AIMONGER
WHITE PAPER

AI Discovery

Sovereign vs Managed Inference, A Board Decision Framework

As firms operationalise GenAI, the question shifts from model choice to inference placement: where does inference run, who can see the data, and what breaks if the vendor changes terms? Boards need a decision framework across managed APIs, private managed endpoints, and sovereign or local inference, tied to data classes, latency, talent, and cost, not ideology.

David Saliba . 2026-07-22 . aimonger.com/whitepapers/sovereign-vs-managed-inference/

The problem in one sentence

Where inference runs determines who can see your data, what you pay when usage spikes, and whether you can exit when terms change.

Boards hear two loud stories: “use the global API” and “build sovereign AI.” Both can be wrong defaults. The calm question is operational: for this workflow and this data class, which placement delivers control, quality, and total cost over twenty-four months?

McKinsey finds AI use widespread while EBIT impact stays concentrated among high performers who redesign workflows. Unresolved placement blocks absorption: security says no to managed APIs, engineering cannot run local models well, and the workflow waits in queue.

Sovereign inference, plainly: models run on infrastructure you control or tightly bound, so prompts and outputs stay inside your operational boundary. **In practice:** HR payroll queries run on a private GPU cluster in Frankfurt; marketing drafts stay on a managed EU-zone endpoint.

Managed API, in short: you send requests to a vendor-operated model over the network; control is contractual and architectural, not physical. **Worked case:** fast time-to-value, OpEx pricing, exit risk if unit prices or terms shift.

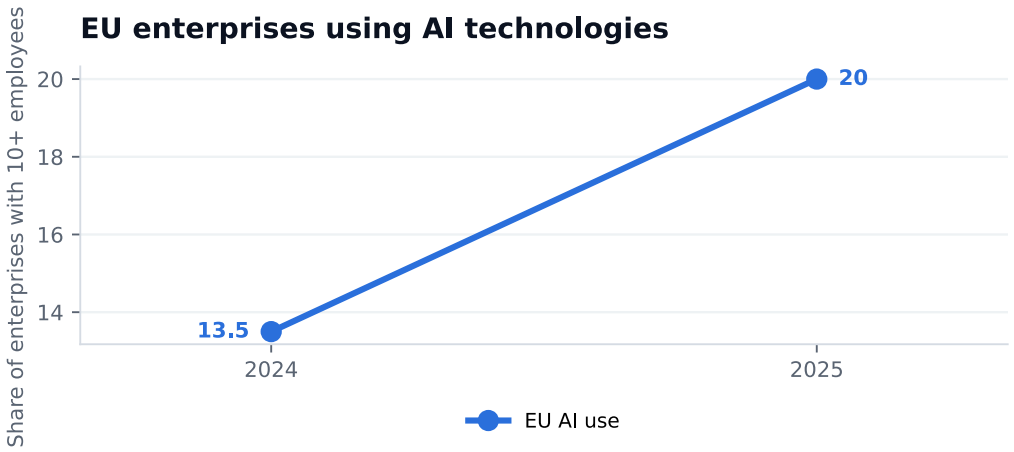
Data class means: a labelled category of information with fixed rules for where it may be processed and by whom. **On Monday:** “Confidential commercial” may use a private endpoint; “regulated personal data” may not use a consumer chat tool, period.

Placement is strategy

In boardrooms, “sovereign AI” can sound like the serious answer before anyone has priced the GPU estate, staffed the patching burden, or measured quality against managed alternatives. The opposite error is just as common: assume a global API is fine for every document because it is convenient.

What you see globally right now is cloud concentration on one side and sovereignty language on the other. US hyperscalers are pouring capital into AI infrastructure, national governments are debating domestic compute, and enterprise vendors are bundling managed agents into suites. leadership teams need a calmer question: which placement fits this data class and workflow?

McKinsey finds AI use widespread and EBIT impact scarce. One reason programmes stall is unresolved placement: security blocks managed APIs, engineering cannot run local models well, and the workflow waits. Gartner's mainstreaming projection for GenAI applications by 2026 means more vendors will offer managed agents and assistants. That increases convenience, and it increases concentration risk.



Eurostat, 2024 vs 2025. AIMonger redraw.

Figure 1. Share of EU enterprises (10+ employees) using AI technologies, 2024 vs 2025. Plain read: adoption is climbing but most firms are still choosing first placements. **Board reading:** you are not late to the sovereignty debate, and you are early if you decide by data class now. Source: Eurostat, December 2025. AIMonger redraw.

Decision matrix

Factor	Managed public API	Private managed endpoint	Sovereign / local
Data control	Contractual	Stronger boundary	Maximum operational control
Time to value	Fast	Medium	Slow unless skilled
Capex / ops burden	Low	Medium	High
Model freshness	Vendor cadence	Vendor cadence	Your upgrade burden
Talent required	Integration	Cloud + security	MLOps + infra + eval
Best for	Low/medium sensitivity	Sensitive enterprise data	High sensitivity / strict residency

Failure modes

- **Sovereign cosplay.** The firm runs a local model on under-patched hardware with no evaluation parity, gaining a sovereignty label while quality and security lag a hardened private endpoint. The label satisfies a policy line and the reality satisfies none of it, because nobody is patching CUDA, running holdouts, or monitoring the inference path. A sovereign estate that nobody operates is a concentration of risk indoors, not a reduction of risk.
- **Managed denial.** Security blocks all managed APIs without funding a working private or sovereign alternative, leaving workflows stuck in queue while staff route around policy in shadow tools. The block is a decision that creates no capability, and the shadow routing it provokes is a

larger data-leak risk than the managed API it refused. Denial without an alternative is governance theatre that moves risk rather than reducing it.

- **Split brain from shadow AI.** Teams adopt consumer chat tools because the official inference path is too slow or too narrow, multiplying data-leak risk and audit gaps the board thought it had closed. Each shadow tool is an ungoverned inference path with its own retention and training-use defaults, and the board has no visibility into any of them. The shadow estate grows in proportion to how badly the official path serves real workflows.
- **Freshness debt on local stacks.** Weights and runtimes fall six months behind managed offerings, so the "sovereign" path cannot meet task quality the business already expects from vendor cadence. The business compares local output to what a colleague got from a managed tool last week, and the local path looks incompetent. Freshness debt converts a control advantage into a quality disadvantage that users route around.

Board checklist

1. **Data classes mapped to placements.** Every information class has an explicit allowed inference placement so teams cannot silently route regulated personal data through a consumer API. The map is the policy artefact auditors ask for, and it is the thing that makes the placement decision enforceable rather than aspirational. Without it, placement is a series of local judgements that no two teams make the same way.
2. **Training-use and retention terms reviewed.** Legal has read vendor contracts for prompt retention, training use, and subprocessors, not assumed "enterprise tier means safe." Enterprise tier is a SKU, not a guarantee, and the training-use clause is the line that determines whether your prompts become someone else's model. Legal review is the control that turns a marketing claim into a contractual obligation.
3. **Exit plan for pricing or term changes.** The board knows how to migrate workflows if unit prices jump or terms shift, including export of logs and connector documentation deadlines. An exit plan is what makes a managed dependency a commercial choice rather than a trap. Without it, a vendor term change becomes a board crisis instead of a procurement negotiation.
4. **Evaluation parity across placements.** The same proposition holdouts run after every model or placement change so quality cannot diverge between Frankfurt local and EU-zone managed without detection. Parity is the control that makes a hybrid estate governable, because it gives the team a single quality ruler across placements. Without it, the placements drift apart and nobody can say which one is degrading.
5. **Named owner per production placement.** Each live inference path has a human owner accountable for patching, cost, and incident response, not "the cloud team" in aggregate. Aggregate ownership is no ownership, because nobody is on call when a specific path degrades. The named owner is the person the incident page calls at 2 a.m., and their name on the roster is what makes the placement production-ready.
6. **Twenty-four-month cost model for sovereign options.** Finance models idle GPU capacity, MLOps labour, and upgrade evaluation, not sticker price on hardware alone. Sticker price is the capex slide; the operating reality is idle capacity, patching hours, and the evaluation labour every upgrade demands. A twenty-four-month model is the comparison a private managed endpoint wins or loses on, and it is the number that prevents a sovereignty decision based on month-one hardware cost.

Data classes first, infrastructure second

Create three to five data classes, for example:

1. **Public and marketing.** Content approved for external publication may use managed APIs with standard contractual controls when no personal or confidential data is present. This is the class where convenience and freshness win, because the data is already public and the control requirements are minimal. Routing it to a sovereign estate wastes capex on data that does not need it.
2. **Internal general.** Day-to-day operational documents may use private managed endpoints where audit and residency terms are stronger than public multi-tenant pools. The class needs a contractual boundary but not physical control, so a private managed endpoint gives stronger terms than a public API without the operating burden of local hardware. This is where most enterprise workloads land.
3. **Confidential commercial.** NDAs, pricing, and unreleased strategy require private endpoints or sovereign paths with customer-managed keys and explicit subprocessors list. Leakage here is competitive damage, so the placement needs a boundary the firm can audit and keys the firm controls. The subprocessor list is what lets legal verify that no surprise affiliate processes the data.
4. **Regulated personal data.** HR, health, and identifiable customer records map to the strictest placement with logging, retention caps, and training-use prohibition by default. This is the class where a consumer chat tool is a reportable incident, not a productivity tool. The placement must satisfy residency, retention, and purpose-limitation rules that a managed public API cannot contractually guarantee.
5. **Secrets and credentials.** API keys, passwords, and authentication material never enter any model prompt, regardless of placement marketing or "enterprise" SKU labels. No inference path needs credentials to answer a question, and any prompt that includes them is a design fault regardless of where it runs. This class is a refusal rule, not a placement choice.

Map each class to allowed placements. Most fights disappear when class 4/5 cannot silently enter a consumer API.

Hybrid estates are normal

Serious enterprises often run:

- **Managed APIs for low-sensitivity drafting.** Marketing summaries and public FAQ drafts use EU-zone managed endpoints where data class allows fast freshness without sovereign capex. The managed path gives vendor cadence on the content that changes fastest and carries the least risk. This is where the convenience of managed inference is a real advantage rather than a compromise.
- **Private endpoints for confidential commercial corpora.** Client contracts and unreleased financials stay on dedicated VPC or customer-key environments with audit rights stronger than public API pools. The private endpoint raises the contractual boundary without raising the operating burden to full sovereignty. It is the placement that buys control where managed public is too open and sovereign is too expensive to operate.
- **Sovereign paths for regulated workloads.** Payroll, clinical, or litigation-support queries run on infrastructure the firm controls when residency, support path, or air-gap rules require it. Sovereignty here is a response to a specific rule, not a posture, and the operating cost is justified by a constraint no other placement can satisfy. The workload is narrow enough that the MLOps burden is bounded.

The control problem is consistent evaluation and identity across placements, not forcing one placement for everything.

For a Malta-origin or mid-market firm, hybrid is usually the practical answer: use managed paths where the data class allows it, reserve private or local inference for the workflows where control gains justify the operating burden.

Open weights as sovereign leverage (and Chinese-lab influence)

Sovereign placement got cheaper and more capable in 2026 because competitive open weights arrived, not because mid-market firms suddenly built foundation-model labs.

As of July 2026, boards should treat three open-weight bands as real placement options when they can staff serve and eval: DeepSeek-V4 Pro/Flash (1M-token open-weight family from April 2026; see DeepSeek's 24 April 2026 preview), Zhipu GLM-5.2 (MIT open weights from June 2026; see Z.ai's 16 June 2026 post), and Alibaba Qwen3.x (hosted Max preview plus a separate open-weight family cadence). CNBC journalism in June–July 2026 reported rapid enterprise and developer interest as Chinese open models undercut closed US token economics, and as temporary US government limits on some Anthropic and OpenAI rollouts made revoke-resistant weights look like insurance. The same journalism reported GLM-5.2 near Claude Opus 4.8 on a watched agentic benchmark at roughly one-fifth the token cost. Treat those comparisons as journalistic snapshots, not primary benchmark proof; prove quality on your own holdout before promotion.

The Chinese influence question is not optional for European boards. It splits into three different decisions:

1. **Self-hosted Chinese-lab weights.** Download and serve inside your VPC so prompts and alpha do not leave during inference. You still record origin, licence, checksum, and patch cadence. Procurement, insurers, and some customers may still object to Chinese-origin weights on policy grounds even when data never touches a Chinese endpoint.
2. **Chinese-hosted APIs.** Cheap tokens with a live data path to a foreign operator. For confidential commercial and regulated classes this is usually the wrong sovereignty story, even when the model brand is fashionable.
3. **US closed APIs under policy stress.** Access and export controls can change which flagship SKUs you may call. That is an argument for dual-run exit evidence and for open-weight trays on alpha workflows, not an argument to abandon evaluation.

> **Chinese-lab open weight, put simply:** a competitive downloadable model from a China-based lab that you may self-host for cost and revoke resistance, while still owning an origin and supply-chain decision. **Operating case:** DeepSeek-V4 or GLM-5.2 on a Frankfurt tray you patch, with the same proposition holdouts as your Azure OpenAI path; marketing drafts stay on the managed endpoint.

Freshness debt and sovereign cosplay still apply. A 700B-class MoE checkpoint on a thin POC tray is not sovereignty. It is a stalled upgrade with a Chinese filename.

Worked example: legal contract review (synthetic)

A professional services group reviews client contracts under NDAs across EU jurisdictions.

Workflow slice	Data class	Placement	Why
Public clause library comparison	Internal general	Managed EU-zone API	Low sensitivity; freshness matters
Client-specific draft under NDA	Confidential commercial	Private managed endpoint	Contractual boundary; audit rights
Litigation strategy notes	Regulated / high consequence	Sovereign or air-gapped	Human gate; minimal retention

On the P&L: one programme, three placements, with unified identity and eval harness, not three orphan pilots.

Private managed endpoint (plain English): vendor runs the model in an isolated environment (dedicated VPC, customer-managed keys) with stronger boundaries than public multi-tenant APIs. **On the P&L:** higher unit cost than public API, often lower than owning GPUs you idle half the time.

Worked example: HR employee queries (synthetic)

An HR team wants a policy assistant for leave, benefits, and payroll FAQs.

Wrong answer: public consumer tool because "it's just HR." Personal data, wrong retention, no audit trail.

Right answer: data class 4 (regulated personal data) maps to private endpoint or sovereign self-hosted weights (for example a sized DeepSeek-V4-Flash or mid-tier open model with evaluation parity); logging and retention defined; training-use prohibited in contract; evaluation parity when model version changes. Prefer self-hosted open weights over any consumer or foreign-hosted API for HR personal data.

On Monday: one visible HR incident from shadow AI costs more than a year of private endpoint fees.

Talent and continuity

Sovereign inference without MLOps depth becomes a museum of outdated weights. Ask:

- **Who patches CUDA and runtime?** Name the team on call for driver and framework updates, because unpatched local stacks are a common source of worse security than hardened managed endpoints. A local model on an unpatched runtime is a vulnerability with an inference endpoint, and the patching cadence is what separates sovereignty from exposure. The name on the on-call rota is the difference between a control and a risk.
- **Who evaluates after each weight upgrade?** Model swaps require rerun of frozen holdouts before promotion; without that owner, quality drifts silently while the board still believes it is "sovereign and safe." Evaluation after upgrade is the control that keeps the local path honest, and it is labour that managed vendors fold into their cadence. Without a named owner, the upgrade ships untested and the drift begins.
- **Who responds when the GPU node dies at 2 a.m.?** Document on-call rotation and spare capacity, since sovereignty that cannot survive a hardware failure is availability risk concentrated indoors. A single node failure that takes the workflow down until morning is an availability story the board will hear about from users, not from monitoring. Spare capacity is the cost of sovereignty that finance must see in the TCO.
- **What is the spare capacity plan?** State reserved headroom for failover and batch peaks, because idle GPUs are expensive but zero spare capacity makes sovereign inference fragile under load. The plan is the answer to the capacity question finance asks when the GPU bill arrives. Idle capacity is not waste; it is the insurance that makes sovereign inference production-grade rather than demonstration-grade.

If answers are vague, you do not have sovereignty. You have concentration risk indoors.

Contract clauses that matter for managed paths

- **Training use of prompts and outputs.** Contract must prohibit vendor training on customer content by default, with opt-in only where legal and security explicitly approve a narrow exception. The training-use clause is the line between a managed service and a data pipeline that feeds someone else's model. Default-off with a named approval is the posture that keeps the clause enforceable rather than aspirational.

- **Retention periods.** Define maximum retention for prompts, completions, and logs so GDPR and sector rules have a contractual backstop, not only a settings page operators never read. A settings page is a configuration that drifts; a contract clause is an obligation that holds. The backstop is what makes retention a legal control rather than an operational hope.
- **Subprocessors and regions.** List every subprocessor and processing region so data-class mapping can be enforced and updated when vendors add new regions or affiliates. The list is the artefact that lets legal verify the placement still satisfies the data-class policy after a vendor reorganisation. An unmapped subprocessor is a placement change the firm did not consent to.
- **Audit rights.** Reserve the right to audit or receive third-party attestations covering the inference path used for regulated or confidential workloads. Audit rights are the contractual foundation of the trust the firm places in the managed path, and without them the firm is relying on vendor assertions. The attestation is the evidence an auditor will ask for when the placement is challenged.
- **Exit assistance and log export.** Require bulk export of logs, configs, and connector documentation within a defined window if the contract ends or prices change materially. Exit assistance is what makes a managed dependency reversible, and the defined window is what makes it operationally usable. Without it, leaving a vendor is a project with no deadline and no cooperation.
- **Incident notification timelines.** Set maximum hours to notify the customer of breaches affecting inference data, aligned with your own regulator and customer contractual obligations. The timeline is the bridge between the vendor's incident response and the firm's regulatory clock. A misaligned timeline means the firm learns of a breach after its own notification deadline has passed.

Appendix: placement decision record

Record one decision per material workflow. Leave the Entry column blank until the placement is agreed.

Field	Entry
Workflow	
Data class	
Placement chosen	
Why not alternatives	
Owner	
Review date	
Exit trigger	

Plain read: empty cells are intentional. Placement without an exit trigger is a lock-in decision disguised as architecture.

Total cost of sovereignty (checklist)

Include:

- **GPU or equivalent capacity including idle.** Model reserved and burst capacity with realistic utilisation assumptions, because sovereign TCO fails when finance plans for one hundred per cent load that never arrives. Realistic utilisation on inference is often twenty to forty per cent, and the idle sixty is the cost of control. The idle line is the number that makes sovereign TCO honest.
- **Power, cooling, or private cloud instance cost.** Data-centre or dedicated cloud charges belong in the comparison, not only the accelerator purchase price on the capex slide. Power and cooling are recurring opex that scale with the hardware whether or not it is fully used. Omitting them is the error that makes sovereign look cheap on the slide and expensive on the invoice.

- **MLOps headcount fraction.** Allocate fractional FTE for patching, monitoring, and upgrades; without it, local inference becomes an unmaintained appliance within two release cycles. The FTE fraction is the people cost that managed vendors absorb into their service price. Without it, the sovereign estate decays on the cadence of the model market and nobody is maintaining it.
- **Evaluation labour per upgrade.** Budget analyst and engineer time to rerun holdouts whenever weights or runtimes change, matching the discipline managed vendors implicitly absorb into their cadence. Every upgrade is a quality question that costs hours to answer, and the cadence is months not years. This is the labour line that turns "we run local models" into "we run local models we trust."
- **Downtime risk allowance.** Quantify revenue or SLA cost of sovereign outages without vendor failover, especially for workflows that cannot fall back to managed paths by policy. A sovereign outage is an incident the firm owns end to end, with no vendor to call. The allowance is the actuarial cost of that ownership, and it is the number that separates a TCO from a wish.
- **Delayed feature cost versus managed freshness.** Estimate opportunity cost when local models lag vendor capability by quarters, affecting quality on tasks competitors already run on newer endpoints. The lag is a quality cost that shows up in user satisfaction and competitive parity, not in the hardware invoice. It is the cost of being behind the vendor cadence on the tasks that matter.

Compare to managed private endpoint cost over 24 months, not month one. Many "sovereign saves money" claims fail this comparison.

Security myths

Myth: local always safer. **Reality:** unpatched local stacks and broad employee access can be worse than a hardened private endpoint.

Myth: managed always leaks training data. **Reality:** contractual terms vary; read them; enforce data classes.

24-month TCO comparison fields

Cost element	Managed public	Private managed	Sovereign/local
Usage			
Reserved capacity / idle			
Platform engineering			
MLOps / patching			
Evaluation after upgrades			
Downtime risk allowance			
Exit / migration			

Approve sovereignty only when the TCO and control gains beat private managed options for the data class in scope.

Hybrid control plane

Identity, evaluation harnesses, prompt/version registries, and cost tags should be consistent across placements. Otherwise the estate becomes three AI programmes that cannot be governed together.

Evaluation parity, briefly: the same proposition tests and holdouts run after every model or placement change, so quality does not silently diverge between environments. **Architecture case:** when you upgrade a local open weight (DeepSeek-V4, GLM-5.2, Qwen3.x, or similar) or the vendor ships a new GPT-5.x / Claude / Gemini revision, rerun the frozen holdout before promoting.

Evidence base: placement is an operating and economics decision

Source	Finding	Implication
Stanford AI Index 2026	Industry produced over 90% of notable frontier models in 2025; organisational adoption high while agent use remains early	Most firms will consume external model markets; placement and terms matter more than training frontier models
FinOps Foundation 2026	98% of surveyed practitioners manage AI spend; private cloud and data-centre spend also increasingly in FinOps scope	Sovereign options must be costed with idle capacity and ops labour, not GPU sticker price alone
NIST AI 600-1	Inventory should include underlying foundation models, versions, access modes, provenance, and supplier considerations	Placement choices belong in the system inventory
EU AI Act	Obligations depend on role and risk; documentation and oversight expectations rise with consequence	Placement must map to data class, residency, and audit needs
Eurostat 2025	AI use remains minority among EU enterprises overall, but majority among large firms	Mid-market firms should avoid ideological sovereignty that they cannot operate

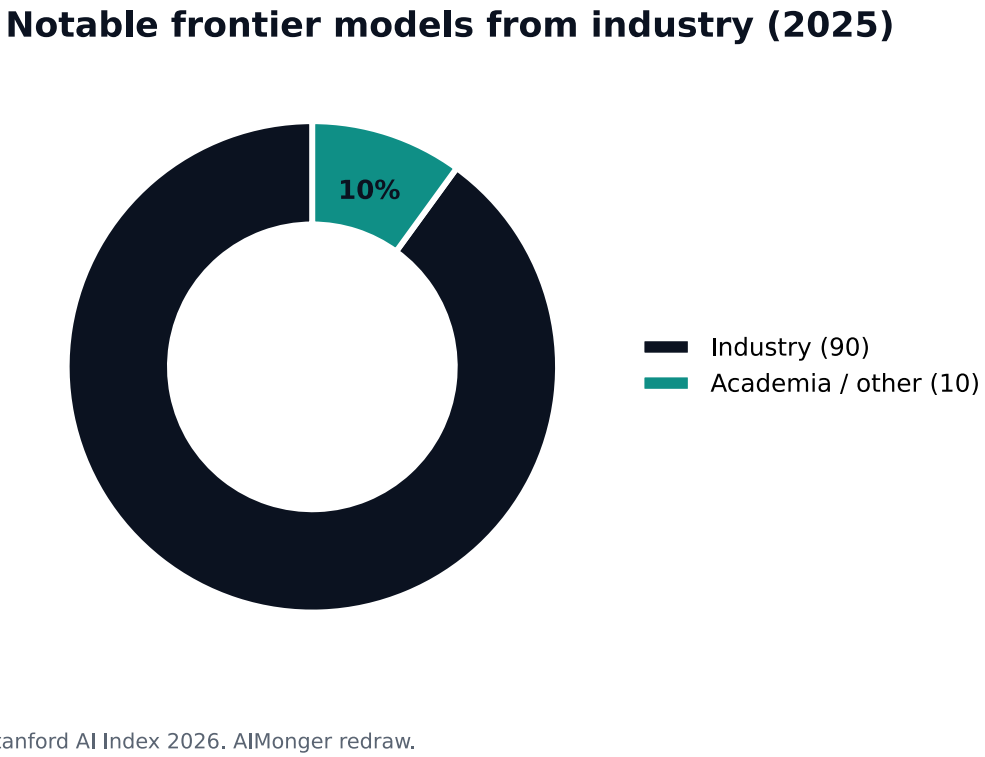


Figure 2. Share of notable frontier models from industry vs other sources (Stanford AI Index 2026). In short: buying inference is the default path; sovereignty is about control of that consumption, not building foundation models in-house. **On Monday:** your build budget should focus on eval, integration, and data, not competing with hyperscaler training runs. Source: Stanford AI Index 2026. AIMonger redraw.

Methodology note

"Sovereign" here means operational control of inference infrastructure and data path, not legal immunity. Managed private endpoints can be more controlled than poorly patched local stacks. Compare total cost of ownership over 24 months, including evaluation labour after each model upgrade.

Counter-position: only local inference is safe

Local can reduce certain multi-tenant risks. It can also concentrate patching, identity, and availability failures indoors. Safety is the product of controls, not location alone. Map data classes to allowed placements and enforce exit clauses for managed providers.

Placement decision record (required)

Workflow, data class, placement, alternatives rejected, owner, review date, contractual training-use terms, exit trigger, evaluation parity process.

Research addendum: sovereignty is a control bundle, not a slogan

An OECD working paper on measuring domestic public-cloud AI compute (October 2025) explicitly declines a single universal definition of sovereign AI compute because national objectives vary across privacy, security, energy, cybersecurity, and industrial policy. Buyers should therefore decompose "sovereignty" into testable controls:

- **Storage location.** Where prompts, outputs, and logs persist at rest, and whether replication crosses borders the data-class policy forbids. Storage location is the most visible control and the one most often assumed rather than verified. Replication to a disaster-recovery region can quietly cross a border the policy forbids, and only an explicit replication check catches it.
- **Processing location.** Where inference compute runs during the request, distinct from where backups or support tooling later copies the data. Processing and storage can be in different regions, and a placement that stores in-region can still process elsewhere under some vendor modes. The distinction is what makes "data stays in the EU" a checkable claim rather than a slogan.
- **Operator jurisdiction.** Which legal regime governs the operator and support staff who can access systems, mattering for subpoena and sector rules beyond physical server geography. A server in Frankfurt operated by a vendor whose parent is in another jurisdiction carries a different legal exposure than a server operated by a domestic firm. Jurisdiction is the control that physical location alone cannot establish.
- **Support and access path.** How vendor support reaches the environment, whether through jump boxes, break-glass accounts, or an offshore NOC, because sovereignty fails if support paths bypass your boundary controls. A break-glass account that a support engineer can open without notice is a hole in the boundary the firm cannot audit. The support path is the back door, and it needs the same scrutiny as the front.
- **Encryption and key control.** Whether customer-managed keys and encryption in transit and at rest are contractually guaranteed, not merely marketing checkboxes on a shared tenant. Key control is the difference between encryption the vendor can undo and encryption the firm alone holds. The contractual guarantee is what makes the checkbox a control rather than a feature flag.
- **Model ownership and portability.** Rights to export weights, adapters, and configs if you exit, so sovereign investment is not trapped in a vendor-specific runtime format. Portability is what makes a sovereign build a reusable asset rather than a sunk cost in a specific runtime. Without export rights, leaving the vendor means rebuilding the model from scratch.
- **Continuity and exit rights.** Documented migration assistance, data export windows, and failover if the provider exits a region or deprecates an endpoint your workflow depends on. Continuity is the control that turns a vendor dependency into a managed risk, and the export window is the deadline that makes migration operationally feasible. Without it, a vendor deprecation is a board incident.

Data residency is not GDPR compliance. GDPR Chapter V regulates transfers to third countries; keeping inference in the EU does not by itself establish lawful purpose, minimisation, retention, or processor terms. Conversely, lawful transfers with appropriate safeguards can exist without every token staying physically in-region.

Managed providers document distinct modes: global pooled processing, geographic/EU-zone processing, and single-region processing. Global endpoints generally do not provide regional processing guarantees. Choose per data class, not as ideology.

Gartner has also forecast that power availability may operationally constrain a substantial share of AI data centres by 2027. Capacity and energy scenarios belong in long-horizon placement decisions alongside legal geography.

Ninety-day placement charter (board-ready)

Phase	Activities
Days 1 to 30	Publish the data-class map and forbidden paths. Inventory all production inference endpoints.
Days 31 to 60	Fill a 24-month TCO worksheet for one regulated and one general workflow. Negotiate training-use and retention clauses.
Days 61 to 90	Run evaluation parity across placements. Record a placement decision per workflow. Set exit trigger dates.

Investment committee approves sovereign capex only when private managed paths fail the control test, and TCO wins on a twenty-four-month view including idle GPU and MLOps labour.

Choose inference placement like you choose banking jurisdictions: deliberately, by data class, with owners and exit plans.

Ideology is not a control. An operating decision is.

References

1. Stanford Institute for Human-Centered Artificial Intelligence, “The 2026 AI Index Report,” Economy chapter. https://hai.stanford.edu/ai-index/2026-ai-index-report/economy
2. FinOps Foundation, “State of FinOps 2026 Report.” https://www.finops.org/insights/state-of-finops/
3. NIST, “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,” NIST AI 600-1, July 2024. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.600-1.pdf
4. Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj
5. Eurostat, “20% of EU enterprises use AI technologies,” 11 December 2025 (20.0% in 2025; 13.5% in 2024). https://ec.europa.eu/eurostat/en/web/products-eurostat-news/w/ddn-20251211-2
6. Gartner, “Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026,” press release, 11 October 2023. https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026
7. McKinsey & Company / QuantumBlack, “The State of AI: Global Survey 2025.” https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

8. OECD, “Measuring Domestic Public Cloud Compute Availability for Artificial Intelligence,” October 2025. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/10/measuring-domestic-public-cloud-compute-availability-for-artificial-intelligence_39fa6b0e/8602a322-en.pdf

9. Regulation (EU) 2016/679 (GDPR), Chapter V / Article 44. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504

10. Gartner, press release on AI data-centre power constraints, 12 November 2024. https://www.gartner.com/en/newsroom/press-releases/2024-11-12-gartner-predicts-power-shortages-will-restrict-40-percent-of-ai-data-centers-by-2027

11. CNBC, “China’s Zhipu is booming with Anthropic and OpenAI held back” (26 June 2026). https://www.cnbc.com/2026/06/26/china-zhipu-z-ai-open-source-anthropic-openai.html

12. CNBC, “Chinese AI models gain ground with U.S. companies as costs surge” (7 July 2026). https://www.cnbc.com/2026/07/07/chinese-ai-models-costs-us-openai-anthropic.html

13. DeepSeek, “DeepSeek V4 Preview Release” (24 April 2026). https://api-docs.deepseek.com/news/news260424/

14. Z.ai / Zhipu, “GLM-5.2: Built for Long-Horizon Tasks” (16 June 2026). https://z.ai/blog/glm-5.2

Published by AIMonger . https://aimonger.com . Malta . Europe