAA · EVALUATION CHECKLIST

FOR ARCHITECTURE AND SECURITY REVIEW

EDITION 2.0
2026

REFERENCE OVERVIEW

What running agentic AI inside your boundary requires

A controlled deployment has to answer four questions: Does the workload need a trust boundary? What controls enforce it? What will it cost to operate? What evidence will security, finance, and procurement need?

The reference defines those controls, shows the failure modes they prevent, and provides the diagrams, inventories, and checklists needed to evaluate them. Where a question has no settled answer, specify the evidence a vendor must provide.

REVIEW ROLES	EVIDENCE REQUIRED	SECTIONS
Architecture and platform reviewers Director/VP-level architecture review	Circulation-ready architecture and decision material.	3, 4, 5, 6, 7, 8, 13
Business and finance reviewers May require additional finance review	Cost range, risk position, and exit path.	1, 2, 11
Security, risk, and commercial reviewers Separate security/ commercial gate may follow	Responsibility, data flow, inventory, and identity.	9, 10, 11

SCOPE

- Use the same control tests for a cloud account or an on-premises deployment.
- Framework mappings are architectural, not legal advice; this guide does not rank vendors or products.
- Figures are either sourced or identified as estimates. Any benchmark included here identifies its source and methodology.

WORKING DEFINITION

Sovereign AI means you can identify where data rests, which identities can access it, and how it can leave. You retain the authority to change those controls.

EVIDENCE

Figures and quotations are sourced. Verify deployment-specific numbers and regulatory requirements before relying on them.

CONTENTS

SECTION 01

Evaluating a controlled deployment

Two questions determine whether a controlled deployment is justified: What control requirements does the workload have, and can the organisation operate them? A sound architecture still fails if governance, security, or finance cannot approve or sustain it.

Start evidence and risk review early

Vendor qualification runs on its own clock and does not begin when the architecture is approved. In most enterprises it can run in parallel from the first serious evaluation, and the elapsed time is dominated by queueing rather than by work.

> **REVIEW COORDINATION**
>
> *"This would be a long, arduous procurement committee led process."* Technology leader, a large investment management firm. Review typically spans an Architecture Review Board, a security questionnaire, legal and privacy, finance, and vendor onboarding. Start these tracks in parallel once the deployment shape is known.

REVIEW TRACK	EVIDENCE REQUIRED	BEGIN WHEN	TYPICAL DELAY
Architecture review board	Layer diagram, data flow diagram, network topology, integration points.	Before approval, with the deployment boundary and integrations defined.	The data flow diagram does not exist yet, so the meeting is booked twice.
Security questionnaire	SOC 2 report, shared responsibility matrix, component inventory, identity matrix, vulnerability handling.	Same week as the technical evaluation. It is the longest pole.	Required evidence is incomplete, delaying scope and review.
Penetration test	A component inventory precise enough to scope against.	As soon as the deployment shape is known.	Scope cannot be set, so the test cannot be quoted or booked.
Legal and privacy	Data processing terms, BAA where PHI is involved, residency commitments, subprocessor list.	In parallel with security, not after it.	Sequential review: legal waits for security, security waits for an artifact.
Finance	A three-year cost model with a stated method and a range, plus the exit cost.	Before the number is spoken out loud in any meeting.	A single-point estimate that moves, which costs more credibility than a wide range would have.
Vendor onboarding	Financial standing, insurance, references, continuity terms.	Immediately. It is administrative and it is slow.	Discovered late, after everyone believed the deal was done.

Table 1.1 · Run the evidence tracks in parallel where organisationally possible.

Whether the workload needs a controlled boundary

Any one of these, answered yes with specifics, is a defensible basis for private deployment. Answering yes to none means you are about to spend capital for reasons you will not be able to articulate in twelve months.

TEST	THE QUESTION, STATED PRECISELY	WHAT "YES" LOOKS LIKE IN EVIDENCE
Data class	Does the workload process data carrying a legal, contractual or classification restriction on *where* it may be processed or *who* may be a subprocessor?	A named clause: PHI under a BAA; ITAR or EAR controlled technical data; classified material; a customer contract prohibiting third-party model providers; residency commitments in a data processing agreement.
Continuity	Must the capability keep working independent of a vendor's roadmap, rate limits, deprecation schedule or pricing changes?	The workload sits in a process with a multi-year lifecycle, clinical, safety or regulatory, where a model being retired is an incident rather than an upgrade.
Economics	Is sustained utilisation high enough that owned capacity beats on-demand? This turns on measured utilisation.	A measured duty cycle from a real pilot, not a projection. Bursty utilisation genuinely favours rented capacity. See §2 and §5.
Locality	Must inference sit physically near the data or the process, for latency, bandwidth, or because the site has no reliable link?	Plant floor, vessel, substation, forward site or imaging suite: any environment where the WAN is the least reliable component.

When a controlled deployment is a poor fit

Check these boundary conditions before committing capital or operational capacity.

- **Utilisation is low or unpredictable.** Idle accelerators can make owned capacity uneconomic. Rent until the duty cycle is proven.
- **No platform team.** Running this is a standing platform function. Organisations without a durable infrastructure team consistently stall after the first model goes live. If the answer to "who runs this" is one DevOps engineer who already has a job, record that as a finding.
- **No existing facility.** With no data centre footprint and no power budget, on-premises means a facilities project with a lead time in quarters. A deployment inside your own cloud account may satisfy the same control requirement far sooner.
- **Restriction requires a named clause.** If no one can name it, record that as a finding.

> **THE THREE FAILURE MODES**
>
> **1. Buying compute before defining the platform.** Accelerators arrive, identity and networking and storage are unsolved, and the rack draws power for two quarters while the platform is designed around it.
>
> **2. Treating it as one model deployment.** The second and third teams each rebuild serving, retrieval and auth their own way. Governance is now per-team, which means it does not exist.
>
> **3. No upgrade story.** The environment freezes at its day-one model and library versions when the path to replace them was not designed. Eighteen months later the stack is the reason the organisation is behind.

Build or adopt

The nine planes in §3 are all well-understood problems with mature open-source components. Assembling, securing and maintaining them is a multi-year platform programme. The review should decide where your engineers spend the next eighteen months. Either answer is defensible when its assumptions are stated.

Staffing is usually the deciding input. If the build option assumes hiring that has already been refused, model the plan the organisation can staff and show what the gap costs.

Decision summary

The recommendation should state its assumptions, boundary, cost range, risk position, and requested decision. It will be read by people who were not in the evaluation.

SIX LINES FOR THE DECISION

A concise recommendation should contain:

1. **The restriction.** The named clause or policy that rules out the default option. One sentence.
2. **The workload.** What it does, who uses it, and the manual process it replaces.
3. **The boundary claim.** Where data comes to rest and what enforces that. Point at the data flow diagram.
4. **The cost.** Three-year range with the method stated, and what the number is sensitive to. See §2.
5. **The risk position.** Vendor continuity, exit cost, and what happens if the supplier disappears. See §11.
6. **The ask.** The decision you want, the date you need it, and what stops if it slips.

Build the evidence pack first

Before asking for approval, assemble the deployment boundary, data-flow diagram, identity model, cost range, compliance mapping, and exit conditions. An answer without an artifact is not a control.

SECTION 02

Cost model and key sensitivities

A credible cost model states the range, assumptions, utilisation, staffing, and exit cost. Cost predictability matters to reviewers who will not attend the architecture review.

Approval thresholds shape the business case

Before modelling anything, identify where the number lands relative to your organisation's approval thresholds. Confirm those thresholds with finance before building the model.

REPORTED THRESHOLD	ADDITIONAL REVIEW TYPICALLY REQUIRED	WHAT THE NUMBER THEN HAS TO INCLUDE
Discretionary authority, described at one mid-market organisation as up to roughly $50k in a day	Technical and finance review remains within existing authority.	A defensible unit cost and a stopping condition.
Around $100k of annual budget, described as the point where a spend needs justifying	Additional business and finance review.	Three-year total, the alternative it displaces, and the cost of doing nothing.
A million dollar solution	Executive sponsorship and usually a formal business case.	Everything above, plus exit cost, continuity risk and a phased commitment.

Table 2.1 · Thresholds vary by organisation and these are illustrative. The useful action is to ask your own finance partner for the bands before you build the model, not after.

The line items

A cost model includes more than the GPU price. The accelerator is a minority of true cost in a self-operated environment, and the items that get omitted are the ones finance will find.

CATEGORY	LINE ITEMS	BEHAVIOUR	COMMONLY OMITTED
Compute	Accelerators (owned or rented), CPU nodes, storage tier for weights, indexes and logs.	Capital or usage	Storage growth from audit retention, which is contractual rather than optional.
Facility	Power, cooling, rack space, power distribution work, floor loading remediation.	Fixed, long lead	This category is commonly omitted from initial cost models. See §5.
Network	Interconnect (NVLink, InfiniBand or RDMA fabric), load balancers, egress bandwidth, private links.	Capital plus usage	Interconnect, which is not optional if you use tensor parallelism.
Platform	Licences or subscription, support tier, non-production environments.	Subscription	Development, staging, disaster recovery, and demo environments add cost; production is rarely the only environment.
Model	External API tokens still consumed, open-weight model evaluation effort, fine-tuning runs.	Usage, volatile	The agentic multiplier. See §5.
People	Platform and infrastructure FTEs, security review effort, on-call rotation.	Fixed, recurring	Staffing can be a major recurring cost and is often the hardest line to get approved.
Assurance	Penetration test, audit evidence work, third-party attestation costs.	Annual	Recurring pen tests, which do not stop after year one.
Exit	Migration away, data extraction, retraining, residual hardware value.	Contingent	Almost always. See §11.

Table 2.2 · Line items for a three-year model. Categories rather than prices. Fill the numbers from your own quotes.

The arithmetic

```
TCO(3yr) = compute + facility + network + platform + models + people + assurance + exit
// compute, if owned
compute  = (accel_count × unit_price) + refresh_reserve + storage(TB × $/TB/mo × 36)
// compute, if rented, at duty cycle d
compute  = accel_count × $/hr × 8760 × 3 × d
// people: the line that decides most models
people   = FTE_count × fully_loaded_cost × 3
// models: apply the agentic multiplier from §5 before believing this
models   = users × sessions/user/mo × tokens/session × M × $/token × 36
```

Own or rent, on a measured duty cycle

Rental for H100-class capacity fell from roughly $8 to $10 per hour at the 2023 to 2024 peak to roughly $2.00 to $3.50 by late 2025 and 2026, against a purchase price around $25k to $40k.[11] Prices vary by quote; verify current pricing before use. Using a simple comparison at around $3.00 per hour, a $25k to $40k purchase implies payback of roughly 11 to 18 months at continuous use, before power, cooling, support and staffing. At 40% utilisation, the same calculation implies roughly 29 to 46 months. At 10%, it implies roughly 114 to 183 months, so rental is generally cheaper over a three-year planning horizon. These are illustrative calculations; include operating costs, residual value and measured duty cycle in the full model. Steady-state inference has the duty cycle that justifies ownership. Bursty training does not.

> **INPUTS TO SHOW FINANCE**
>
> Show finance the two or three inputs that most affect the answer. In practice, they are duty cycle, p90 context length, and FTE count. A model that shows how the number changes with those inputs is easier to scrutinize; a point estimate can lose credibility when it moves.

Sources of variance

SOURCE OF VARIANCE	WHY IT MOVES	HOW TO BOUND IT
Agentic token growth	An agent re-sends an accumulating conversation every turn. Token volume grows super-linearly across a task, and multi-agent systems multiply it again.[6]	Per-team token quota at the gateway, enforced rather than reported. Hard ceilings per agent run.
Context length drift	Retrieved documents and tool output push p90 context far above what was modelled, and KV cache scales with it.	Cap retrieved context. Measure p90 in the pilot rather than estimating it.
Idle accelerators	Owned capacity bought for a peak that arrives twice a quarter.	Size to steady state, overflow the peak to rented capacity, and say so in the proposal.
Runaway agents	A loop, a retry storm, or a tool that returns more than expected.	Attribute cost per request in the audit record, so it surfaces within a day. See Table 7.2.
Staffing reality	The model assumed an FTE the organisation will not approve.	Model the operation with the headcount you actually have, then show what the gap costs.
Environment sprawl	Development, staging, disaster recovery and a demo environment.	Count them at the start. They add cost and are often omitted from the initial design.

The cost controls to require of any platform

These are the mechanisms that turn a variable bill into a predictable one. Ask for a demonstration of each control.

- **Per-team and per-agent quota enforced at the gateway.** Reporting after the fact is accounting. Enforcement is a control.
- **Token and cost attribution in the audit record**, to a team or cost centre, per request. See Table 7.2.
- **A hard stop**, not only an alert, when a quota is exhausted, with a documented behaviour for what the user sees.
- **Priority classes under contention**, so that a scarce accelerator degrades a batch job rather than a clinical one.
- **Chargeback data your finance system can ingest** without a manual step.

> **VENDOR COST MODELS**
>
> Ask every vendor for a total cost model on your own workload shape rather than a price list. If they cannot produce one, that is useful information and you should record it. A pilot with metering turned on is the only reliable way to get the inputs, and it is worth negotiating the metering before the pilot rather than after.

The budget line this usually comes out of

Where an AI platform has no budget line of its own, the spend often has to attach to one that already exists. Cybersecurity, compliance and infrastructure budgets are the usual candidates, because the governance and access-control properties of this architecture are the things those budgets already fund. That framing is legitimate when the controls are real. Decide early which budget you are asking against, because it changes which artifacts from §10 you need first.

> **COST OF THE ALTERNATIVE**
>
> Finance will ask what the alternative costs. The defensible version has three parts: the manual effort the workload replaces, measured rather than estimated; the spend already going to overlapping tools, which is often several assistant subscriptions; and the exposure created by ungoverned use of public tools that is already happening. The third is the hardest to quantify and the easiest to evidence, because your egress logs already contain it.

SECTION 03

What the platform has to do

A private AI environment is nine cooperating planes. A cluster that only serves a model covers one plane fully and part of a second. The accelerators are the visible part of the cost and a small part of the risk.

Take the table into a design session and mark each row *owned*, *partial*, or *absent*. The absent rows are your project plan, and usually the reason a previous attempt stalled.

#	PLANE	WHAT IT HAS TO DO	WHAT BREAKS WITHOUT IT	USUAL OWNER
1	**Accelerator & scheduling**	GPU nodes, driver and firmware lifecycle, partitioning (MIG or time-slicing), topology-aware scheduling, quota and preemption between teams.	One team's batch job starves production inference. No way to give finance and clinical separate guarantees.	Infrastructure
2	**Model serving**	Inference engines, KV-cache and batching policy, autoscaling, model-level routing, warm-pool management for large weights.	Latency collapses under concurrency. Loading a 140 GB model on demand becomes a multi-minute cold start.	Platform
3	**Artifact supply chain**	Private container and model registry, provenance and signature verification, offline mirroring of packages and weights, licence tracking, SBOM.	An isolated cluster cannot install anything. It may also install from an unvetted source.	Platform / SecEng
4	**Data & retrieval**	Object storage, vector index, connectors to source systems, and retrieval that respects the permissions of the *requesting user* rather than those of the indexing service.	Enterprise RAG most often fails when a correctly secured document is returned to someone who could not open it in the source system. See §8.	Data platform
5	**Agent runtime & tool sandbox**	Isolated execution for agent-authored code, an allowlisted tool broker, credential brokering, human-approval gates, and hard blast-radius limits per agent.	This is the highest-risk plane. An agent with broad credentials and a shell is an unbounded liability.	Platform / SecEng
6	**Identity, secrets & policy**	Enterprise SSO for humans, workload identity for services, first-class identity for *agents*, short-lived credentials, central secret custody, machine-readable policy.	Long-lived API keys in environment variables. The audit trail cannot identify which identity made a request.	IAM / SecEng
7	**Model access gateway**	A single ingress for all model traffic: authentication, per-team quota, routing between self-hosted and external models, token and cost accounting, request and response capture.	No chokepoint means no enforcement and no cost visibility. Shadow usage routes around whatever you built.	Platform
8	**Observability & audit**	Metrics and traces for operations, plus a separate tamper-evident audit record for governance, with defined retention and an export path for evidence requests.	Operationally blind, and unable to answer an auditor. These are two different systems with two different retention policies. See §7.	SRE / GRC
9	**Lifecycle**	Model and library upgrade path, version pinning with rollback, evaluation gates before promotion, backup and disaster recovery for weights, indexes and audit history.	Failure mode three from §1: the environment freezes and slowly becomes the constraint it was built to remove.	Platform

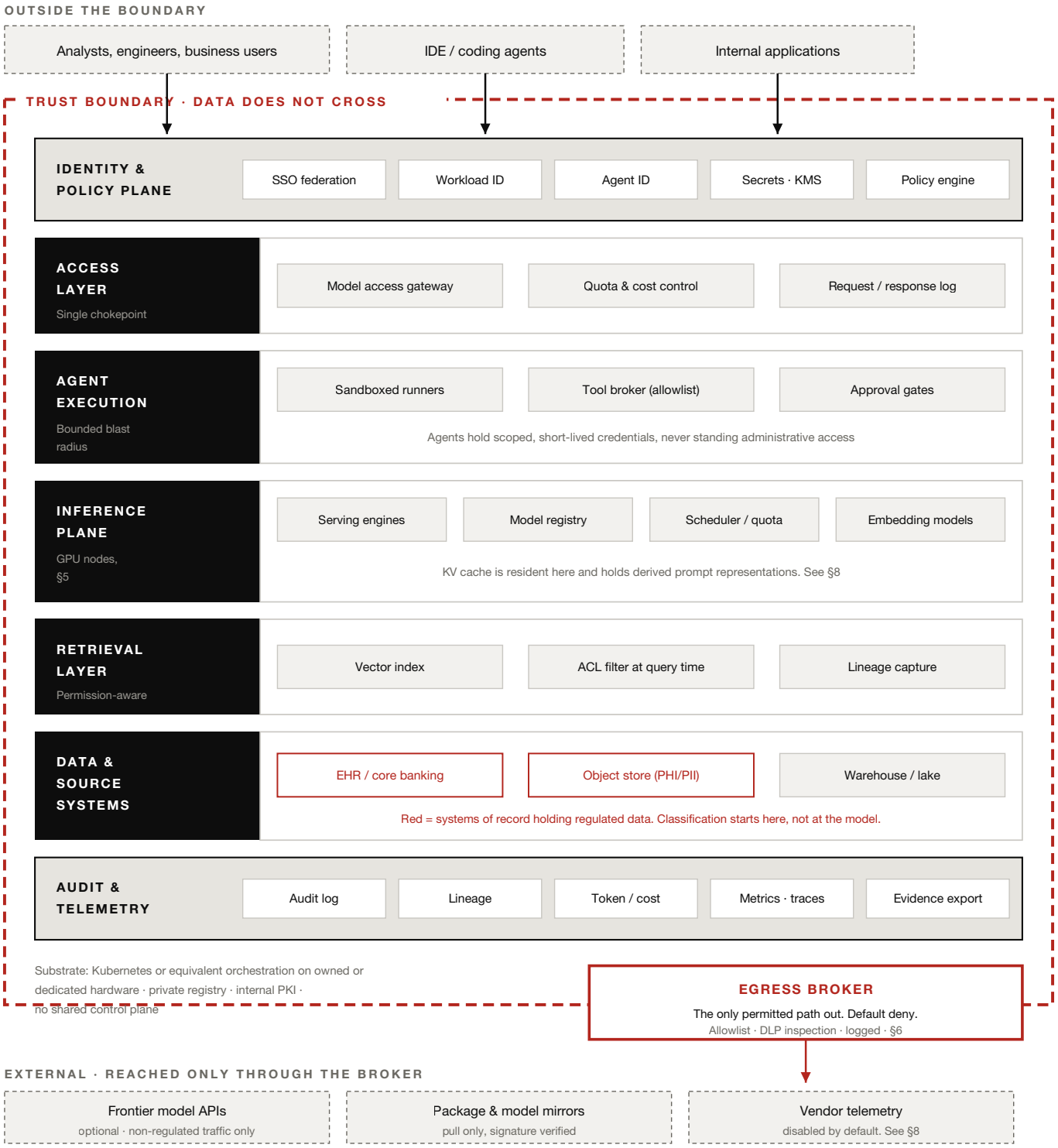
Table 3.1 · The nine planes. Mark each row owned, partial or absent. The absent rows are the project plan.

> **FOR PLATFORM EVALUATION**
>
> When the table is used to evaluate a platform, it becomes a scoring sheet. For each row, ask which party operates it, which party is accountable when it fails, and what evidence it produces. Those three answers populate the shared responsibility matrix in §10, which is the artifact your security team will ask for first.

SECTION 04

Reference architecture

Security review asks for three drawings: the components, how data moves between them, and how the network is arranged. The red line below is the assertion you are making to your reviewers. Sections 6 through 8 are what make it true.



In a fully disconnected deployment the broker is replaced by a manual, reviewed import process and the external row does not exist as a live path.

In a private-cloud deployment the boundary is your own account and network rather than your own building. The layers and controls are unchanged.

Figure 4.1 · Layer view. Two cross-cutting planes (identity, audit) wrap five functional layers. One aperture. Section references point to the control that enforces each claim.

4.2 Data flow

In security reviews this is frequently the document that gates the next meeting. One security lead put the requirement in four words: *"Show me the data flow."* Produce it before the review, with a classification on every hop and a named control on every arrow. Figure 4.2 is the generic shape. This diagram is illustrative; redraw it for the deployment.

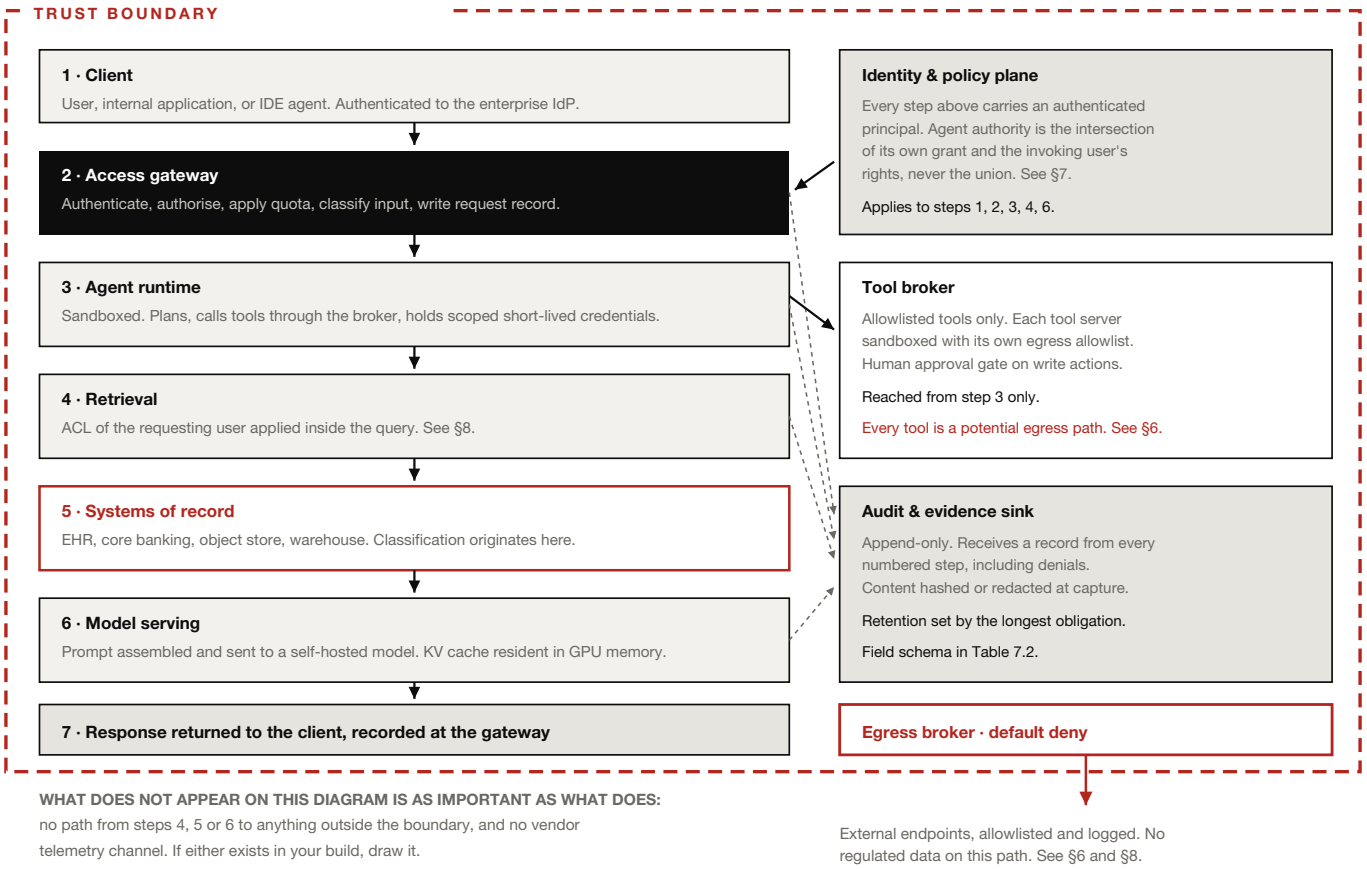


Figure 4.2 · Data flow. Redraw this against your own deployment before the security review and mark each hop with its data classification.

STEP	WHAT MOVES	CLASSIFICATION	CONTROL	EVIDENCE
1 › 2	User prompt, session context, application identity.	Inherits the user's clearance. May contain regulated data.	TLS, enterprise SSO, per-team quota.	Gateway request record with principal and session ID.
2 › 3	Authorised request and a scoped, short-lived credential.	Unchanged.	Agent identity distinct from user identity, with intersected permissions.	Delegation chain in the audit record.
3 › 4	Retrieval query carrying the requesting user's identity.	Unchanged.	ACL applied inside the query rather than as a post-filter.	Document IDs returned and the ACL decision applied.
4 › 5	Read against source systems, or against an index derived from them.	**Regulated at source.**	Source-system permissions, connector service account scoped to read.	Source system access log, correlated by trace ID.
5 › 6	Assembled prompt including retrieved records.	**Regulated. This is the highest-sensitivity hop.**	Self-hosted model inside the boundary. KV cache isolation, no cross-tenant prefix reuse.	Model, version and configuration hash.
6 › 7	Generated response.	Classified at the level of its inputs.	Output classification, DLP where the response leaves the boundary.	Response record, hashed or redacted at capture.
3 › tools	Tool arguments, and whatever the tool returns.	Whatever was passed. Frequently underestimated.	Allowlist, sandbox per tool server, approval gate on writes.	Tool name, argument hash, result status, endpoint contacted.

Table 4.1 · The rows a security architect will read first are 4 to 5 and 5 to 6. Have an answer for both before the meeting.

4.3 Network topology

Security reviews ask for the exact ingress path: what terminates TLS, which load balancer type, where certificates come from, which ranges are permitted. “I will have to check” is a common answer in this conversation and it delays the review. Figure 4.3 is the shape to fill in. This figure is illustrative; confirm it for the deployment.

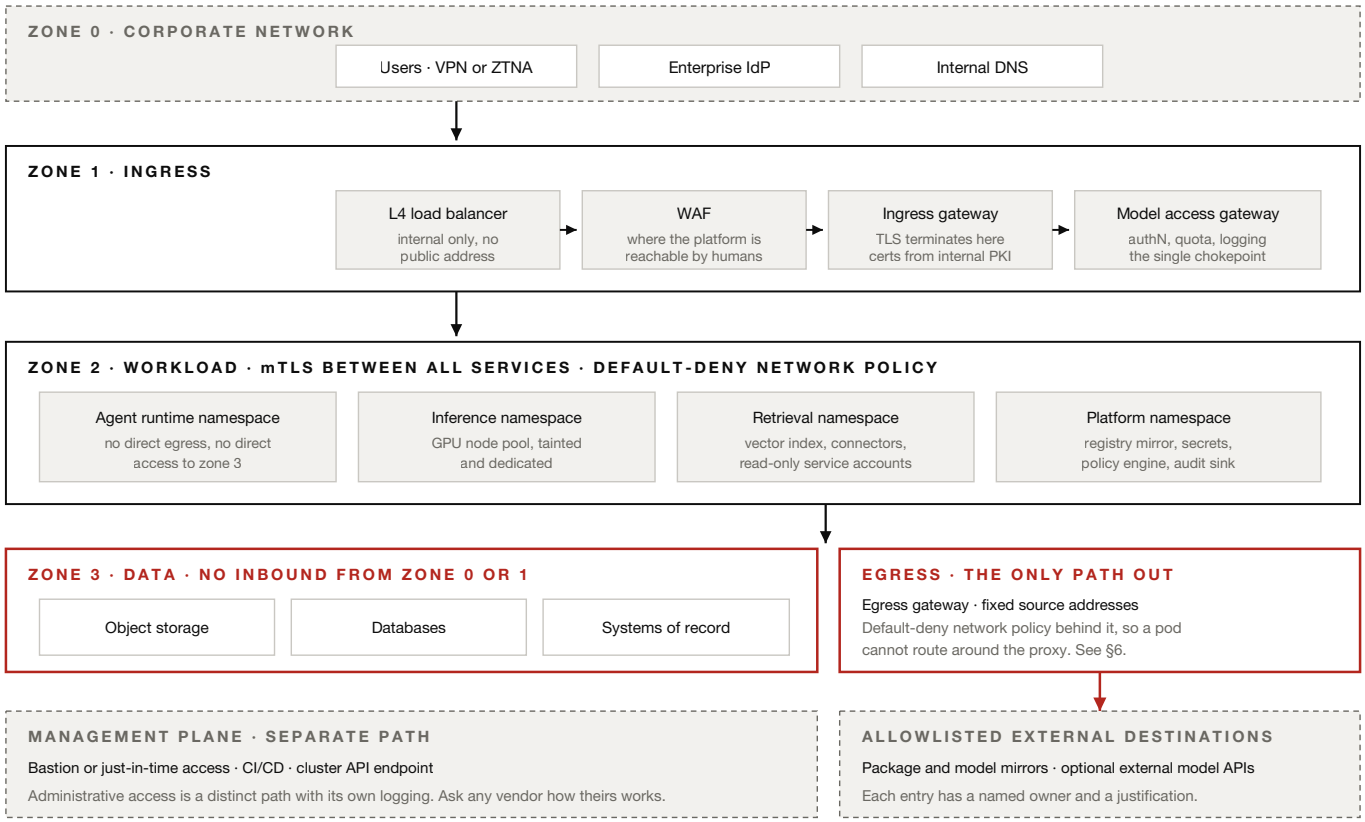


Figure 4.3 · Network zones. The zone boundaries are the parts a security architect will interrogate. Fill in your own load balancer types, address ranges and certificate authority before the review.

The network questions to have answered in writing

Each of these has stalled a review somewhere. Written answers make the review more efficient.

- Which load balancer, at which layer, and is it internal only or does it have a public address?
- Where does TLS terminate, and is traffic re-encrypted behind it?
- Which certificate authority issues internal certificates, and what is the rotation period?
- What are the ingress source ranges, and is there a WAF in front of any human-facing surface?
- Which ports and protocols are open between each pair of zones?
- How is service-to-service traffic authenticated, and is mTLS enforced or merely available?
- What are the fixed source addresses for egress, so the far side can allowlist them?
- How does administrative access reach the cluster, and is it a separate path from user traffic?
- Which DNS resolves platform names, and does it work when the external link is down?
- How does the container registry get updated, and through which path?

SECTION 05

GPU and compute sizing

Sizing conversations go wrong because they start from the parameter count. Parameter count determines whether the model *loads*. Concurrency and context length determine whether the service *works*. Run the arithmetic below with your own numbers.

The memory equation

```
VRAM = weights + KV_cache + overhead   // per replica
// 1 · weights, the part everyone estimates correctly
weights = params × bytes_per_param
   BF16/FP16 = 2 · FP8/INT8 = 1 · INT4/NVFP4 ≈ 0.5
// 2 · KV cache, which decides whether you survive concurrency
kv_per_token = 2 × layers × kv_heads × head_dim × bytes_per_elem
KV_cache = kv_per_token × context_len × concurrent_seqs
// 3 · overhead: activations, CUDA context, fragmentation
overhead ≈ 15% to 20% × (weights + KV_cache)
```

The leading `2` is the key and value tensors. `kv_heads` is *not* the attention head count on any modern model: grouped-query attention shares key and value projections across query heads, and is the reason 70B-class models are servable at all.[1] Using the query-head count here overestimates KV memory by roughly an order of magnitude, the most common arithmetic error in these discussions.

Worked example: a 70B model at long context

Taking a Llama-3.3-70B-class configuration with 80 layers, 8 KV heads, head dimension 128 and an FP16 cache. Confirm the configuration before relying on this example:

```
kv_per_token = 2 × 80 × 8 × 128 × 2 = 327,680 bytes ≈ 0.31 MiB / token
  // one conversation at 128k context:
            = 0.31 MiB × 131,072 ≈ 40 GiB of KV cache
```

Forty gigabytes for one sequence, against 138 GB of weights at BF16.[2] A pair of 80 GB accelerators holds the model and almost nothing else. Ten concurrent long-context conversations need a different machine. Parameter count alone is insufficient; include KV-cache capacity in the sizing case.

> **SIZING IMPLICATION**
>
> At production concurrency, KV cache can become the limiting resource, depending on context length, concurrency, model architecture and cache precision. When capacity is insufficient, the serving engine may preempt requests, recompute or swap their KV state, and degrade latency. An optimistic allocation can still fail. Size for *context* × *concurrency* first, then check the weights fit.

Published memory requirements

MODEL / PRECISION	WEIGHTS ONLY	PLANNING TARGET	PRACTICAL PLACEMENT
Llama 3.3 70B, BF16	138 GB	180 GB	2× H100 80GB holds weights only. Use at least 4×80GB for practical serving; TP4 or TP8 are supported profiles[2][5]
Llama 3.3 70B, FP8	69 GB	90 GB	Fits 1× H100 80GB only with tuning. Comfortable on 1× H200 141GB[2]
gpt-oss-120b, MXFP4	~61 GiB	80 GB	117B total, 5.1B active (MoE). Single 80GB accelerator[3]
DeepSeek-R1 671B, FP8	~700 GB	8× H200	Published recipe: 8× H200 FP8 (TP8 plus expert parallel), or 4× B200 at FP4[4]
Llama 3.1 405B, BF16	~810 GB	n/a	Approximately 243 GB at 4-bit with planning overhead. 8× H100 = 640 GB, so BF16 does not fit one node

Table 5.1 · Vendor-published where available. The first column is approximate model-weight memory. Planning targets reserve runtime and KV-cache headroom; they do not account for your concurrency. Confirm the resulting placement against the serving runtime.

A CAUTION ABOUT QUANTISATION

The cheapest lever available and the easiest to over-apply. FP8 is broadly accepted for production serving. Aggressive 4-bit schemes shift accuracy in workload-specific ways that aggregate benchmarks do not reveal. If the workload is clinical, financial or safety adjacent, treat a precision change as a model change and put it through the same evaluation gate.

Interpreting throughput measurements

Published tokens-per-second figures are not comparable without four qualifiers: precision, batch size, concurrency and the input to output token split. Published results vary because these variables are often different or unstated. Three reported points for a 70B-class model at BF16, four-way tensor parallel:

CONFIGURATION	WORKLOAD	THROUGHPUT	BEHAVIOUR
4× H100 80GB	200 in / 200 out	~7,000 tok/s	at 500 concurrent users, time-to-first-token under 5 s[5]
4× H100 80GB	1,000 in / 200 out	~2,600 tok/s	at 250 concurrent. Longer prompts cost throughput, sharply[5]
4× A100 80GB PCIe	200 in / 200 out	~570 tok/s	saturates around 50 users. TTFT exceeds 10 s at 100[5]

Table 5.2 · An order-of-magnitude gap between accelerator generations on the same model. Use these figures as a workload shape, not a specification; benchmark your own prompt distribution before committing capital.

BENCHMARK EVIDENCE

No vendor benchmark predicts your latency, because none of them ran your prompt distribution. Ask any supplier for throughput and time-to-first-token at a stated concurrency, precision and token split, and treat an answer without all four qualifiers as no answer. If the numbers do not exist, the correct next step is a metered pilot on your own traffic rather than a negotiation about someone else's chart.

The agentic multiplier

Agentic workloads require different sizing assumptions. An agent plans, calls tools, reads results and re-sends an ever-growing conversation every turn, so tokens accumulate super-linearly across a task and long contexts become the norm. The figures cited here with a stated methodology come from Anthropic: agents use roughly **4× the tokens of a chat interaction**, and multi-agent systems roughly **15×**.[6] Do not apply a larger multiplier without a documented measurement method.

SIZING WORKSHEET

1. Use projected *concurrent* users as the input. Registered users and daily actives do not measure concurrency.
2. Multiply by 4 for single-agent workloads, or 15 if you are deploying multi-agent systems.[6]
3. Take the 90th-percentile context length you actually expect, including retrieved documents and accumulated tool output. Agentic workloads sit far higher here than teams predict.
4. Compute `kv_per_token × p90_context × concurrent_seqs`.
5. Add weights at your chosen precision, then reserve a documented 15 to 20% planning margin for runtime and KV-cache headroom. This is an engineering allowance, not a universal vendor requirement.[7]
6. Divide by per-accelerator memory, rounding up to the parallelism topology rather than to the nearest GPU.
7. Add a documented contingency for production demand that the pilot did not model.

Parallelism and interconnect

Official vLLM guidance supports the following design choice for parallel inference: use **tensor parallelism within a node** and **pipeline parallelism across nodes**, and explicitly, *if the GPUs in the node have no NVLink interconnect, use pipeline parallelism rather than tensor parallelism.*[8] Multi-node tensor parallelism needs InfiniBand or RDMA. Over conventional Ethernet it is a well-known way to buy idle accelerators.

Power, cooling, and rack constraints

More often the binding constraint than the hardware budget, and the item most likely to be missing when facilities reviews the proposal.

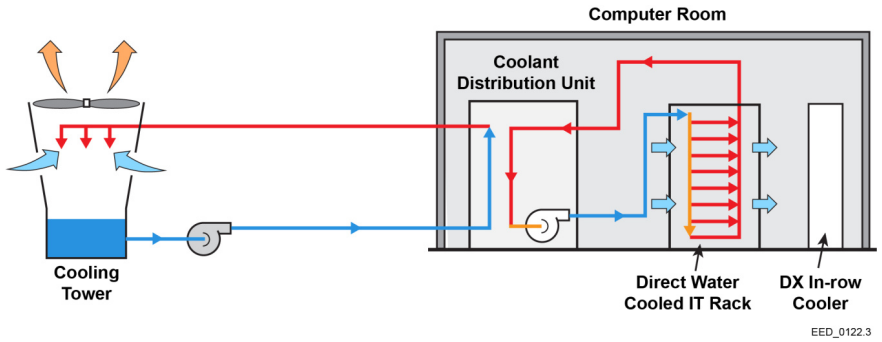


Figure 5.4 · Direct liquid cooling makes the facility coupling explicit: heat leaves the IT rack through a coolant distribution unit and is rejected at the cooling tower. Source: U.S. Department of Energy, Cooling Water Efficiency Opportunities for Federal Data Centers (Figure 2). `https://www.energy.gov/cmei/femp/cooling-water-efficiency-opportunities-federal-data-centers`

ITEM	FIGURE	CONSEQUENCE
One 8-GPU H100 system	~10.2 kW	A single node exceeds the design draw of a whole legacy rack[9]
4 such systems in a 42U rack	~40.8 kW	NVIDIA lists four as the maximum supported density; greater than four is not recommended[9]
Typical enterprise rack design (illustrative)	~10 kW	Roughly four times the draw; plan for additional power and cooling work
Blackwell rack-scale (NVL72)	~120 kW	Liquid cooling is required for this rack-scale design[10]

Table 5.3 · Bring these four numbers to the facilities conversation before the procurement conversation. Lead times are measured in quarters.

FACILITIES QUESTIONS TO ANSWER BEFORE PROCUREMENT

Each is a question a data-centre manager can answer in a meeting, and each has killed an otherwise sound proposal at a later and more expensive stage.

- What is the per-rack power ceiling today, and what would raising it require: PDUs, busway, a utility conversation, or a new hall?
- Is the cooling design air or liquid, and what inlet temperature and delta-T can it actually hold at full load?
- What is the floor loading limit? Dense GPU racks are heavy enough that this is a genuine constraint rather than a formality.
- What is the UPS and generator headroom, and does the AI load fall inside or outside protected power?
- What lead time applies to power distribution work, and does it sit inside or outside the project timeline you have committed to?
- If the next hardware generation needs liquid cooling, what is the upgrade path for this room?

Capacity planning under a fixed ceiling

Owned capacity does not burst. Decide in advance what happens in a contended hour and encode it. If whichever team submits first wins, regulated workloads receive no dependable priority control.

- **Priority classes** per workload, agreed with the business rather than assigned by the platform team.
- **Preemption rules** that state which jobs can be evicted and with how much notice.
- **An overflow path** to rented capacity, with the data classification rules that govern whether a given workload may use it.
- **A queue with a published wait time**, so that degradation is visible rather than experienced as an outage.

Figure 5.5 · A rack-scale reference point: a single scalable unit makes accelerator density, power distribution and cooling planning physical design inputs. Source: NVIDIA, DGX SuperPOD Reference Architecture, Figure 2. `https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-b200/latest/dgx-superpod-architecture.html`

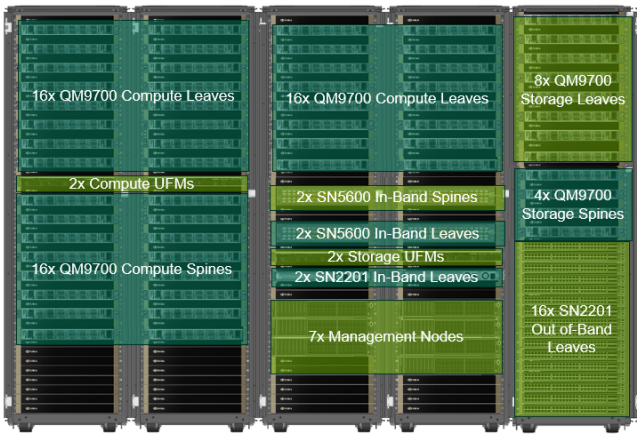


Figure 5.6 · The supporting racks are part of the system: management nodes, storage, switches and control appliances sit alongside compute. Source: NVIDIA, DGX SuperPOD Reference Architecture, Figure 3. `https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-b200/latest/dgx-superpod-architecture.html`

SIZING CASE REQUIREMENTS

Size from context and concurrency rather than parameter count. Apply an explicit agentic multiplier and say where it came from. Confirm power and cooling before procurement. Make the own-versus-rent case on a measured duty cycle. Those four checks make the sizing case auditable, whatever conclusion it reaches.

SECTION 06

Network isolation and egress control

Isolation has three postures. Conflating them weakens the review. Pick one deliberately, state its trade-off, and be able to demonstrate the control.

POSTURE	DEFINITION	BUYS YOU	COSTS YOU
Air-gapped	No inbound or outbound network path. No telemetry, no external model calls, no live package installs. Updates arrive by reviewed, manual import.	Air-gapped operation provides the strongest boundary claim and may be required for some IL6-class or mission-specific environments, subject to authorization. It removes an entire category of exfiltration risk.	Patching becomes a standing operational programme. Every dependency must be mirrored in advance. Highest running cost by a wide margin.
Isolated with brokered egress	Default-deny outbound, with a short allowlist through an inspected proxy: identity provider, artifact mirror, mail gateway.	The practical middle ground for most regulated enterprises. Keeps patching and model updates viable while egress stays enumerable.	The allowlist is a living control that decays without ownership. Requires enforcement that workloads cannot bypass.
Private link	Traffic avoids the public internet via private connectivity to a provider.	Removes public exposure and simplifies some network assurance.	**Does not change legal jurisdiction or compellability.** If the sovereignty requirement is legal rather than technical, this posture does not satisfy it, a distinction reviewers in defence and EU-regulated sectors will make immediately.

Table 6.1 · Under NIST SP 800-53 these are expressions of AC-4 (information flow enforcement) and SC-7 (boundary protection). Naming the controls helps the conversation with your security architect.

> **DEMONSTRATE DISCONNECTED OPERATION**
>
> Fully disconnected operation is frequently requested by federal contractors and defence suppliers, and vendors in this market often cannot answer immediately. Treat it as a test with a pass condition. Can the environment install, upgrade, and recover with the external link physically removed for thirty days? Is there a documented import process with signature verification? Does any component phone home for licensing? Ask for a demonstration in a disconnected environment, and record the answer either way.

Enforcing egress so it cannot be bypassed

The pattern that holds up: route all outbound traffic through a dedicated egress gateway, register the permitted destinations, and refuse everything else. In a service mesh this is registry-only mode with an egress gateway. The mesh sidecar **must be backed by network policy or node-level firewall rules, so a workload cannot simply route around it.**[12] A proxy that a pod can bypass is not a control.

For disconnected clusters the corresponding requirement is a private OCI registry inside the boundary that mirrors everything the environment needs: base images, Helm charts, OS packages, Python wheels, and model weights, populated by a reviewed sync process. Model weights are frequently forgotten in mirroring plans and are among the largest artifacts you will move.

> **TOOL CALLS ARE EGRESS PATHS**
>
> Every tool you grant an agent is a potential egress channel, and tool ecosystems such as MCP have become a significant new egress surface. Exfiltration via poisoned tool descriptions, where instructions embedded in a tool's own metadata redirect data outward, is documented rather than theoretical.[13]
>
> The controls are unglamorous and effective: block outbound by default from tool-server processes; allowlist explicit endpoints per tool; run each tool server in its own sandbox; filter tool calls at the broker; and alert on any outbound connection from a tool process that is not on the list. Treat a new tool as a firewall change, because that is what it is.

What to be able to demonstrate

- An enumerated list of every permitted egress destination, with a named owner and a business justification per entry.
- Evidence that default-deny is real: a packet capture or a denied-connection log from an attempt to reach an unlisted host.
- Proof that workloads cannot bypass the proxy, at the network layer rather than by convention.
- A review cadence for the allowlist, and the date it last shrank.
- The fixed egress source addresses, so counterparties can allowlist you in return.

SECTION 07

Identity, access and audit

Three populations need identity in this environment, and most designs handle only the first well.

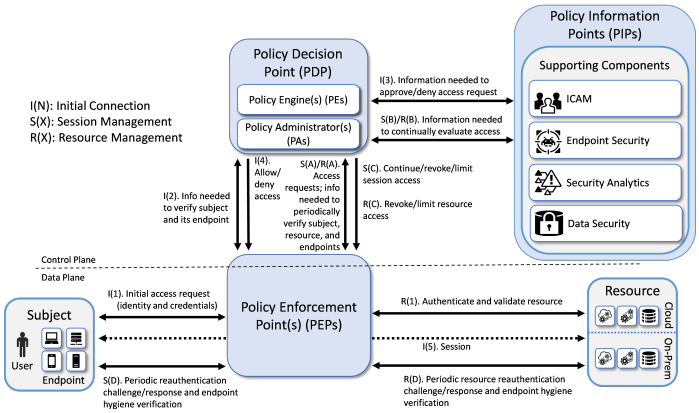


Figure 7.1 · General zero-trust architecture. The policy decision point combines policy engines and administrators; enforcement points sit between subjects and resources while identity, endpoint, analytics and data-security signals inform the decision. Source: National Institute of Standards and Technology, Implementing a Zero Trust Architecture, Figure 1. `https://pages.nist.gov/zero-trust-architecture/VolumeB/architecture.html`

1 · HUMANS

Federated to the enterprise IdP by SAML or OIDC. **No local accounts on the AI platform.** The moment there is a second user directory, joiner-mover-leaver breaks and your access review has a blind spot. Group membership from the IdP drives role assignment.

2 · WORKLOADS

Services authenticate by attested, cryptographically verifiable identity with short-lived credentials rather than long-lived secrets. SPIFFE and SPIRE are the de facto open standard, and bridge to enterprise authorisation via OAuth 2.0 token exchange.[14]

3 · AGENTS

Agent identity requires a separate model. An agent acts *on behalf of* a user but is not that user, and must never hold more authority than the person who invoked it. Standards work is live but **not ratified**, so treat vendor claims of compliance accordingly.[15]

The rule that keeps agent authority bounded

An agent's effective permissions should be the *intersection* of its own scoped grant and the permissions of the invoking human, never the union, and never a standing service account that exceeds both. This single rule bounds agent authority and preserves attribution. An agent's capability also shrinks automatically when the invoking user changes role.

The SSO and identity integration matrix

Protocol mismatch discovered during integration is a common and avoidable delay. A platform that supports SAML but not OIDC is a normal situation and usually workable, but it needs to be known before the security review rather than during it, because the fallback is often local accounts and that breaks the first rule above. Ask any vendor to complete this table with a version number and a date, and treat blanks as blanks. Confirm the answer with each vendor and for the selected deployment.

CAPABILITY	WHAT TO REQUIRE	IF IT IS ABSENT	ASK FOR
SAML 2.0	SP-initiated and IdP-initiated flows, signed assertions, metadata exchange, certificate rotation without downtime.	Rare. If absent, most enterprise IdPs cannot integrate at all.	A working integration against your own IdP in a trial tenant.
OIDC	Authorisation code flow with PKCE, discovery document, refresh handling, standard claims.	Common gap. The fallback is often SAML, which is acceptable, or local passwords, which is not.	A stated position: supported, roadmap with a date, or not planned.
Group and role mapping	IdP group claims map to platform roles without manual assignment, and unmapped groups get no access.	Every role change becomes a ticket, and access reviews stop being reliable.	The mapping configuration, and what happens to a user removed from a group.

CAPABILITY	WHAT TO REQUIRE	IF IT IS ABSENT	ASK FOR
Provisioning and deprovisioning	SCIM or equivalent, so that a leaver loses access within the standard SLA.	Access deletion timeliness is a common audit test.	The deprovisioning path and its measured latency.
Federation broker	Where a broker such as Keycloak sits between the platform and the IdP, its own hardening, upgrade and availability story.	A single point of failure for all access, often unowned.	Who operates the broker, and what happens when it is down.
MFA	Inherited from the IdP rather than implemented separately, with step-up for administrative actions.	Two MFA systems, one of which is weaker.	Confirmation that the platform cannot bypass IdP MFA.
Multi-region behaviour	Whether role assignments and workload identities persist across regions or must be duplicated.	Access works in one region and silently fails in another.	A tested answer rather than an assumed one.
Break-glass	An emergency access path with its own logging, approval and a mandatory review after use.	An undocumented admin account that nobody reviews.	The procedure, and the log of the last time it was exercised.

Table 7.1 · Complete this before the security questionnaire arrives. A blank row is an answer and it is better found now.

SEPARATE TELEMETRY AND AUDIT RECORDS

Operational telemetry and the governance audit record are different systems with different consumers, retention periods and integrity requirements. Metrics and traces exist to keep the service up and can be sampled, mutable and short-lived. The audit record exists to answer a regulator, and must be complete, append-only and retained on schedule. Building one and calling it both is a finding waiting to happen.

The audit record, field by field

These are the minimum fields an auditor needs to reconstruct an incident. Retention should be set by the longest obligation that applies to you, which is usually a sector rule rather than an AI-specific one.

FIELD GROUP	CONTENTS	WHY AN AUDITOR ASKS
`identity`	Invoking principal (human or agent), delegation chain, session and trace ID, source application.	"Who did this, and under whose authority?" The delegation chain makes agent actions attributable to a person.
`model`	Model name, version, precision, endpoint, and the version hash of the system prompt and configuration.	Reproducing a decision requires the exact configuration. A changed system prompt is a changed system.
`content`	Prompt and output, **hashed or redacted rather than verbatim** where regulated data is involved, plus data classification of inputs and outputs.	Demonstrates what was processed without the log itself becoming a regulated data store. See §8.
`retrieval`	Document and record identifiers returned, the ACL decision applied, and the source system for each.	The only way to answer "could this user have seen that document?" after the fact.
`actions`	Each tool invoked, argument hash, result status, and any external endpoint contacted.	Agent actions have real-world effects. This records their potential impact.
`policy`	Every allow and deny decision with the rule ID that produced it, and any human approval, override or escalation.	Proves the control ran, rather than merely that it was configured.
`cost`	Input and output token counts, attributed to a team or cost centre.	Not a compliance field, but it ends the "who is spending this" argument and is the earliest signal of a runaway agent. See §2.
`integrity`	UTC timestamp, monotonic sequence, hash of the preceding record, written to append-only or object-locked storage.	Distinguishes a log from an audit trail. If an administrator can silently edit it, it is not evidence.

Table 7.2 · A practical audit schema. The integrity row is the one most often missing, and the one that determines whether the rest is worth anything.

AUDIT EXPORT REQUIREMENTS

A recurring question in regulated sectors is whether the system can generate a report of what was used, by whom, and under which policy, on demand and without an engineering project. Logging the fields above is necessary and not sufficient. Ask for the export: which format, which date ranges, whether it can be filtered by user, model and policy decision, how long it takes, and whether the evidence package is reproducible. If the answer is that a report could be built, that is a roadmap item rather than a capability, and it should be recorded as one.

Access review, the control auditors test

SOC 2 testing tends to concentrate on user access: who was granted access, who approved it, when a leaver was removed, and whether the evidence exists. Build for that specifically. Every access grant should have an approver recorded, every removal should have a timestamp, and a quarterly review should produce an artifact rather than a conversation. The platform's contribution is an export that a GRC team can reconcile against the HR system without manual work.

SECTION 08

Where regulated data comes to rest

Teams try to draw this boundary around the model. That is the wrong object. The model is a processor. Classification travels with the *data*, from the system of record through retrieval into the prompt, and onward into a series of places that rarely appear on the architecture diagram.

Regulated data comes to rest in at least twelve locations in a working agentic system. Eight are obvious. Four are frequently overlooked.

#	WHERE DATA LANDS	WHAT IS ACTUALLY THERE	CONTROL
1	Prompt / user input	Whatever the user typed or pasted, verbatim.	Input classification; gateway policy
2	Retrieved context	Source-system records injected into the prompt by RAG.	ACL-preserving retrieval (below)
3	**Embeddings & vector index** *Frequently overlooked*	Not automatically anonymised. Research demonstrates substantial reconstruction of original text in the evaluated setting.[17] **Vectors derived from PHI should be treated as potentially regulated data** unless a documented de-identification determination establishes otherwise.	Treat the vector store as a regulated data store, inside the boundary
4	**KV cache in GPU memory** *Frequently overlooked*	Holds derived key/value representations of prompt content for the life of the sequence, and may persist longer where prefix caching is enabled across requests.	Tenancy isolation; disable cross-tenant prefix reuse; wipe on release
5	Conversation / session history	Accumulated context persisted between turns.	Retention policy; encryption at rest
6	**Gateway request/response logs** *Frequently overlooked*	Your audit control can become the largest unclassified copy of regulated data in the estate.	Hash or redact at capture, never after. See §7
7	**Traces & vendor telemetry** *Frequently overlooked*	APM and LLM-observability tooling frequently captures prompt bodies by default, and frequently ships them to a third-party SaaS endpoint, crossing the boundary invisibly.	Audit every agent and SDK for outbound telemetry; default off
8	Agent scratchpad	Intermediate reasoning and accumulated tool output.	Same classification as its inputs; ephemeral storage
9	Tool call arguments	Data passed to a tool, and onward to wherever that tool sends it.	Tool-level egress control. See §6
10	Evaluation datasets	Real records copied into test fixtures, typically outside change control.	De-identify or hold under production controls
11	Fine-tuned weights	Training data can be memorised and extracted from the resulting model.	Classify the weights at the level of the training data
12	Backups & snapshots	DR copies of every store above, often with different retention and weaker access control.	Include in the classification scope, not just the primaries

Table 8.1 · Data-at-rest inventory for an agentic system. A boundary claim is only as strong as its treatment of rows 3, 4, 6 and 7.

DEFINE THE ISOLATION BOUNDARY

Where several teams, business units or customers share one environment, ask exactly where the isolation boundary sits: separate namespaces, separate clusters, separate node pools, or separate GPUs. Rows 3, 4 and 6 above are where shared-tenancy assumptions usually fail, because a vector index, a KV cache and a log store are all easy to share by accident. Require a written isolation model with a stated threat it defends against. A vendor that cannot describe its isolation boundary in one paragraph has not decided on one.

Retrieval permissions

Enterprise RAG most often fails when a correctly secured document is returned to someone who could not have opened it in the source system, because the indexing service read everything with a privileged account and the permissions were never carried into the index.

Filtering after retrieval is not a control. Retrieving the top *k* results and then removing the ones the user cannot see is a thin application-layer defence that a logic bug or a prompt injection defeats, and it leaks through result counts and latency even when it works. Permissions must be enforced *in the retrieval layer*, as part of the query, so that unauthorised content is never a candidate.[18]

Determine whether an inaccessible document is evaluated as a candidate and then filtered out, or is never a candidate at all. The two answers are architecturally different. The query should show which one is implemented; query-time enforcement should be the primary authorization control, because post-filtering alone is insufficient against an injection.

De-identification limits

Under HIPAA the two established routes remain those at 45 CFR §164.514(b): Safe Harbor, meaning removal of the enumerated identifier categories, or Expert Determination. Confirm this interpretation against current HHS guidance.[27] Two cautions in an AI context:

- **Free text defeats naive redaction.** Clinical notes and ticket bodies carry identifiers in unstructured prose. Pattern-based scrubbing has a real error rate, and one miss is a disclosure.
- **De-identification is not idempotent across a corpus.** Individually de-identified records can be re-identified in combination, which is what a retrieval system is built to do.

Backup, restore and the recovery objectives

Disaster recovery here covers more stores than teams expect, and it is another question vendors are commonly asked and commonly defer. Set your own objectives first, then test any supplier against them.

WHAT HAS TO BE RECOVERABLE	WHY IT IS EASY TO MISS	YOUR RPO	YOUR RTO	TESTED HOW
Model weights and adapters	Large, slow, sometimes assumed re-downloadable when the link is closed.			Restore into a clean cluster.
Vector index	Rebuildable in principle, but the rebuild takes hours and re-reads regulated source data.			Time a full rebuild once.
Audit history	Retention is contractual. Losing it is a finding.			Restore and verify the hash chain.
Configuration and policy	Often lives only in the cluster.			Rebuild from source control alone.
Secrets and keys	A restore that cannot decrypt is not a restore.			Exercise key recovery.

Table 8.2 · An untested restore is a plan, not a capability.

On business associate agreements

A covered entity that self-hosts an open-weight model on infrastructure it controls does not need a BAA *for the model*, because no third party receives PHI. The Privacy, Security and Breach Notification Rules continue to apply in full, and a BAA is still required wherever an external party does touch PHI, which in practice includes managed services, remote support and telemetry vendors. Confirm with counsel before relying on this. This follows from the definition of a business associate rather than from guidance addressing self-hosted language models specifically.

BOUNDARY CHECKS

1. Is the vector store classified at the level of its source data?
2. Can prompt content survive in KV cache across tenants or requests?
3. Does any log, trace or telemetry stream carry regulated content past the boundary, including to a vendor SaaS endpoint?
4. Does retrieval enforce source-system permissions in the query itself?
5. Are backups and evaluation datasets inside the classification scope?

SECTION 09

SOC 2, HIPAA, and everything else

Two frameworks come up in nearly every enterprise conversation: SOC 2 and HIPAA. ITAR appears in defence and aerospace. The rest of the alphabet is real but conditional. Prioritise the frameworks that apply to the workload and have an accountable owner. This is architectural mapping, not legal advice.

SOC 2

SOC 2 is often treated as table stakes. It is an attestation about a service organisation's controls against the AICPA Trust Services Criteria,[26] and the practical questions are about scope and period rather than about whether a report exists. Confirm the mapping against the current AICPA criteria.

WHAT TO ASK FOR	WHY IT MATTERS	WHAT A WEAK ANSWER LOOKS LIKE
Type II, not Type I	Type I describes control design at a point in time. Type II tests operating effectiveness over a period, usually six or twelve months.	A Type I report offered as though it were equivalent.
The period, and a bridge letter	A report ending eight months ago covers nothing recent. A bridge letter covers the gap to today.	An in-date report with no bridge letter and no next audit date.
The scope	Which systems, which criteria (Security is mandatory; Availability, Confidentiality, Processing Integrity and Privacy are optional), and which entities.	A report scoped to a corporate SaaS product while the deployed component sits outside it.
Subservice organisations	Carve-out or inclusive method. Carve-out means their cloud provider's controls are excluded and you inherit that assessment work.	Carve-out without a list of the subservice organisations.
Complementary user entity controls	The controls the report assumes *you* operate. This is the shared responsibility matrix in another form, and it is where unassigned obligations hide.	Nobody on either side has read that section.
Exceptions and management response	Every real report has some. Their nature and the remediation plan is the signal.	A report with no exceptions and a very short testing narrative.

Mapped to this architecture, the criteria that carry the most weight are the CC6 series (logical and physical access), CC7 (operations, monitoring and incident response) and the Confidentiality criteria. In practice the tested controls are user access provisioning and removal, change management, monitoring and alerting, and vulnerability management. Access deletion timeliness is a frequent test, which is why §7 treats deprovisioning as an architectural property rather than a process detail.

> **EFFECT OF SELF-HOSTING ON SOC 2 SCOPE**
>
> When the software runs inside your own account or facility, a large part of the control environment is yours, and the vendor's report covers less of the surface than either party assumes. That is an advantage for you, because the controls you already have attested apply. It also means the shared responsibility matrix in §10 is the document that determines what the report needs to cover. Produce that first and the SOC 2 conversation gets much shorter.

HIPAA

The second framework that reliably appears, usually as a direct question about whether PHI can be processed and what happens to it. The architectural answer is stronger than the paperwork answer: if no third party receives PHI, the surface for a business associate relationship is much smaller. See §8 for the BAA reasoning and its caveats.

SECURITY RULE SAFEGUARD	WHAT IT ASKS OF THIS ARCHITECTURE	WHERE IT IS ENFORCED	EVIDENCE ARTIFACT
Access control §164.312(a)	Unique user identification, emergency access, automatic logoff, encryption and decryption.	§7 identity, no local accounts, break-glass procedure.	Access review export; break-glass log.
Audit controls §164.312(b)	Hardware, software and procedural mechanisms that record and examine activity.	§7 audit record, including the integrity fields.	Audit-log integrity attestation; evidence export.
Integrity §164.312(c)	PHI is not improperly altered or destroyed.	Append-only audit storage; backup and restore testing per §8.	Restore test records; hash-chain verification.
Person or entity authentication §164.312(d)	Verify that a person seeking access is who they claim to be.	Enterprise IdP federation with MFA inherited, per Table 7.1.	MFA coverage report.
Transmission security §164.312(e)	Guard against unauthorised access to PHI in transit.	TLS at ingress, mTLS between services, egress default-deny per §6.	Network topology diagram; egress allowlist.
Proposed update *Status in motion; confirm before relying on it*	A 2025 proposed rule would remove the "addressable" category and mandate encryption, MFA, network segmentation and regular testing. **Not final**, and final action has slipped.[20]	Already satisfied by §4 segmentation and §7 identity if you build to this reference.	Segmentation diagram; MFA coverage report.

Table 9.1 · Architectural mapping. Confirm the proposed rule's status at the time of your review.

The rows in Table 8.1 are where HIPAA scope is usually underestimated. Vectors derived from PHI should be treated as potentially regulated data unless a documented de-identification determination establishes otherwise. Gateway logs holding verbatim prompts are a PHI store. Backups of either are in scope. A covered entity that has classified those three correctly is in a much better position than one that has a BAA and has not.

> **EFFECT ON HIPAA SCOPE**
>
> A covered entity that keeps PHI inside its own boundary reduces the number of parties it has to place under agreement, and reduces the number of places it has to attest to. It does not reduce the obligations themselves. The Privacy, Security and Breach Notification Rules apply in full either way.

ITAR, and the defence case

Narrow but strict where it applies. Under ITAR, releasing or otherwise transferring covered technical data to a foreign person in the United States can be a deemed export.[28] Design the system so foreign persons cannot access the controlled data or its administrative paths unless export counsel confirms an applicable authorization or exemption.

- **Personnel, not just infrastructure.** Administrative access by a vendor support engineer who is not a US person is the failure mode, which makes the management plane in Figure 4.3 a compliance boundary rather than an operational detail.
- **Review the control plane.** A managed service whose operator can reach the environment may require authorization, personnel controls and additional contractual safeguards, regardless of where the data physically sits.
- **Practical consequence.** A disconnected deployment, US-person-only operations and a documented import process may be required, depending on the data and authorization.

Conditional frameworks

These come up occasionally and matter enormously when they do. Confirm status directly before citing any of them, because several change on a monthly cadence.

FRAMEWORK	WHEN IT APPLIES	WHAT IT ASKS OF THIS ARCHITECTURE	EVIDENCE
NIST SP 800-53	US federal-adjacent reviews; a common vocabulary elsewhere.	Egress and boundary work maps to AC-4 and SC-7. NIST is developing control overlays for securing AI systems, including generative and agentic cases.[24]	Control matrix aligned to the overlay.
FedRAMP	Selling to or through US federal agencies.	The 20x programme uses phased adoption; confirm its current phase before relying on it.[21] The GenAI-specific prioritisation framework was removed in January 2025, so there is no GenAI fast lane.[22]	Applies to a service you consume or offer, not to a purely internal deployment.
DoD Impact Levels	Defence workloads.	IL4 covers CUI. IL5 adds control enhancements, US-controlled facilities, US-person access and separation from non-government tenants. IL6 is classified to SECRET.[23]	Personnel and facility attestations; tenancy separation evidence.
FIPS 140-3	US government cryptography requirements.	Cryptographic modules must be validated. Confirm current CMVP status for every module in the path.[29] In practice this constrains which libraries and images can be used, and the engineering work to comply is itself the obstacle.	CMVP certificate numbers for every module in the path.
ISO/IEC 42001	Organisations wanting a certifiable AI management system.	Process and governance rather than technical controls. Wraps the environment and depends on §7 evidence to demonstrate operation.	A certificate, which is useful when a reviewer wants an artifact rather than an assurance.
EU AI Act	EU deployment or EU-facing systems.	Article 12 requires automatic event logging and Article 19 requires retention of at least six months.[16] A 2026 Omnibus agreement proposed deferring some high-risk obligations; Article 50 transparency obligations were not part of that proposed deferral.[25] Confirm adoption and applicable dates before relying on this summary.	Log retention policy and its demonstration.
GDPR	EU personal data.	Lawful basis, data minimisation, subject access and erasure. Erasure is the hard one here: rows 3, 4, 6 and 11 of Table 8.1 all hold derived copies.	A tested deletion path that reaches the vector index and the logs.
SOC 1 / ICFR	Where the system touches financial reporting.	Controls over financial reporting rather than security. Rare for AI platforms, and it appears when the workload feeds a reported number.	Scope statement showing whether the platform is in or out.

Table 9.2 · Confirm every changing status against the primary source at the time of your review. Several of these changed within the last twelve months.

SECTION 10

The security review pack

Security review stalls on missing documents far more often than on failed controls. The items below are commonly requested during security review. Prepare them in advance so the review can focus on the architecture.

ARTIFACT	WHAT IT HAS TO CONTAIN	PRIMARY REVIEWER	COST OF NOT HAVING IT
Shared responsibility matrix	Every control domain, with the party accountable for operating it and the party accountable when it fails. Template below.	Security architect, GRC	The most damaging gap in this list. Without it, every other answer is provisional.
Data flow diagram	Every hop, with classification and control. Figure 4.2.	Security lead, privacy	Frequently the literal gate for the next meeting.
Network topology	Zones, load balancer types, TLS termination, certificate authority, egress addresses. Figure 4.3.	Network and security engineering	Answered live with “I will have to check”, which costs a round trip.
Component inventory	Every deployed application and service, with version, exposure and authentication. Template below.	Pen test scoping	The test cannot be scoped, quoted or booked.
Identity matrix	Protocols, group mapping, provisioning, MFA. Table 7.1.	IAM team	Protocol mismatch discovered during integration, with a fallback that weakens access control.
Vulnerability handling	Scanning cadence, who scans, how results are shared, remediation SLA by severity, exception process and window.	Security operations	Regulated buyers have fixed exception windows and will not proceed without a stated SLA.
SBOM and licences	Components, versions, licences, and which are commercially licensed rather than open source.	Legal, supply chain	An open-source component that needed a commercial licence, found late.
Update and patch path	How container images and models are updated, through which registry, on what cadence, and how the customer is notified.	Platform and change management	Change advisory board cannot approve what it cannot describe.
Incident and breach terms	Detection, notification timeline, forensic support, and who communicates to regulators.	Legal, GRC	Negotiated during an incident, which is the worst possible time.
DR and recovery objectives	RPO and RTO per store, and the date of the last tested restore. Table 8.2.	Infrastructure, GRC	An untested plan presented as a capability.

Table 10.1 · The review pack. Assign an owner and a date to each row at the start of the evaluation rather than when it is requested.

> **A TYPICAL EVIDENCE REQUEST**
>
> *“Is there a shared responsibility matrix documented, and is that something that can be shared?”* Security architect, a large retailer. The question is routine and the honest answer from many vendors is that the document is different for every customer and has to be drafted. That is true and it is still a delay. A generic matrix that is then marked up per deployment costs days rather than weeks.

Shared responsibility matrix, template

Use this allocation as a baseline for the deployment. Responsibilities change depending on whether the platform runs in the customer's cloud account, the customer's data centre, or a vendor-managed environment. Confirm each row in writing before the security questionnaire, and treat every row marked shared as a row that needs a named person on each side.

CONTROL DOMAIN	TYPICALLY VENDOR	TYPICALLY CUSTOMER	THE QUESTION THAT SETTLES THE ROW
Physical and facility security	None, when self-hosted	All	Whose building, whose badge system, whose attestation?
Cloud account or hypervisor	None, when self-hosted	All	Who holds the account root credentials and the billing relationship?
Kubernetes control plane	Shared	Shared	Who upgrades it, on whose schedule, and who approves the window?
Node OS patching	Shared	Shared	Who patches the host, and what is the SLA by CVE severity?
Platform component images	Builds and signs	Approves and deploys	Who scans images, who remediates a finding, and within how long?
Registry and update path	Publishes	Mirrors and admits	Through which registry, and what happens when egress is closed?
Network policy and firewall	Specifies	Implements	Who writes the rules and who can change them without review?
Ingress, TLS and certificates	Specifies	Issues and rotates	Which CA, what rotation period, and who is paged when a cert expires?
Egress allowlist	Declares what it needs	Approves and enforces	Who owns each entry, and when was the list last reduced?
Identity provider integration	Supports protocols	Operates the IdP	Which protocols are actually implemented today? Table 7.1.
User lifecycle and access review	Provides the mechanism	Operates the process	Who evidences that a leaver lost access, and in what form?
Secrets custody and rotation	Provides the mechanism	Holds the keys	Can the vendor read a customer secret? The answer must be written down.
Data classification and retention	Provides controls	Sets policy	Who decides retention for each of the twelve stores in Table 8.1?
Backup and restore	Documents the method	Executes and tests	Who runs the restore test, how often, and where is the evidence?
Audit retention and export	Emits records	Retains and produces	Who produces the evidence package when a regulator asks?
Model selection and approval	Makes available	Approves for use	Who approves a model or precision change, and against what gate?
Agent tools and approval gates	Provides the broker	Defines the allowlist	Who approves a new tool, and is it treated as a firewall change?
Vulnerability scanning	Scans own components	Scans the environment	Are results shared in both directions, and on what cadence?
Penetration testing	Supports scoping	Commissions and remediates	Who pays, who fixes, and who re-tests?
Incident detection and response	Notifies	Declares and responds	What is the notification clock, and what triggers it?
Regulatory attestations	Own attestations	Own attestations	Which report covers which components? See §9 on scope.

Table 10.2 · A generic allocation for a self-hosted deployment. Use it as a baseline, then confirm each row for the actual environment. The right-hand column is the part that does the work.

Component inventory, template

A penetration test cannot be scoped without this, and the request is usually phrased as "we need a list of applications." The categories below are stable across deployments. The instances are not, so produce it against your own build.

CATEGORY	WHAT TO LIST	EXPOSURE	AUTHENTICATION	IN PEN TEST SCOPE
Ingress and gateway	Load balancers, WAF, ingress controller, model access gateway.	Internal or external	SSO, mTLS	Yes, always
Web and API surfaces	Every console, notebook interface, API endpoint, webhook receiver.	Internal	SSO	Yes
Agent and tool servers	Each tool server, its runtime, and its outbound destinations.	Internal only	Workload identity	Yes
Inference services	Serving engines, model endpoints, embedding services.	Internal only	Workload identity	Yes
Data services	Vector store, object storage, databases, caches, message queues.	Internal only	Workload identity	Yes
Platform services	Registry, secrets manager, policy engine, audit sink, scheduler.	Internal only	Workload identity	Yes
Observability	Metrics, logging, tracing, dashboards, any agent that ships data.	Internal	SSO	Yes, and check outbound telemetry
Management plane	Cluster API, bastion or just-in-time access, CI/CD runners.	Restricted	MFA, short-lived	Yes, usually the highest-value target
Third-party components	Anything your own teams deploy on the platform.	Varies	Varies	Decide explicitly, not by omission

Table 10.3 · Add version, image digest and owning team per instance. That version doubles as the SBOM index.

Questions for vendors

Each of these is asked routinely in security reviews and deferred often enough to be worth preparing for. Ask them in writing, early, and record the answer including "not yet". A supplier that says not yet with a date is more useful than one that says yes without evidence.

ARCHITECTURE

- How is the platform structured inside our cluster: which namespaces, node pools and privileged workloads?
- What is the registry update path, and does it work with egress closed?
- Can this run fully disconnected for thirty days, demonstrably?
- How is multi-tenancy isolation enforced, against which threat?
- What is the exact ingress topology: load balancer type, TLS termination, certificate authority?

OPERATIONS

- What is the update cadence, how are we notified, how do we roll back?
- What is the DR story, with RPO and RTO per store, and when was the last tested restore?
- What does the observability surface show: traffic, latency, errors, per-team consumption?

EVIDENCE

- Can the system produce a compliance report on demand, filtered by user, model and policy decision?
- Which SOC 2 report covers which components, and what are the complementary user entity controls?

PERFORMANCE AND COST

- What throughput and time-to-first-token, at what concurrency, precision and token split?
- What does this cost at our scale, and which inputs is that most sensitive to?

ACCESS

- Can any vendor employee reach our environment, under what conditions, and is that logged where we can see it?
- Which identity protocols are implemented today, as opposed to planned?

> **INTERPRETING VENDOR ANSWERS**
>
> Every vendor in this market has gaps. What distinguishes them is whether the gap is known, dated and disclosed, or discovered by you in month four. Write the answers down, date them, and re-ask the ones that mattered before you sign.

SECTION 11

Vendor continuity and exit

If the supplier is young, expect the question. It is raised directly and often, in terms like *"our organisation has a hesitancy for startups, we've had a lot of startups burn us in the past."* It is a reasonable position and it is answerable with contract terms and architecture rather than with reassurance.

Supplier-risk categories

The concern usually falls into four categories: the supplier disappears; the supplier is acquired and the product changes; support quality falls below what the workload needs; or the organisation becomes unable to operate the system without the supplier. Each has a different mitigation, and conflating them produces a vague answer.

RISK	WHAT TO REQUIRE	EVIDENCE TO ASK FOR	ARCHITECTURAL MITIGATION
Supplier ceases trading	Source code escrow with a defined release condition, plus the build instructions and dependency manifests needed to actually use it. Confirm the escrow terms with legal before relying on them.	The escrow agreement, the deposit date, and confirmation that the deposit has been verified rather than merely lodged.	Open formats and standard substrate, so the environment keeps running without the supplier. See §12.
Acquisition or pivot	Change of control terms, a minimum support period, and price protection for the committed term.	The clause. Most standard agreements do not contain it unless asked.	No proprietary data formats, so your data is portable on day one.
Support quality	Response and resolution SLAs by severity, named escalation path, and a named account contact.	The last quarter's actual response times, not the target.	Runbooks in your possession, so routine operations do not require a ticket.
Operational dependence	Your team must be able to run, upgrade and recover the environment without vendor involvement.	A documented handover, and a rebuild performed by your staff.	The exit test below.
Financial standing	Whatever your procurement process normally requires from a supplier of this size.	Funding position, customer count, and insurance certificates.	Phase the commitment so that exposure grows with confidence.
Key person	Knowledge held by more than one person on both sides.	Named secondary contacts, and documentation quality you can inspect.	Internal ownership of the runbook.

Table 11.1 · Escrow alone is weak. Escrow plus a verified deposit plus a demonstrated rebuild is a control.

THE EXIT TEST

Run this test: *Can your own team rebuild this environment from source control, your own registry and your own mirrored artifacts, with no call to the vendor and no external network access?* Schedule it once during the first year, before you need it. It tests continuity, disaster recovery, the artifact supply chain and the disconnected posture in a single exercise, and it produces an artifact procurement will accept.

Phasing the commitment

Supplier risk is usually managed structurally rather than through contract language alone. Use a short initial term with a defined success measure, a second phase that expands scope only after the exit test has been run, and a multi-year commitment only once the operational dependence question has an evidenced answer. This also matches the budget thresholds in §2: smaller phases limit initial exposure while the evidence accumulates.

REFERENCE CHECKS

Expect to be asked for references in your own sector and region, and expect to be asked for them by the procurement team rather than the technical one. If a supplier cannot provide a comparable reference, that is worth knowing early, and it is a fair thing to ask for in writing at the start of an evaluation rather than at the end.

SECTION 12

From a cloud account to your own facility

Portability keeps the location of the workload as a deployment decision rather than an architectural one. It also lets the organisation defer that choice while the evidence is gathered.

Figure 12.1 · Cooling towers are part of the deployment surface when inference moves into owned facilities; power, heat rejection and maintenance become architecture inputs. Source: U.S. Department of Energy, Cooling Water Efficiency Opportunities for Federal Data Centers. `https://www.energy.gov/cmei/femp/cooling-water-efficiency-opportunities-federal-data-centers`

The portability contract

All five are required for a move to remain a scheduling exercise. Losing any one creates a rewrite.

LAYER	COMMIT TO	BECAUSE
Orchestration	Kubernetes or equivalent, with no provider-specific control plane	The workload definition moves unchanged. Only the substrate differs.
Packaging	OCI images and open model formats (safetensors, GGUF)	Artifacts stay importable into an environment with no internet access.
Storage	S3-compatible object storage	Data-access code is unchanged between a cloud bucket and an on-prem object store. Verify the edges, below.
Inference API	An OpenAI-compatible endpoint in front of every model	Applications stop caring whether the model is hosted or self-hosted.
Identity	Standard federation and workload identity, never cloud-native IAM roles	Cloud IAM has no on-premises analogue. The most common hard-stop in a migration.

Migration failure points

MANAGED SERVICES WITH NO EQUIVALENT

Proprietary databases, managed vector search and hosted model services have no drop-in on-prem counterpart. Each is a rewrite, and they surface late because they were adopted for convenience.

INSTANCE-METADATA CREDENTIALS

Cloud IAM roles attached to compute do not exist on your own hardware. Attested workload identity is the substitute, and retrofitting it touches every service.[14]

OBJECT STORAGE AT THE EDGES

Core S3 operations port cleanly. The periphery does not. Check lifecycle policies, object lock, versioning and replication early, object lock especially, since audit retention may depend on it.

ELASTICITY

On-prem capacity is fixed. Bursts queue, degrade by policy, or overflow to rented capacity, and the choice has to be encoded as priority classes before the first contended week.

A staged path

STAGE	DO	EXIT CRITERION
1 · Contain	Deploy inside your own cloud account with the boundary, gateway, identity and audit of §4 in place.	Egress is enumerated and every model request attributed. Most of the governance value is realised here.
2 · Decouple	Adopt the portability contract. Remove provider-specific dependencies one at a time, starting with identity.	The stack builds and runs in a second environment with no code changes. Demonstrate it.
3 · Measure	Instrument real utilisation and token volume for at least one full quarter of production use.	You have a measured duty cycle for the §2 economics rather than a projection.
4 · Relocate	Stand up the facility, mirror the artifact supply chain, migrate steady-state inference first.	Sustained workloads run on owned capacity while bursty work stays rented.

MIGRATING AN EXISTING DEPLOYMENT

The question is usually consolidation rather than replacement. From assistant subscriptions the driver is cost and governance. From a managed ML platform, notebooks and model registries have counterparts while managed endpoints, feature stores and proprietary pipeline definitions do not. From shadow usage, inventory it from egress logs: that inventory is both the migration scope and the business case.

Repatriation is widely reported, including in vendor-sponsored surveys. One 2026 survey of 203 enterprise IT decision-makers reported that 79% had moved some AI workloads from public cloud to private or on-premises infrastructure, or were in the process of doing so, within 24 months.[19] Treat sponsored figures as directional and disclose the sponsor if you cite them.

SECTION 13

Evaluation checklist

Use this checklist in the review. Each item should be answerable with a demonstration.

PROCUREMENT & APPROVAL

- ☐ Security questionnaire and vendor onboarding started in parallel with the technical evaluation.
- ☐ Every artifact in Table 10.1 has an owner and a date.
- ☐ The shared responsibility matrix is marked up and agreed in writing.
- ☐ The approval threshold the number falls into is known.
- ☐ A one-page recommendation exists that can be forwarded without rewriting.

COST

- ☐ A three-year model covers all eight categories, including people and exit.
- ☐ The inputs the answer is most sensitive to are stated.
- ☐ Per-team quota is enforced at the gateway, not merely reported.
- ☐ An agentic token multiplier was applied, and its source is stated.

BOUNDARY & EGRESS

- ☐ Every egress destination is enumerated, owned and justified; default is deny.
- ☐ Workloads cannot bypass the egress proxy at the network layer.
- ☐ Each agent tool has its own egress allowlist and its own sandbox.
- ☐ Vendor telemetry is off by default, verified in traffic.
- ☐ Disconnected operation has been tested if it has been claimed.

IDENTITY & ACCESS

- ☐ All human access federates to the enterprise IdP. No local accounts.
- ☐ Table 7.1 is complete, with a supported-today answer in every row.
- ☐ Workloads use attested identity with short-lived credentials.
- ☐ Agent permissions are the **intersection** of the agent grant and the user's rights.
- ☐ Joiner-mover-leaver propagates within the standard SLA, with evidence.

DATA BOUNDARY

- ☐ The vector store is classified at the level of its source data.
- ☐ Retrieval enforces source-system ACLs **in the query**, not as a post-filter.
- ☐ Prompts and outputs are hashed or redacted at capture.
- ☐ KV-cache reuse across tenants is understood and deliberately configured.
- ☐ The multi-tenancy isolation boundary is written down.
- ☐ Backups and evaluation datasets are inside the classification scope.

AUDIT & EVIDENCE

- ☐ The audit record is append-only and an administrator cannot silently alter it.
- ☐ Telemetry and the audit record are separate systems with separate retention.
- ☐ Evidence export is routine, filterable by user, model and policy decision.
- ☐ Retention satisfies the longest applicable obligation, tested by restore.

CAPACITY

- ☐ Sizing was computed from context × concurrency, not parameter count.
- ☐ Power, cooling and rack density were confirmed **before** procurement.
- ☐ Contention degrades by documented priority policy, not at random.

OPERATIONS & CONTINUITY

- ☐ A model and library upgrade path exists, with version pinning and rollback.
- ☐ An evaluation gate blocks promotion of a new model or precision.
- ☐ DR covers weights, indexes and audit history, with RPO and RTO per store.
- ☐ A restore has actually been performed, and the date is recorded.
- ☐ Your own team can rebuild the environment from source and mirrored artifacts.
- ☐ Escrow, change-of-control and support SLA terms are in the contract.

FIVE CLAIMS REQUIRING EVIDENCE

1. **Enumerated egress.** Every path out is known, allowlisted and logged. The default is deny.
2. **Attributed access.** Every model request carries an authenticated identity, human or agent. Nothing accepts anonymous or shared-key traffic.
3. **Permission-preserving retrieval.** Retrieval cannot return what the requesting identity could not read at source.
4. **Tamper-evident record.** Governance events are append-only and administrators cannot silently alter them.
5. **Bounded agency.** Every agent has a scoped identity, an allowlisted tool set, and an evidenced worst-case scenario.

Sources

Quantitative, technical, and regulatory claims are linked to the references below. Derived estimates and time-sensitive statuses are identified in the body.

1. NVIDIA, *Mastering LLM Techniques: Inference Optimization*: KV cache, grouped-query attention. developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/
2. NVIDIA NIM, *Support Matrix*: per-model minimum and recommended GPU memory and precisions. docs.nvidia.com/nim/large-language-models/latest/reference/support-matrix.html
3. gpt-oss-120b model card: MXFP4 checkpoint size, MoE active parameters. huggingface.co/openai/gpt-oss-120b
4. vLLM recipes, *DeepSeek-R1*: 8×H200 FP8 (TP8 plus expert parallel); 4×B200 at FP4. recipes.vllm.ai/deepseek-ai/DeepSeek-R1
5. Silex Data, *Evaluating Llama 3.3 70B Inference: H100 vs A100*: concurrency sweeps. Third-party benchmark. blog.silexdata.com/blog/evaluating-llama-33-70b-inference-h100-a100/
6. Anthropic Engineering, *How we built our multi-agent research system*: agents ≈4× chat tokens; multi-agent ≈15×. anthropic.com/engineering/multi-agent-research-system
7. vLLM, *Conserving Memory*: `gpu_memory_utilization` and the practical overhead margin. docs.vllm.ai/en/latest/configuration/conserving_memory/
8. vLLM, *Parallelism and Scaling*: tensor parallelism within a node, pipeline parallelism across nodes. docs.vllm.ai/en/latest/serving/parallelism_scaling/
9. NVIDIA, *DGX SuperPOD Data Center Design, Electrical*: per-system draw and reference rack density. docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html
10. NVIDIA, *GB200 NVL72*: rack-scale power envelope and liquid cooling. nvidia.com/en-us/data-center/gb200-nvl72/ docs.nvidia.com/dgx/dgxgb200-user-guide/hardware.html
11. CloudZero, *H100 GPU cost*; Introl, *GPU cloud price collapse, December 2025*: purchase price, rental rates, payback. Secondary sources; estimates vary widely. cloudzero.com/blog/h100-gpu-cost/ introl.com/blog/gpu-cloud-price-collapse-h100-market-december-2025
12. Google Cloud, *Egress gateway best practices*: registry-only mode, and network policy to prevent sidecar bypass. docs.cloud.google.com/service-mesh/docs/istio-apis/security/egress-gateways-best-practices
13. Model Context Protocol, *Security Best Practices*: tool-description poisoning and egress controls for tool servers. modelcontextprotocol.io/docs/2026-07-28/tutorials/security/security_best_practices
14. SPIFFE/SPIRE: attested workload identity, short-lived SVIDs; bridging to enterprise authorisation via OAuth 2.0 token exchange (RFC 8693). redhat.com/en/topics/security/spiffe-and-spire OAuth 2.0 Token Exchange (RFC 8693): rfc-editor.org/rfc/rfc8693
15. IETF OAuth WG, *Identity Assertion JWT Authorization Grant* (Cross-App Access). **Internet-Draft, not a ratified standard.** datatracker.ietf.org/doc/draft-ietf-oauth-identity-assertion-authz-grant/
16. EU AI Act, Articles 12 and 19: automatic event logging; minimum six-month log retention. eur-lex.europa.eu/eli/reg/2024/1689/oj?locale=en
17. Morris et al., *Text Embeddings Reveal (Almost) As Much As Text*: reconstruction of source text from embedding vectors. arxiv.org/abs/2310.06816
18. Participant-aware access control for enterprise AI (arXiv:2509.14608): ACL propagation into retrieval. arxiv.org/abs/2509.14608
19. Cloudian, *2026 Enterprise AI Infrastructure Survey*, n=203. **Vendor-sponsored; disclose when citing.** storagenewsletter.com/wp-content/uploads/2026/03/Cloudian-AI-Infrastructure-Survey_Report.pdf
20. HHS / Federal Register, *HIPAA Security Rule NPRM*, 6 January 2025, and HHS fact sheet. **Proposed, not final.** hhs.gov/hipaa/for-professionals/security/hipaa-security-rule-nprm/index.html federalregister.gov/public-inspection/2024-30983/health-insurance-portability-and-accountability-act-security-rule-to-strengthen-the-cybersecurity-of
21. FedRAMP 20x programme pages: phases and authorisation classes. Status changes frequently; confirm before citing. fedramp.gov/20x/
22. FedRAMP changelog and Winvale reporting: the Emerging Technology Prioritization Framework (GenAI) was removed in January 2025. fedramp.gov/changelog/ info.winvale.com/blog/fedramp-emerging-technology-prioritization-framework-eliminated
23. DISA Cloud Computing SRG impact levels; Microsoft, *DoD IL5*: US-person access and tenant separation requirements. learn.microsoft.com/en-us/azure/compliance/offerings/offering-dod-il5 cyber.mil/dccs/dccs-documents/
24. NIST, *Control Overlays for Securing AI Systems* (COSAiS): concept paper August 2025; overlays in development. csrc.nist.gov/projects/cosais
25. Gibson Dunn, *EU AI Act Omnibus agreement*: deferral of high-risk obligations; Article 50 transparency obligations not deferred. **Confirm final adopted text.** gibsondunn.com/eu-ai-act-omnibus-agreement-postponed-high-risk-deadlines-and-other-key-changes/
26. AICPA, *Trust Services Criteria*: the criteria a SOC 2 report is issued against, including the CC series. aicpa-cima.com/resources/download/2017-trust-services-criteria-with-revised-points-of-focus-2022
27. HHS, *Guidance Regarding Methods for De-identification of PHI*: Safe Harbor and Expert Determination, 45 CFR §164.514(b). hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html
28. US Department of State, *ITAR*, 22 CFR Parts 120 to 130, including the deemed export rule. **Confirm with export counsel.** ecfr.gov/current/title-22/chapter-I/subchapter-M
29. NIST, *Cryptographic Module Validation Program*: FIPS 140-3 validated modules. csrc.nist.gov/projects/cryptographic-module-validation-program

---

SOURCE STATUS

Hardware figures, programme statuses, and regulatory dates reflect the cited sources at the time of writing. Items marked as in motion require confirmation before use.

PUBLISHER

This reference was published by Shakudo. Its architecture and evaluation questions are intended to be applied to any supplier, including Shakudo.