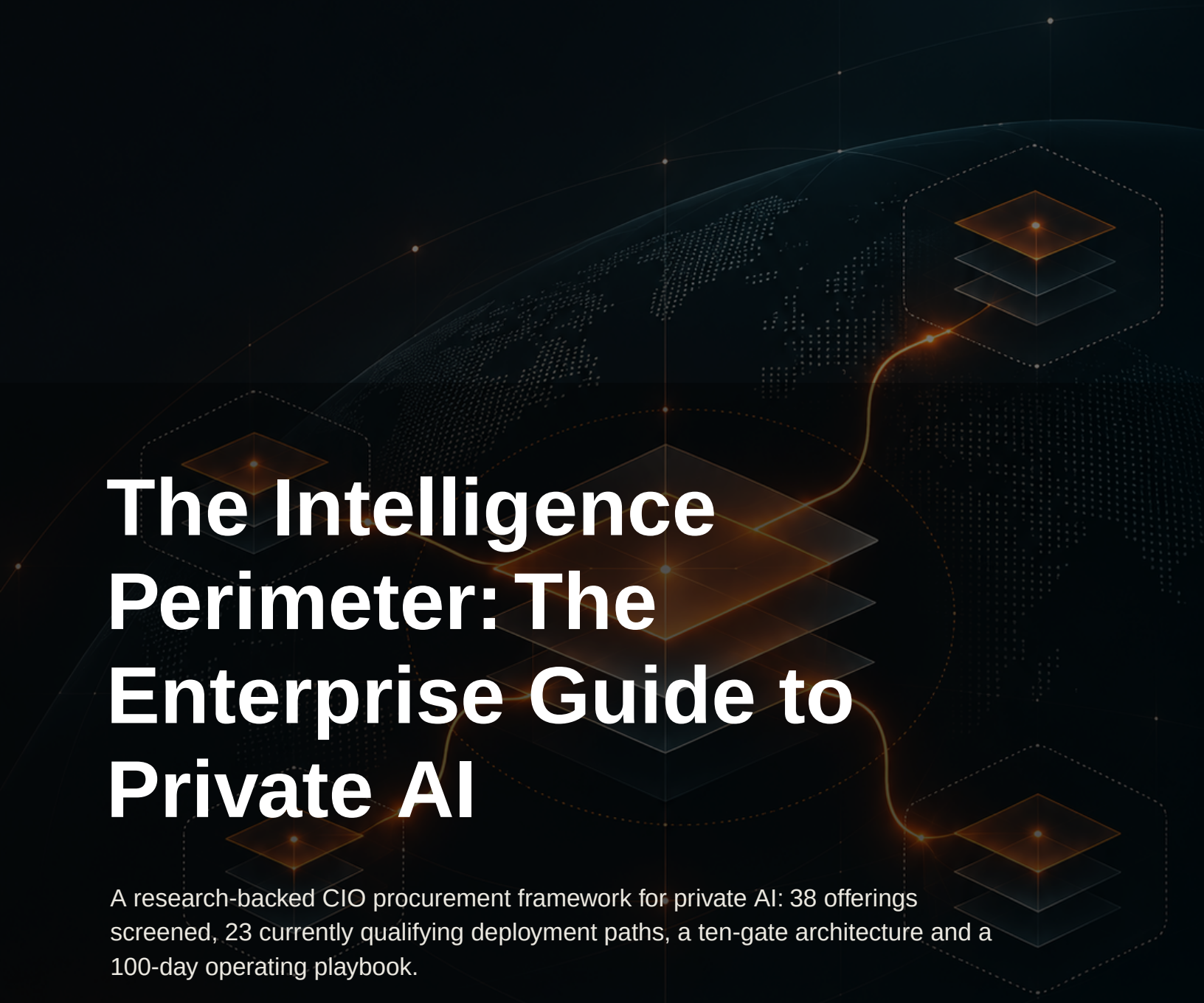
FRONTIER.

FRONTIER RESEARCH / DECISION-GRADE ANALYSIS



The Intelligence Perimeter: The Enterprise Guide to Private AI

A research-backed CIO procurement framework for private AI: 38 offerings screened, 23 currently qualifying deployment paths, a ten-gate architecture and a 100-day operating playbook.

FR-033 / Dion Hinchcliffe / November 16, 2026

REPORT PROFILE

Built to change a decision.

A research-backed CIO procurement framework for private AI: 38 offerings screened, 23 currently qualifying deployment paths, a ten-gate architecture and a 100-day operating playbook.

REPORT	**FR-033**
AUTHOR	**Dion Hinchcliffe**
PUBLISHED	**November 16, 2026**
PROFILE	**frontier-ro-0.1**

This PDF is a publication rendering of the canonical Frontier Research Object. The web report remains the live source of truth; citations, figures, methodology, and rights metadata travel with this edition.
FR-033 | Frontier Research | 16 November 2026 | Dion Hinchcliffe, Chief Research Officer

The executive finding: Bring intelligence under enterprise authority

The center of gravity in enterprise AI is moving. The first wave gave employees access to intelligence through rented APIs and embedded assistants. The next asks a harder question: where is the enterprise actually allowed to think? A global insurer has claims files in Frankfurt, private research in Boston, a contact center in Manila, and regulatory promises in half a dozen jurisdictions. Its AI cannot acquire a single operating home by decree. Yet its applications need a predictable way to use models without sending sensitive work through whichever external endpoint happened to be fashionable last quarter.

Private AI is the ability to place and operate model execution, enterprise context, agent tools, and their evidence within a boundary the enterprise can specify, test, and enforce. The boundary may be a customer-operated data center, a colocation footprint, a disconnected sovereign estate, or a contractually isolated and locally operated managed environment. The legal, operational, and technical consequences differ sharply among them. A private network endpoint alone does not establish sovereignty; neither does an encryption checkbox. The actual control lies in who can inspect, move, alter, suspend, or recover each element of a workload.[1][2][3]

The Frontier thesis is that the enterprise should own its intelligence perimeter while preserving the ability to move the work. A private AI estate that cannot be provisioned rapidly, independently audited, or migrated from public infrastructure will quickly turn into an expensive machine room. An estate that meets those tests can become the computing home for sensitive workflows, high-volume inference, proprietary models, and agents whose privileges must be bounded by business policy.

The market is already credible. Engineered AI factories from Dell, HPE, Lenovo, Cisco, and Supermicro join serving platforms from NVIDIA, Red Hat, VMware, Nutanix, and SUSE; customer-site offerings from AWS, Google, IBM, and Oracle; and enterprise application and model stacks from Cohere, Mistral, Palantir, H2O.ai, Cloudera, Domino, Aleph Alpha, and SambaNova. These products do not offer the same level of control, and a successful buyer cannot procure them as though they are interchangeable. This report uses a strict, dated gate to distinguish offerings documented as available now from components, previews, and public-cloud-only promises.[4][5][6][7][8][9][10][11]

Decision in one minute. Favor private placement where the fully loaded cost per accepted business outcome improves, where binding data or intellectual-property controls demand it, or where local latency and continuity create a competitive advantage. Retain public AI where it is the better legal, financial, or technical choice. Standardize identity, policy, workload contracts, evidence, and evaluation across both. A CIO's scorecard should reward outcomes under enforceable authority, not the number of racks purchased.

EXPLORE THE EVIDENCE — THE PRIVATE AI CONTROL ATLAS. Open the navigable Data Insight for all 38 screened product paths (23 admitted, four enabling components, 11 held), six operating layers and ten buyer gates. Its source-backed dossiers expose direct company/product links and explicitly scoped CEO, revenue and employee context. Go directly to the interactive operating fabric.

Open the interactive Frontier Private AI Operating Fabric atlas

Why private AI is now an operating strategy

The early private-AI debate was dominated by reassuring phrases about data never leaving a building. The new operating question is wider. Modern agents retrieve records, store embeddings, retain key–value caches, search vector indexes, invoke tools, request external models, record traces, and update systems of record. The prompt and output are only two stops on a much longer journey. Enterprises now need residency and access controls for weights, prompts, retrieved passages, embeddings, caches, intermediate reasoning state where retained, tool payloads, observability traces, backups, evaluation samples, and support logs. A compliance review that examines only where the model endpoint runs has left most of the architecture unseen.[12][1]

There are four practical placement contracts. Customer-operated means the enterprise operates equipment and the day-two platform in its site or colocation facility; this offers direct control but also makes hardware failures and specialist staffing the customer's problem. Vendor-managed on-site keeps the workload in a customer facility while delegating operations, hardware maintenance, and potentially the control plane to a supplier. Disconnected sovereign emphasizes independence from public networks and external control planes, often at the cost of slower updates and narrower model choice. Dedicated private cloud can provide single-tenant capacity and data locality without customer-operated hardware, but it rarely satisfies requirements that prohibit an external operator from controlling the service. Each contract may be legitimate; their trust boundaries must be made explicit.[13][14][15][16]

This is why data residency, operational sovereignty, confidentiality during use, and model sovereignty must be negotiated separately. Data can remain in a country while administrators abroad hold access paths. A model can run in a confidential enclave but still call an external retrieval service. A closed-weight model may have a private deployment option without allowing the customer to inspect or redistribute its weights. A customer may own the GPU but depend on a supplier's licensing service to keep the inference application running. Ask which layer holds power, and how to prove it under an incident.

The regulatory landscape gives these questions urgency without simplifying them. The EU GDPR imposes conditions on international personal-data transfers, including onward transfers; the EDPS explains that permitted remote access can itself constitute a cross-border transfer. The EU AI Act rolls out obligations by activity and risk category; the timetable changes by provision. The EU Data Act creates cloud-switching obligations and phases out certain switching charges from January 2027, but the law does not turn a proprietary model interface or an imported data-governance policy into a portable one. Multi-geographic enterprises should obtain local legal counsel on the exact data, service, and personnel involved.[2][3][17][18]

Ten gates before a vendor even enters the room

Our qualifying framework deliberately separates five hard gates, which a candidate must pass for a specified workload, from five operating-fit tests, which compare the options that remain. There is no universal certificate saying an offering is private enough for every enterprise. Residency and sovereignty are requirements defined by the buyer's jurisdiction, contracts, regulator, and threat model.

As Figure 1 shows, the five hard gates determine whether a deployment is even eligible; only then should procurement compare provisioning speed, economics, performance, migration and regional operations.

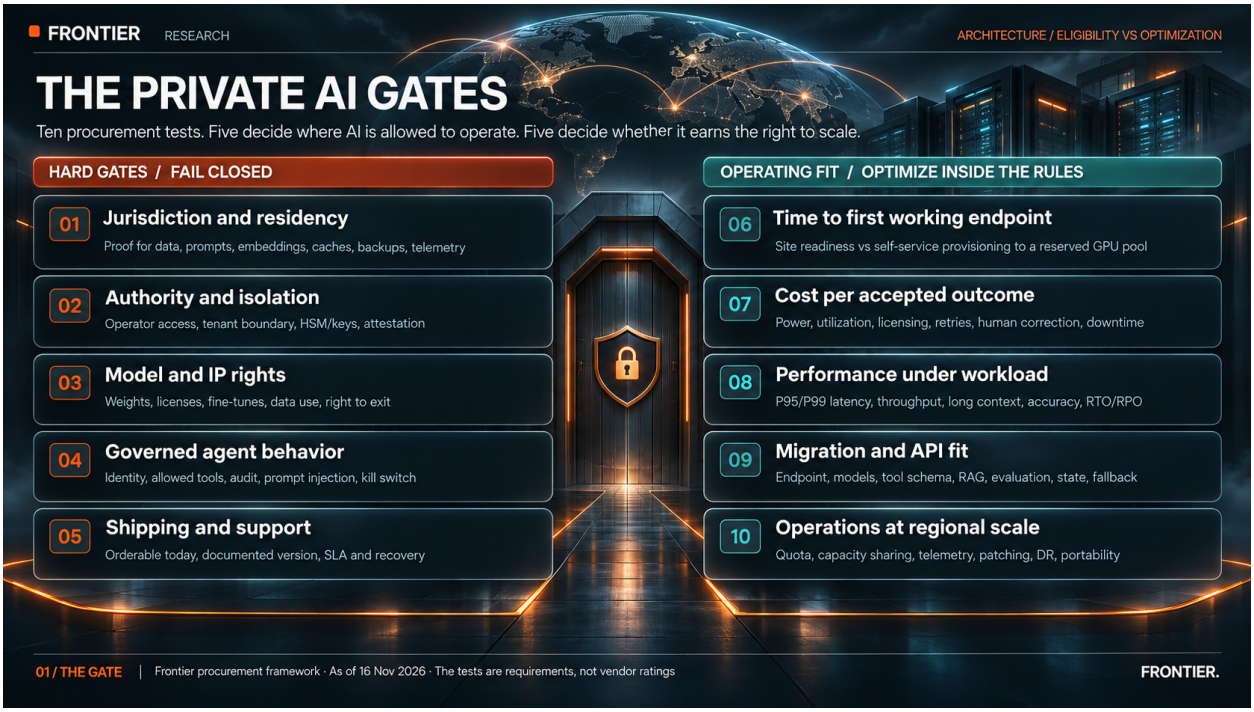


FIG. 01 Five hard gates establish whether a proposed private-AI deployment is eligible for a specific workload. Only candidates that pass proceed to five operating-fit tests covering provisioning speed, cost per accepted outcome, performance, migration and regional operations. Requirements are buyer-specific; no vendor scores are implied.

Frontier Research — original Private AI procurement framework, 16 November 2026.

Gate 01 — Jurisdiction and residency. Trace every byte of sensitive payload, retrieved context, training data, cache, trace, backup, and administrative access. Obtain a data-flow diagram for the precise product version and configuration, including failover, troubleshooting, model updates, monitoring, and support. A local model whose logging service exports prompts abroad has failed this gate.

Gate 02 — Authority and isolation. Define the operator, controller, processor, tenant, and key owner. Inspect IAM, network controls, privileged support procedures, HSM/KMS placement, attestation evidence where claimed, and the actual retention behavior of operational data. A buyer must be able to revoke access and prove the revocation worked.

Gate 03 — Model and intellectual-property rights. Identify the exact model version and license; rights to use, fine-tune, retain, redistribute, export adapters, store KV or embeddings, and run commercial workloads. Open weights do not mean every license is permissive. Proprietary frontier models may be absent from the private catalog even when the platform's cloud edition offers them.

Gate 04 — Governed agent behavior. An agent's tool identity must be narrower than its user's general permissions; the organization needs allowlisted tools, transaction approval points, workload isolation, traceable decisions, red-team testing, and an emergency suspension procedure. Data isolation alone does not stop an authorized but mistaken agent from making a consequential system change. Our prior research on AI governance and the autonomous enterprise explains why the operating boundary belongs around the agent's actions as well as around inference.[19][20][1]

Gate 05 — Shipping and supported today. Demand an orderable SKU or subscription, current release notes, supported hardware, production-eligible license terms, deployment instructions, support SLA, and referenceable customer evidence or verifiable production delivery. A slide showing a feature planned for the next quarter does not meet the threshold. A pre-release endpoint licensed only for evaluation cannot be treated as a production product.

Gate 06 — Time to a working endpoint. Separate the calendar time to obtain site power, racks, cooling, hardware, and operator clearance from self-service provisioning within an already available pool. These may be months and minutes, respectively. For an established estate, require a developer to provision a new approved, identity-governed endpoint without an infrastructure ticket marathon; measure elapsed time, not demonstration clicks. AWS AI Factories, for example, documents a roughly three-to-six-month deployment after a site is ready, despite subsequent AWS-console familiarity.[13]

Gate 07 — Cost per accepted business outcome. Include equipment or reservation charges, financing, power and PUE, cooling, network, storage, licensing, human operations, security, recovery capacity, model updates, evaluation, retries, and rework. Normalize by verified successful tasks at the required quality and risk threshold. Prices per GPU-hour or token are useful ingredients, not the answer.

Gate 08 — Performance under real workloads. Compare p50, p95, and p99 time-to-first-token and final completion; simultaneous users, queueing, long-context prefill, KV-cache pressure, retrieval latency, tool wait, and output-quality failures. Include regional traffic and a failure of a whole rack or availability zone. Never use a vendor benchmark as a replacement for your own acceptance test.

Gate 09 — Migration and API fit. Test the actual application and its model contract, not only the ability to send POST /v1/chat/completions. Schema, system prompts, tool-call formats, constrained JSON, tokenization, embeddings, streaming, model version pins, context limits, multimodal payloads, safety controls, identity, and tracing often diverge across runtimes. NVIDIA NIM exposes both OpenAI- and Anthropic-style endpoints on supported backends, but backend-specific functions still vary; the word compatible is always bounded by a versioned test matrix.[12]

Gate 10 — Operations across countries. Test GPU quotas, tenant metering, patch and vulnerability cadence, country-by-country data retention, DR and restore, private registry synchronization, incident evidence, capacity expansion, upgrade rollbacks, backup portability, and the commercial right to exit. An international organization also needs a repeatable local landing zone, not a bespoke project in every country.

Market map: 38 offerings considered; 23 admitted now

We screened 38 named offerings as of November 16, 2026, using publicly accessible vendor documentation, product availability statements, implementation material, and, where present, customer delivery evidence. We admitted 23 full product or infrastructure paths for a specified private deployment scope; four additional platforms/components can make the architecture work but are not complete private-AI delivery contracts on their own; eleven were held because of preview status, absence of a verified private data plane, or incomplete offering scope. This is a public-evidence screen, not a field audit, procurement certification, performance ranking, or assertion that every vendor can serve every jurisdiction or model.

These are 38 distinct evaluated product and deployment paths from 35 distinct supplier names, not 38 separate corporations. Google Cloud, NVIDIA, and Oracle each appear twice because an admitted on-premises or disconnected product is evaluated separately from another hosted, adjacent, or preview offering that did not pass the same gate. A supplier can also span architectural roles, while one customer deployment may combine several. Figure 2 maps functions and control boundaries, not product scores or market share.

Figure 2 shows the six layers through representative product paths. Its stylized provider marks and names are identifiers, not independently authenticated logo files, endorsements or comparative scores. The key distinction is who supplies the infrastructure, who operates the software, and whose control plane still has access.

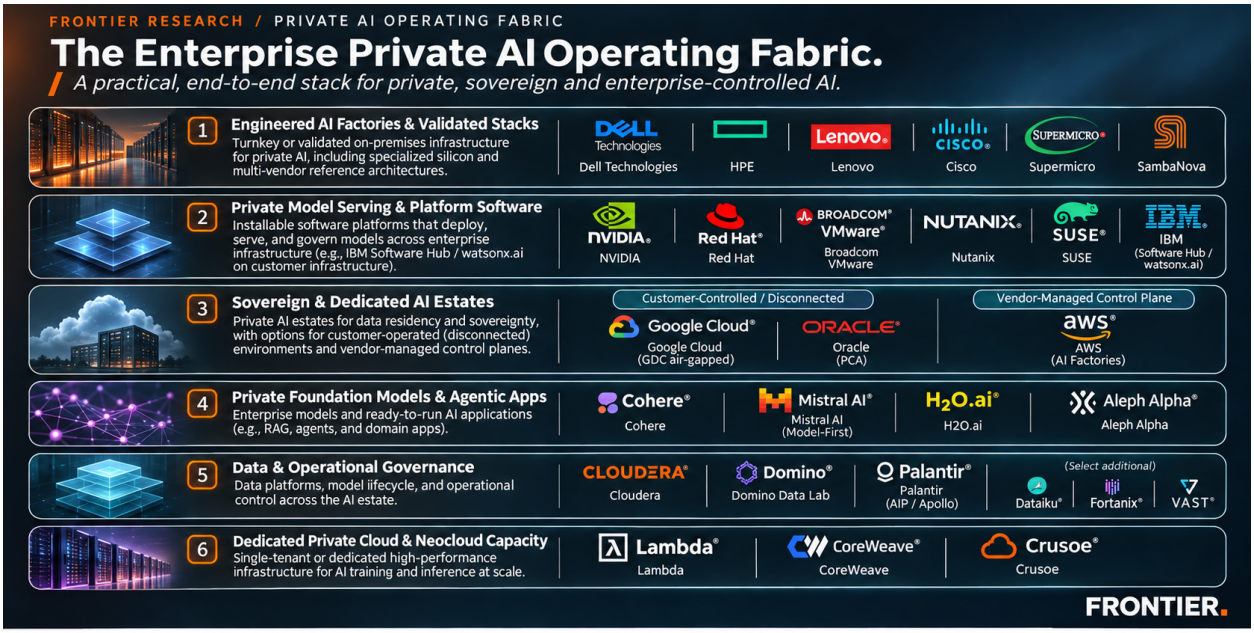


Figure 2. The Private AI Operating Fabric. A six-layer taxonomy of the enterprise private AI stack, with representative providers delivering production-ready offerings today. Providers are grouped by their primary role and deployment model, from infrastructure and serving platforms to sovereign estates, models and applications, data governance and dedicated capacity.

Source: Frontier Research (2026). Provider identifiers are typographic representations of company names and marks.

FIG. 02 Six operating lanes clarify ownership and control: SambaNova supplies on-premises RDU systems, IBM Software Hub runs on enterprise OpenShift, Google GDC air-gapped and Oracle PCA have different operator models, AWS AI Factories remains AWS-managed and region-connected, and Lambda Private Cloud provides hosted single-tenant GPUs. Provider identifiers illustrate documented products and enabling components, without implying endorsement or certification.

Frontier Research — independently compiled product-scope taxonomy, 16 November 2026. Company marks are illustrative identifiers; trademarks belong to their respective owners.

Navigate Figure 2, rather than merely read it. Open the interactive six-layer Private AI atlas to click individual providers, inspect the exact documented offering, compare product boundaries and download the canonical evidence pack. The financial context shown in each dossier belongs to the disclosed corporate or reporting scope, not automatically to the Private AI product.

1. Engineered factories and validated stacks — Dell, HPE, Lenovo, Cisco, Supermicro, SambaNova. The category includes integrated customer-site hardware and supported multi-vendor reference architectures; it does not imply that every supplier manufactures the whole stack. Dell, HPE, Lenovo and Supermicro offer validated infrastructure designs, with differing software and service contracts. Cisco's Secure AI Factory is a validated reference design combining Cisco UCS compute, networking, security and observability with NVIDIA accelerators and certified storage partners such as NetApp, Pure Storage and VAST; the exact bill of materials and cross-vendor support remain procurement tests. SambaNova's SambaRack and SambaStack are an on-premises, proprietary-RDU system with integrated serving, so they belong here rather than among hosted GPU rentals. Confirm each production SKU, site prerequisites, power, support ownership and upgrade rights before ordering.[4][5][6][7][21][22][23]

2. Private serving and platform software — NVIDIA, Red Hat, VMware, Nutanix, SUSE, IBM. These are installable software and serving control layers that run on compatible infrastructure; they do not all include a customer-site rack. NVIDIA NIM and NVIDIA AI Enterprise provide model serving and scheduling components. Red Hat's AI Inference is built around vLLM and developing llm-d capabilities, some of which remain preview. VMware Private AI Foundation brings model services into VMware Cloud Foundation; Nutanix Enterprise AI offers governed endpoints and orchestration, and SUSE AI Factory builds on private Kubernetes infrastructure. IBM Software Hub/watsonx.ai is installed on customer-supplied Red Hat OpenShift and GPU capacity, with its own software prerequisites, operators and private-registry processes. Buying the software does not transfer hardware ownership or day-two responsibility to IBM. Treat the serving platform, the hardware foundation and the security operating model as separately evidenced contracts.[24][25][26][8][27][11]

3. Sovereign and customer-site estates — Google Distributed Cloud air-gapped, Oracle Private Cloud Appliance and AWS AI Factories. Their shared characteristic is physical placement close to the customer, not the same sovereignty or operating contract. Google Distributed Cloud air-gapped can keep eligible models and workloads in a disconnected environment, but remains a Google-provided and supported system whose operator and service-access terms must be checked. Oracle Private Cloud Appliance is a customer-owned and operated engineered appliance, with its own local OCI-compatible control plane; it does not imply that every OCI regional AI service is available inside it. AWS AI Factories places dedicated capacity in customer facilities or colocation but remains AWS-managed and connected to a selected AWS Region. It should not be used as evidence of a fully customer-operated, disconnected estate. IBM's installable watsonx.ai software is placed in Layer 2. Establish who controls keys, remote access, firmware, network dependencies and emergency operations before describing any of these paths as sovereign.[10][14][15][28][13][11]

4. Private foundation models and agentic applications — Mistral AI, Cohere, H2O.ai and Aleph Alpha. These suppliers offer different levels of the stack. Mistral is primarily a foundation-model and private-model-deployment option; its Studio workers can be self-hosted without guaranteeing that the Studio control plane is local. Cohere North adds an enterprise agent application experience around privately deployed models. H2O.ai h2oGPTe includes governed retrieval and agent/RAG functionality for private and air-gapped settings. Aleph Alpha PhariaAI supplies a private enterprise AI platform with an on-premises release; buyers should still specify the included user experiences, model families, orchestration and support tier. Model weights, application UX, governance, and a wholly local control plane are separate entitlements, even when sold under one product family.[29][30][31][32][33][34]

5. Data, model and operational governance — Cloudera, Domino, Palantir, Dataiku, Fortanix and VAST. These are complementary control functions, not interchangeable data platforms. Cloudera combines governed data services with on-premises inference; Domino focuses on the model development, lifecycle and MLOps record. Palantir's AIP/Foundry/Apollo combination spans operational ontology, agentic business workflows and continuous software/model deployment, including disconnected estates; Apollo is a cross-environment delivery control plane, not just a data-governance tool. Dataiku addresses governed AI workflows and access, Fortanix addresses confidential execution and controlled key release, and VAST supplies the data platform. The latter three remain qualified components in Appendix A rather than individually admitted end-to-end private AI stacks. The customer's architecture determines how these roles interoperate.[35][36][37][38][39][40][41]

6. Dedicated private cloud and GPU capacity — Lambda Private Cloud. Lambda offers hosted bare-metal, single-tenant GPU clusters for reserved terms, with tenant network isolation, hardware support and optional managed Kubernetes. It is a private-capacity rental in the provider's data centers, rather than an enterprise-owned or customer-site GPU appliance. That placement can be attractive for stable demand but does not satisfy an explicit requirement for customer-operated hardware, a disconnected site, or local operator jurisdiction by itself. SambaNova's on-premises RDU racks are addressed in Layer 1 because their deployment and control model is fundamentally different. Price and test the exact isolation, operations, exit, egress and support contract rather than treating the word private as a location guarantee.[16][23]

Four qualified building blocks, not standalone factories. Canonical's supported Charmed Kubeflow can underpin air-gapped MLOps, but customers still need an inference gateway and enterprise policy architecture. Dataiku LLM Mesh governs access to hosted and local models and supports policy and routing; some Dataiku AI Services require an outbound connection even with customer-chosen models. VAST AI OS contributes storage, vector/data services, and integration around inference; its DataEnclave confidential runtime requires its own availability check. Fortanix, with Rafay, provides attested confidential model execution and controlled key release; it complements an inference infrastructure stack rather than replacing it. These names belong in the broader provider atlas with their component scope exposed.[42][43][39][41][44][45][40]

The first striking exclusion is Microsoft's Foundry Local on Azure Local. The architectural approach is relevant and publicly documented for enterprise on-premises inference, but the current offer remains preview/access-request in its own documentation. Likewise, the on-customer-hardware Gemini on GDC connected API is restricted to nonproduction preview, despite a separately shipping GDC air-gapped offering. More familiar services such as Databricks Mosaic AI Model Serving, Snowflake Cortex, OpenAI/Anthropic managed enterprise APIs, and standard hosted inference endpoints have strong governance features but lack documented, general customer-operated on-premises delivery of the same service. Their exclusion is about the requested boundary, not their value elsewhere.[46][47][48][49][50][51][52]

Put the applications on a moving platform

A successful private AI architecture starts with a workload contract that travels across environments. Picture it as a shipping container for an enterprise task: purpose, sensitivity, allowed models and tools, evaluation cases, response format, latency budget, audit rules, rollback, and where the task may legally execute. The container need not own a particular accelerator. Its route can change; its promises to the business must not.

Create a single enterprise inference/API facade that can target public models, privately deployed open models, and where rights permit, privately deployed commercial models. Preserve the original public service as a fallback while a private pool proves its quality. For many applications using open or OpenAI-like chat APIs, much of the connection wiring can be moved behind a gateway with limited code changes. It is still necessary to test tool invocation, JSON constraints, streaming, embeddings, tokenizer behavior, batching, rate limits, safety settings, model versions, memory/cache design, prompt differences, telemetry, and error handling. A new endpoint that returns different tool decisions has changed the application even if the HTTP calls look identical.[12][53]

Use three migration classes. Lift: stateless inference calls that can pass golden evaluations on a local model with an endpoint and credentials update; target migration in days once capacity exists. Adapt: RAG, agent and multimodal work requiring retrieval, access-control, and state changes; plan a few sprints and explicit security sign-off. Rebuild or retain public: workloads dependent on unavailable proprietary models, cloud-native tools, managed fine-tuning, very large contexts, or commercial datasets that cannot be locally licensed; preserve a cloud path or redesign the task rather than forcing an artificial conversion. The classification is per workload, not per application portfolio.

The most frequently missed migration layer is context. If the model moves to a local rack but retrieval still queries an external SaaS index, the sensitive facts may continue leaving the intended perimeter. If an agent still invokes unapproved remote tools, locality has not delivered the promised risk posture. Move the embeddings, vector index, policy filters, agent state, cache, and observability evidence as a unit when the workload's classification requires it. Rebuild indexes in the permitted region, compare retrieval recall, and test whether every tool is allowed to operate there.

Model weights deserve similar attention. Track model origin, license, the exact digest, tokenizer and prompt template, fine-tuning adapters, quantization, inference backend, accelerator architecture, and benchmark approval. A vendor's cloud catalog may have models that are not available for private deployment. A platform's statement about freedom of model choice means little without a bill of materials for the model formats and accelerators actually supported on the purchased release.

Where token routing helps—and where it can hurt

Our earlier AI Token Routing Research and the accompanying interactive market map surveyed 59 offerings across eight categories. The strategic lesson carries directly into Private AI: applications should be able to request a capability, quality band and policy class while the enterprise retains the authority to decide which allowed model and estate performs the work. It is an unusually useful design pattern when public and private capacity coexist.[54][55]

A well-designed router first evaluates policy, then chooses among authorized models using measured quality, cost and latency, without silently crossing a geographic or security boundary. Sensitive banking requests might be pinned to Frankfurt private inference; a non-sensitive internal search query might use a cheaper external model. An overloaded private cluster might shift only preapproved low-sensitivity requests to public capacity, never silently downgrade confidentiality. When no permitted backend can meet the service contract, the correct response is a documented rejection or human escalation, not a clever fallback that violates residency.

Routing can also make a bad architecture worse. It creates an extra control plane, another place for logging or prompt leakage, distributed tracing that crosses boundaries, policy conflicts between the router and model endpoint, session and KV-cache discontinuity, inconsistent structured output, and failure modes that look like model errors. Caching can mix sensitive contexts across tenants if isolation is not enforced. Evaluation drift across model versions can cause agents to make different decisions. The router must be inside the same trust and audit design as the models it invokes. The cost case must include the router itself, not merely the lower-priced tokens it buys.

The recommended pattern is federated private-first placement: regional local inference pools expose standard service contracts; an enterprise policy router decides whether a task may execute in each pool; telemetry aggregates permitted performance and cost metrics without exporting restricted prompt bodies. Keep country-level authority over key management and incident response. Test full round-trip portability in both directions every quarter. This is more realistic than a fantasy of a universal public/private API with identical behavior.

Economics: the denominator that matters

Private AI tends to win when a workload is steady, high-volume, repeatable, and well matched to the hardware; or when legal, latency, or IP requirements remove public placement from the feasible set. Public AI remains economically compelling for unpredictable demand, rapid access to a frontier model, short experiments, and organizations without an operations team. Dedicated hardware can be wasted for long periods when demand arrives in bursts. Extra privacy controls and redundancy cost money; a local estate is not automatically cheaper. Vendor savings claims are scenario-specific marketing claims until reproduced with accepted business outputs.

Use a three-year, risk-adjusted cost per accepted business outcome:

Private three-year unit cost = (hardware + financing + support + software licenses + facility/power/cooling + networking/storage + security + people + upgrade/failure/DR costs + migration + model evaluations) / accepted outcomes.

Compare that with the equivalent public path including provider charges, tokens, reserved commitments, storage, retrieval, egress, retries, monitoring, evaluation, human review and compliance overhead. Hold quality, availability and risk acceptance constant. Include a realistic idle-load scenario, peak-load scenario, currency and energy sensitivity, and full recovery cost. The FinOps Open Cost and Usage Specification provides a useful normalization vocabulary, but data from private GPUs needs a customer-defined allocation model for business services and outcomes. [56]

A deliberately illustrative example: a global enterprise runs a multilingual claims-review service in four jurisdictions. The first private site uses the correct-size model on a reserved accelerator pool, and regional branches subscribe to approved inference endpoints. During the first quarter, GPU utilization is 30% and private unit cost loses to a public model; by month nine, better batching and shared non-sensitive capacity raise utilization to 65%, and private unit economics improve. If the private model fails the accuracy and escalation tests, none of that matters: the organization has bought cheap failures. No numbers in this example are a vendor benchmark or a forecast for a specific customer. The exercise demonstrates why utilization, acceptance quality and regional policy must be priced together.

The procurement test should specify p95 latency, accepted-output rate, cost per accepted outcome, GPU utilization at target concurrency, operating labor, and recovery after a lost node. Negotiate price protection and the ability to reallocate licensed compute across permitted countries. Include physical power and supply lead times in the capital approval, and avoid permanent hardware purchases before the workload mix has stabilized.

Five illustrative enterprise patterns

A multinational bank: a jurisdictional perimeter. A hypothetical bank has sensitive customer dossiers in Frankfurt, London and Singapore, and routine public research in the United States. It could use a supported self-hosted inference layer such as Red Hat or VMware for regulated local tasks, with a policy router that rejects cross-border moves, while using public APIs for a separate low-risk research assistant. The real selection criteria are legal data-flow proof, local key custody, transaction authority, response quality and cross-site recovery. The named products illustrate a topology; we make no product recommendation or suitability certification.[57][26][2]

A life-sciences company: proprietary models near proprietary data. A hypothetical multinational pharmaceutical firm might assemble Dell or HPE infrastructure with NVIDIA NIM and protected local data to test molecule-screening assistants against confidential laboratory data. The core task is maintaining weight and dataset rights, exact provenance, reproducibility, isolation and audit. An on-site deployment might reduce copies of experimental data, while a public foundation model remains a benchmark if licensing and policy permit. The company should not mistake a confidentiality claim for a complete validated research environment.[4][5][12]

A sovereign public authority: disconnected by design. A government service with strict sovereignty requirements might compare Google Distributed Cloud air-gapped with a fully assembled, customer-operated alternative. IBM watsonx.ai can run on enterprise OpenShift/GPU infrastructure; Nutanix and SUSE provide private serving foundations but require an eligible local model and application layer, for example a licensed Mistral deployment or an Aleph Alpha PhariaAI stack, plus the necessary retrieval, identity and governance controls. Google supplies an integrated disconnected estate, subject to its specific operator contract; the other paths require deliberate integration and support ownership. The decisive tests are prolonged network isolation, independent key and administrator control, offline updates and evaluation, model rights and incident recovery. These are illustrative topologies, not equivalent turnkey SKUs or endorsements.[14][11][58][27][29][34]

A manufacturing group: private inference at the edge. A hypothetical automotive supplier might put visual inspection and maintenance assistants near plants in Mexico, Germany, Poland and Thailand. Lenovo or Cisco validated AI pods could serve supported models with low latency, while a central policy catalog and region-local evidence stores maintain controls. This is a deployment pattern, not proof that a robot or visual AI will meet the plant's safety and quality obligations. Test local outage behavior, physical and network security, model drift and field-service SLAs.[59][7]

A smaller enterprise: earn the right to buy hardware. A midmarket company can start with a dedicated rented GPU service or existing virtualized infrastructure and a supported model serving stack rather than purchasing a large factory. If the system delivers a stable workload and a repeatable quality gate, graduate it to a compact private footprint. Its first discipline should be workload and policy contracts and realistic support costs. Lambda Private Cloud illustrates the single-tenant rental path; Nutanix or supported self-hosted model software illustrates the equipment-operated path. Neither product is being endorsed.[16][58]

A 100-day private-AI conversion playbook

Days 0–15: make the enterprise intelligence register. Inventory the top 30–50 AI-enabled workflows by business outcome, volume, sensitivity, countries, model dependencies, agent tools and existing public provider. Record who owns each workflow and the system of record. Separate prompt payload, source documents, embeddings, KV cache, traces, vector indexes and model rights. Establish the current accepted-output baseline, service level, and full unit cost. Exit condition: each pilot has a known data path and a named accountable business owner.

Days 16–30: establish two approved landing zones. Choose one ordinary on-site or colocation pool and one higher-control profile where needed, or document why a dedicated managed estate is sufficient. Implement SSO/IAM, HSM-backed keys where required, signed model artifacts, Kubernetes or serving control plane, a private registry, logging and audit retention. Define a standard model endpoint template, tenant quota, rollback and incident procedure. Exit condition: a developer can request an approved endpoint and a security officer can verify its placement and permissions.

Days 31–45: prove the first migration. Select two high-volume, bounded workflows and one sensitive benchmark. Replay representative public-model requests through the private facade. Evaluate tool calls, JSON formatting, multilingual behavior, retrieval precision, latency, error rates, and human review. Record every difference. Where a private model falls short, improve the workflow or retain the permitted cloud path rather than changing the acceptance criterion. Exit condition: quality and policy gates pass with production-like traffic.

Days 46–60: add routing without loosening the border. Place a policy decision ahead of model selection. Label every request with tenant, region, sensitivity, business service, model class and audit identifier. Set safe failure behavior, local-only cache rules, and non-sensitive burst routing. Use separate keys and access policies for local and external inference. Exit condition: a deliberately prohibited cross-border request fails reliably and leaves reviewable evidence.[54]

Days 61–75: harden day-two operations. Run a model rollback, a compromised credential exercise, a lost GPU node, a full cluster rebuild from an immutable manifest, and a multi-region outage drill. Test quota exhaustion, driver/security patching, prompt-injection failures, external telemetry restrictions, and restoration of model artifacts from local backups. Exit condition: operations meets the agreed recovery targets without a vendor engineer improvising from scratch.

Days 76–90: compare economics and expand deliberately. Benchmark at representative 25%, 60%, and 85% utilization, including burst queues. Calculate accepted outcomes per dollar, incident exposure, staffing and failure cost; identify workloads that remain better in public AI. Keep the public path for those that legitimately win. Exit condition: the investment committee sees a credible 36-month comparison and specific workloads ready for scale.

Days 91–100: adopt an operating charter. Publish the private-AI service catalog, inclusion/exit criteria, allowable models, regional placement policy, change windows, audit thresholds, metrics, procurement rights, and quarterly portability test. Build demand forecasts with finance and business owners. Establish exceptions that require CISO/data-controller approval and a human business sponsor. Exit condition: the next team can obtain a governed model endpoint without recreating the architecture debate.

The RFP that exposes the uncomfortable truths

Require every bidder to answer these questions in writing, against the exact release and deployment topology:

1. Which model serving and agent features are generally available today? Supply release notes and customer contracts; flag preview items.
2. Can this particular deployment run entirely with the public network disconnected for 72 hours? Which licenses, logs, models, patches and support tools stop working?
3. Who can access prompts, context, weights, traces and keys—including the vendor's support engineers and subcontractors?
4. Where do full prompts, embeddings, KV state, content indexes, telemetry and backups reside at rest and during processing? Identify every third-country support route.
5. Can the customer retain models, adapters, indices, prompts, metrics and policies if the contract ends? Describe the exit procedure and license rights.
6. How long from an empty, power-ready site to the first supported endpoint, and how long from an existing GPU pool to an approved developer endpoint?
7. Which accelerators, model families, context sizes, quantization methods, multimodal functions and serving engines are supported in this exact SKU?
8. Can an OpenAI-style client migrate with verified parity for tool calls, structured JSON, streaming, authentication, RAG, caching and error handling? Supply the supported compatibility matrix.
9. Can one business unit see another's prompts, model files, cache or traces? Provide isolation and penetration-test evidence.
10. How are agent tools authorized, executed, observed, stopped and compensated after mistakes?
11. How are model supply-chain records, SBOMs, CVEs, firmware trust and secure downloads managed in an isolated network?
12. What are the p95/p99 accepted-task latency and throughput at the target concurrency? Include a complete test harness and quality failure rate.

13. How is per-workload consumption metered across tenants, and how does the tool calculate cost per accepted outcome rather than GPU-hour alone?
14. What is the supported path for a failed GPU, deleted cache, lost zone and an entire site outage? Show RTO, RPO and tested restore artifacts.
15. What minimum commitments, refresh penalties, cross-country license restrictions, maintenance exclusions and accelerator reservations apply?
16. Which models remain unavailable for private deployment despite appearing in the vendor's cloud catalog? Name them.

Score only candidates that pass the workload's five hard gates. Within that set, run the same workload harness, with a common dataset, prompts, ground-truth rubric, security attack suite and total-cost assumptions. A bid that declines a material answer should not be credited by implication. Ensure every country-level controller can review the relevant records independently of headquarters.

What changes in the next three years

Private AI will increasingly resemble a network of governed regional utilities: a few high-capacity sites, smaller specialist inference pools near data and people, a federated catalog of approved models, and a policy engine deciding which tasks can travel. The most expensive mistake is likely to be building an isolated GPU island that cannot exchange workloads with the rest of the enterprise. The second is attaching a shiny private label to an environment whose operator, model rights, or telemetry still cross forbidden boundaries.

The stakes expand as AI agents take on more of the transaction path. Cost, latency and sovereignty begin to compete on every job. Specialized accelerators and confidential computing will create more placement options, but not eliminate model-rights and operations constraints. The most capable proprietary model may remain accessible only through a vendor service; the local alternative might be a different open-weight model needing prompt and evaluation changes. CIOs should plan for an intelligent, measured mix of the two instead of a hypothetical clean migration of every workload.

This report sets the research baseline for a versioned Private AI Provider Atlas—a navigable Data Insight that will expose every reviewed offering, eligibility decision, jurisdictional and operational evidence, product dependencies and sourcing. The atlas is the second phase of this project; it is not yet a published URL. When verified, the report will be amended with prominent inline links and navigable graphics, and only then will the publication PDF be produced. The existing AI Token Router Atlas provides the precedent for source-backed, inspectable records.[55]

Private AI succeeds when an enterprise can choose where intelligence runs, prove what happens there, and shift the work without surrendering the rules. That means defining the boundary before buying hardware, forcing proof before trusting a product claim, and measuring the unit economics of verified work. The reward is room to innovate with sensitive data, predictable capacity where it matters, and the freedom to choose tomorrow's model without rebuilding the enterprise around it.

Appendix A — The market screen, product by product

Eligibility legend: ADMITTED = documented, active product path available for a defined private workload (not a universal certification); COMPONENT = shipping supporting layer requiring an additional stack to provide the complete service; HOLD = preview-only, lacking a demonstrated on-prem/customer-controlled data plane for the requested scope, or only adjacent facilities/hosted APIs. All outcomes are as of 16 November 2026. Vendor capabilities are described from attributable technical or official product sources and have not been independently installed by Frontier.

1. Dell Technologies — AI Factory with NVIDIA · ADMITTED. Validated on-prem AI factory; customer scale and designs documented; exact reference stack and costs must be contracted. [4]
2. HPE — Private Cloud AI / AI Factory · ADMITTED. Shipping with air-gapped large system; later 2026 agent/governance features must be verified individually. [5]
3. Lenovo — Hybrid AI Factory / Sovereign Inference · ADMITTED. Validated on-prem/edge configurations; specific IBM/NVIDIA integrations require scoped SKUs. [6]
4. Cisco — Secure AI Factory / AI PODs · ADMITTED. Validated on-premises reference architecture built from Cisco UCS, networking and security, NVIDIA accelerators, and certified partner storage; validate the complete bill of materials and split-vendor service accountability. [7][21]
5. Supermicro — Enterprise AI Factory Clusters · ADMITTED. Turnkey certified NVIDIA clusters; enterprise control software is largely bundled partner stack. [22]
6. NVIDIA — AI Enterprise / NIM / Run:ai · ADMITTED. Supported on-prem and fully disconnected model serving; needs compatible infrastructure and governance. [24]
7. Red Hat — AI Inference / OpenShift AI · ADMITTED. Commercial self-hosted serving with multiple accelerators; some distributed llm-d capabilities preview. [25]
8. Broadcom VMware — Private AI Foundation with NVIDIA · ADMITTED. VCF private AI service with self-service, GPU quotas, model serving; assumes suitable VCF platform. [26]
9. Nutanix — Enterprise AI / Agent Gateway · ADMITTED. NAI 2.8 and governed model/agent endpoints available; customer should check specific Kubernetes options. [8]
10. SUSE — AI Factory with NVIDIA · ADMITTED. Generally available, private Kubernetes/LINUX AI platform; hardware and model bundles require design. [27]

11. IBM — Software Hub / watsonx.ai on premises · ADMITTED. Installable AI software and local model serving on customer-supplied Red Hat OpenShift/GPU infrastructure; software and cluster operations remain separate responsibilities. [11]
12. Google Cloud — Distributed Cloud air-gapped (Gemini) · ADMITTED. Google-provided and supported disconnected hardware/software estate with eligible local models; clarify supplier operator privileges, model version and country-specific contract rather than assuming customer-run infrastructure. [10][14]
13. AWS — AI Factories with Bedrock/SageMaker · ADMITTED. Dedicated capacity at customer-owned or colocated facilities, but AWS-managed and connected to an AWS Region; not a disconnected customer-operated environment. Site-ready delivery remains a multi-month commitment. [13]
14. Oracle — Private Cloud Appliance GPU · ADMITTED. Customer-owned and operated OCI-compatible engineered appliance with a local control plane and GPU shapes; it does not include every managed OCI Generative AI regional service. [15][28]
15. Cohere — North / enterprise private model deployments · ADMITTED. On-prem, air-gap and customer VPC options; distinguish North, Cohere models and dedicated Model Vault. [31]
16. Mistral AI — Enterprise self-hosted model deployments · ADMITTED. Enterprise on-prem licensing and support; Studio self-host workers may still depend on external Mistral control plane. [29]
17. Cloudera — AI Inference on premises · ADMITTED. On-prem AI inference serving integrated with Cloudera data estate; confirm local version and model coverage. [35]
18. H2O.ai — Enterprise h2oGPTe / H2O AI Hybrid Cloud · ADMITTED. Air-gapped model and agent application with governance and routing; model endpoints differ by topology. [33]
19. Domino Data Lab — Enterprise MLOps / Agentic AI · ADMITTED. Documented private VPC and on-prem model hosting, governance and reproducibility; verify license/support tier. [36]
20. Palantir — AIP with Apollo · ADMITTED. Operational ontology, agent workflows and Apollo-managed continuous deployment across on-premises and disconnected environments; spans application execution and software delivery, beyond data/ML governance. [37][38]
21. Aleph Alpha — PhariaAI · ADMITTED. On-prem release v1.261100 documented; evaluate model family, language support and integration fit. [34]
22. SambaNova — SambaRack + SambaStack · ADMITTED. Customer-site proprietary RDU hardware factory with integrated model serving and OpenAI-compatible APIs; benchmark the supported models and platform lifecycle separately from hosted neocloud rentals. [23]
23. Lambda — Lambda Private Cloud · ADMITTED. Hosted provider-operated bare-metal GPU clusters reserved to one tenant with isolated networking and support; optional managed Kubernetes, but not on-premises customer-owned infrastructure. [16]
24. Canonical — Charmed Kubeflow · COMPONENT. Supported air-gapped MLOps substrate; inference gateway and policy stack must be assembled. [42]
25. Dataiku — LLM Mesh / Agent Management · COMPONENT. Governed routing to local models; some AI Services require Internet even with BYO LLM; verify offline requirements. [39]

26. VAST Data — AI OS / DataEnclave · COMPONENT. Mature data plane; confidential DataEnclave announced, verify purchase status and model dependencies. [41]
27. Fortanix / Rafay — Confidential AI & CoCo orchestration · COMPONENT. Available confidential workload attestation and keys; complementary to, not replacement for, inference platform. [40]
28. Microsoft — Foundry Local on Azure Local · HOLD — preview. This is Microsoft's Azure Local/Arc-enabled Kubernetes cluster inference offering, distinct from the separate Foundry Local device SDK; the enterprise cluster deployment remains preview/access-request, so its exact product name and exclusion are retained. [46]
29. Google Cloud — Gemini on GDC connected API · HOLD — preview. Explicit pre-GA restrictions prohibit commercial production use; distinct from GDC air-gapped. [47]
30. Databricks — Mosaic AI Model Serving · HOLD — locality. Cloud-hosted AI/ML serving; not proof of customer-owned/offline model infrastructure. [48]
31. Snowflake — Cortex AI · HOLD — locality. Managed cloud AI/SQL service; no verified on-prem AI data plane in cited product. [49]
32. OpenAI — Closed API and ChatGPT Enterprise · HOLD — locality. Private data terms and gateway services do not establish local hosting of proprietary models. [50]
33. Anthropic — Claude enterprise/cloud APIs · HOLD — locality. Enterprise privacy options are valuable but local Claude weights are not established for ordinary customers. [51]
34. Hugging Face — Inference Endpoints · HOLD — adjacent. Dedicated managed endpoints are useful but self-hosting/air-gap control varies; not full stack by default. [52]
35. Equinix — Colocation / interconnection · HOLD — adjacent. Facility and connectivity foundation; AI platform/controls must be procured separately. [60]
36. Oracle — OCI managed Generative AI service · HOLD — adjacent. Dedicated regional model service does not imply availability on each Private Cloud Appliance. [61]
37. NVIDIA — Public developer build API · HOLD — adjacent. Hosted developer API is not local inference; downloadable NIM is separately admitted. [12]
38. Meta — Llama open-weight model family · HOLD — adjacent. A model family is not a fully managed private AI product; enterprise support requires integrator/stack. [62]

Appendix B — Decision rights, evidence quality, and what this edition excludes

We based the market gate on published product documentation, release notes and vendor technical material, supported by open standards, industry institutions and our prior Frontier research. Our eligibility categories describe deployment and operating scopes, not relative model quality, ROI, security certification, national compliance or buyer suitability. Claims such as vendor savings percentages, managed service uptime, full disconnected operation and absolute portability must be independently tested under contract. We do not cite competing technology analyst firms. Some supplier documents can change frequently; a later product release may move an offering into or out of a category.

This edition does not claim to measure market share, purchase prices, deployment counts across the entire industry, independently benchmarked inference throughput, or every jurisdiction's legality. It does not certify confidential computing from marketing descriptions, resolve laws on remote operator access, or replace legal advice. Its procurement scores are buyer-configured questions, not algorithmic product recommendations. Future Atlas revisions will track version-specific changes explicitly, with source dates and published cutoffs.

Rights and independence: Company names and the stylized, recognizable logo treatments in Figure 2 are used solely to identify example offerings; all trademarks and brand rights remain with their owners. This is an independent Frontier editorial visualization, not a vendor-authenticated logo library, certification or endorsement. The forthcoming interactive Atlas will require separately sourced logo assets and mark provenance. Frontier's model of a private-AI workload remains a decision aid, not a guarantee of business outcome, legal compliance or technical suitability.

SOURCES / CITATIONS

Trace the argument.

Every source below is linked. Source numbering follows first appearance in the report.

01 **NIST, AI Risk Management Framework: Generative AI Profile (AI 600-1), July 2024**
https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

02 **European Union, GDPR Regulation 2016/679, Article 44 cross-border personal data transfers**
https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng/

03 **European Data Protection Supervisor, International transfers and remote access**
https://www.edps.europa.eu/data-protection/data-protection/reference-library/international-transfers_en

04 **Dell, AI Factory with NVIDIA portfolio and deployed customers**
https://www.dell.com/en-us/newsroom/announcements/dell-technologies-closes-the-gap-between-ai-ambition-and-ai-outcomes

05 **HPE, Private Cloud AI production deployments and air-gapped configuration, March 2026**
https://www.hpe.com/us/en/newsroom/press-release/2026/03/hpe-accelerates-secure-scalable-production-ready-ai-through-new-innovations-with-nvidia.html

06 **Lenovo, Sovereign AI Inference with IBM and NVIDIA, October 2026**
https://lenovopress.lenovo.com/lp2558-own-your-own-intelligence-the-lenovo-sovereign-ai-inference-ibm-nvidia

07 **Cisco, Secure AI Factory with NVIDIA and on-prem AI PODs**
https://www.cisco.com/site/us/en/solutions/artificial-intelligence/secure-ai-factory/index.html

08 **Nutanix, Enterprise AI 2.8 and MCP gateway GA August 2026**
https://www.nutanix.com/press-releases/2026/nutanix-gives-enterprises-the-freedom-to-run-production-agentic-ai-their-way

09 **Amazon Web Services, AI Factories now available, Dec 2025**
https://aws.amazon.com/about-aws/whats-new/2025/12/aws-ai-factories/

10 **Google Cloud, Distributed Cloud air-gapped and Gemini on-site**
https://cloud.google.com/distributed-cloud-air-gapped

11 **IBM, installing watsonx.ai on premises in Software Hub 5.3**
https://www.ibm.com/docs/en/software-hub/5.3.x?topic=watsonxai-installing

12 **NVIDIA, NIM model endpoint API including Anthropic and OpenAI compatibility**
https://docs.nvidia.com/nim/large-language-models/latest/api-reference.html

13 **AWS, AI Factories FAQ: customer site, Bedrock, SLAs, 3-6 month site delivery**
https://aws.amazon.com/about-aws/global-infrastructure/ai-factories/faqs/

14 **Google Cloud, Distributed Cloud air-gapped isolation and control**
https://docs.cloud.google.com/distributed-cloud/hosted/docs/latest/gdcag/overview

15 **Oracle, Private Cloud Appliance with customer-controlled GPU on premises**
https://docs.oracle.com/en-us/iaas/private-cloud-appliance/pca/home.htm

16 **Lambda, Private Cloud dedicated single-tenant GPU clusters**
https://docs.lambda.ai/private-cloud/

17 **European Commission, EU AI Act implementation timeline and risk-based controls**
https://ai-act-service-desk.ec.europa.eu/en/ai-act/eu-ai-act-implementation-timeline

18 **European Commission, Data Act explained: cloud switching and fees**
https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained

19 **Frontier Research, Your AI Policy Cannot Stop an Agent: The State of Enterprise AI Governance**
https://www.thisisfrontier.ai/research/your-ai-policy-cant-stop-an-agent-ai-governance-platforms

20 **Frontier Research, The Autonomous Enterprise**
https://www.thisisfrontier.ai/research/the-autonomous-enterprise

21 **Cisco, AI POD for Splunk on-prem now available, September 2026**
https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2026/m09/cisco-delivers-trusted-ai-at-scale-through-new-splunk-advancements.html

22 **Supermicro, Enterprise AI Factory turnkey validated NVIDIA clusters**
https://www.supermicro.com/en/pressreleases/supermicro-announces-new-ai-factory-cluster-solutions-based-nvidia-enterprise

23 **SambaNova, SambaStack on-prem and private enterprise inference stack**
https://sambanova.ai/products/sambastack

24 **NVIDIA, NIM LLM air-gapped deployment and offline model caching**
https://docs.nvidia.com/nim/large-language-models/latest/deploy-air-gap.html

25 **Red Hat, AI Inference System Card, architecture and supported model/accelerator choices**
https://access.redhat.com/ai/system-card/red-hat-ai-inference

26 **Broadcom VMware, Private AI Foundation with NVIDIA, self-service and NIM serving**
https://blogs.vmware.com/cloud-foundation/2026/07/31/deploying-nvidia-runai-on-vmware-cloud-foundation/

27 **SUSE, AI Factory with NVIDIA GA July 2026**
https://www.suse.com/c/suse-ai-factory-with-nvidia-is-now-generally-available/

28 **Oracle, Private Cloud Appliance GPU L40S shapes and requirements**
https://docs.oracle.com/en-us/iaas/private-cloud-appliance/pca/compute/compute-shapes.htm

29 **Mistral, enterprise models can be deployed on-prem or in customer cloud, licenses vary**
https://mistral.ai/products/ai-cloud/inference/

30 **Mistral Studio self-hosted workflow caveat: external control plane**
https://docs.mistral.ai/studio/deployments/self-hosted/introduction

31 **Cohere, private deployment options: customer VPC, on-prem, air gap**
https://docs.cohere.com/docs/deployment-options-overview

32 **Cohere, North 2, on-prem/VPC enterprise agent platform**
https://cohere.com/north

33 **H2O.ai, Enterprise h2oGPTe on-prem, private and air-gapped deployment**
https://h2o.ai/products/h2ogpte

34 **Aleph Alpha, PhariaAI on-prem production release Nov 2026**
https://docs.aleph-alpha.com/phariaai-home/latest/release-notes/pharia-ai/1.261100.0.html

35 **Cloudera, AI Inference on premises expanded February 2026**
https://www.cloudera.com/about/news-and-blogs/press-releases/2026-02-09-cloudera-unveils-next-phase-of-ai-inferencing-and-unified-data-access-capabilities.html

36 **Domino, self-managed private and on-prem LLM/agent hosting with governance**
https://domino.ai/platform/agentic-ai

37 **Palantir, AIP/Apollo on private networks and air-gapped environments**

https://www.palantir.com/platforms/

38 **Palantir, Apollo deployment in disconnected and air-gapped estates**

https://www.palantir.com/docs/apollo/core/introduction

39 **Dataiku, AI Services model connection and internet requirement caveat**

https://doc.dataiku.com/agent-management/latest/ai/setup.html

40 **Fortanix and Rafay: CoCo support available now Nov 2026**

https://www.fortanix.com/company/pr/2026/11/rafay-fortanix-help-neoclouds-sovereign-ai-clouds-enterprises-deliver-confidential-ai-as-a-service

41 **VAST Data, AI OS data and agent stack**

https://kb.vastdata.com/documentation/docs/vast-ai-os

42 **Canonical, supported Charmed Kubeflow on bare metal, air-gapped and hybrid**

https://canonical.com/mlops/kubeflow

43 **Dataiku, LLM Mesh governed access to locally hosted and external models**

https://www.dataiku.com/product/llm-mesh

44 **VAST, DataEnclave announcement, deployment status to verify**

https://www.vastdata.com/press-releases/vast-data-introduces-dataenclave-to-bring-leading-ai-models-and-enterprise-data-together-on-trusted-infrastructure

45 **Fortanix Confidential AI with model attestation and controlled key release**

https://www.fortanix.com/resources/solution-briefs/fortanix-confidential-ai-for-ai-model-owners

46 **Microsoft Foundry Local on Azure Local, preview only**

https://learn.microsoft.com/en-us/azure/azure-sovereign-clouds/private/foundry-local/what-is-foundry-local-on-azure-local

47 **Google Cloud, Gemini on GDC connected API is preview-only, commercial use not allowed**

https://docs.cloud.google.com/distributed-cloud/gemini-on-gdcc/latest/docs/overview

48 **Databricks, Mosaic AI Model Serving and Databricks cloud product**

https://docs.databricks.com/aws/en/machine-learning/model-serving/

49 **Snowflake, Cortex AI managed SQL/LLM service**

https://docs.snowflake.com/en/user-guide/snowflake-cortex/overview

50 **OpenAI, privacy/security and enterprise API controls**

https://openai.com/enterprise-privacy/

51 **Anthropic, enterprise privacy and safety commitments**

https://www.anthropic.com/enterprise

52 **Hugging Face Inference Endpoints product, private dedicated VPC options**

https://huggingface.co/docs/inference-endpoints/index

53 **vLLM official documentation, prefix caching and shared context performance**

https://docs.vllm.ai/en/stable/design/prefix_caching/

54 **Frontier Research, Rent the Models. Own the Rules: The Enterprise Guide to AI Token Routing**

https://www.thisisfrontier.ai/research/rent-the-models-own-the-rules-ai-token-routing

55 **Frontier Research, 2026 AI Token Routers Data Insight and 59-offering atlas**

https://www.thisisfrontier.ai/data-insights/2026-ai-token-routers

56 **FinOps Foundation, Open Cost and Usage Specification (FOCUS)**

https://focus.finops.org/what-is-focus/

57 **Red Hat, AI Inference product and licensed self-managed serving**

https://www.redhat.com/en/products/ai/inference

58 **Nutanix, Enterprise AI and hybrid inference control plane**

https://www.nutanix.com/products/nutanix-enterprise-ai

59 **Lenovo, Hybrid AI Advantage factory and agentic inference delivery, June 2026**

https://news.lenovo.com/pressroom/press-releases/lenovo-redefines-enterprise-ai-economics-with-agentic-ai-and-inferencing-innovations/

60 **Equinix, data center colocation and private AI ecosystem infrastructure**

https://www.equinix.com/solutions/artificial-intelligence

61 **Oracle, OCI Generative AI availability across regions and dedicated cluster shapes**

https://docs.oracle.com/en-us/iaas/Content/generative-ai/model-endpoint-regions.htm

62 **Meta AI, downloadable Llama model family and license terms; model weights are distinct from enterprise private AI platforms**

https://ai.meta.com/resources/models-and-libraries/llama-downloads/

63 **Domino, self-managed Enterprise/Premium license options**

https://domino.ai/pricing

64 **HPE, Private Cloud AI GA vs roadmap availability, October 2026**

https://www.hpe.com/us/en/newsroom/press-release/2026/10/hpe-teams-with-nvidia-to-advance-secure-agentic-ai-across-enterprises-and-industrial-environments.html

FRONTIER.

TAKE THE NEXT STEP

Push this further.

Research is useful when it changes a decision. Bring us the question, the constraint, or the argument you want tested.

RESEARCH / BRIEFINGS

inquiries@thisisfrontier.ai

Briefings, advisory work, executive discussions, and research inquiries.

REPRINTS / LICENSING

reprints@thisisfrontier.ai

Reprints, enterprise distribution, excerpts, and licensing.

FRONTIER

www.thisisfrontier.ai

Research, Data Insights, News, and the live Frontier corpus.

All rights reserved. Contact Frontier for redistribution or reprint rights.

SEE FURTHER. MOVE SOONER.