MARKET RESEARCH BRIEFING

AI Sovereignty for the European Enterprise

Regulation, Economics, and the Build-vs-Hyperscaler Decision

Why "sovereignty" has become a board-level procurement criterion, why the word bundles five different things with five very different price tags — and where the honest answer, despite the politics, is still the hyperscaler

An evidence briefing for C-level executives, CIOs, CTOs, and strategy and risk leaders at enterprises headquartered outside the United States — with the EU as the primary lens and the wider EEA, the UK, and Switzerland as secondary reference points. Figures current to early July 2026. July 2026.

1. Executive Summary

Three years ago, "AI sovereignty" was a phrase for ministers and conference panels. In 2026 it shows up in enterprise RFPs, in public-tender qualification gates, and on board risk agendas. No single event caused the shift. A US legal regime that reaches data wherever a US-controlled company stores it; a European regulatory wall arriving in layers, each with a date attached; a series of political shocks that made the dependence tangible; and, above all, AI workloads that concentrate a company's most sensitive data in whichever infrastructure runs the model. Each played its part. The difficulty is that "sovereignty" is sold as one thing and priced as five. This briefing separates those layers, puts numbers on each, and tests the business case from both sides: against sovereignty rhetoric that ignores total cost of ownership and capability gaps, and against hyperscaler "sovereign cloud" offerings whose legal immunity is asserted rather than proven.

The thesis in one paragraph

The regulatory pressure is real, datable, and growing, but very little of it legally *requires* sovereign infrastructure for private-sector workloads today; much of what gets marketed as compliance is in fact risk posture. The sovereign supply side has finally materialized: AWS opened a dedicated European cloud in January 2026, European providers and model builders are shipping, and EU compute programs are under construction. It remains small, more expensive, and behind on capability. Meanwhile the economics of building in-house are unforgiving below roughly two-thirds sustained utilization, and every quarter of falling inference prices works against owned silicon. For most European enterprises the rational posture is therefore workload-level triage rather than company-level ideology: full sovereignty for the few workloads where law or survival demands it, hyperscalers with hard controls and a rehearsed exit for everything else, and candor about which is which.

What the evidence says (sourced)

- **Regulation keeps arriving, and keeps slipping.** The EU AI Act entered into force in August 2024; general-purpose AI rules applied from August 2025; transparency duties and the penalty regime arrive in August 2026. The “Digital Omnibus,” agreed in May 2026 and given final approval by Parliament (16 June) and Council (29 June 2026), pushes high-risk obligations out to December 2027 and August 2028 — a delay of more than a year (European Commission; Gibson Dunn; DLA Piper, 2026).
- **The legal core of the problem is jurisdiction, not geography.** The US CLOUD Act attaches to the corporate entity, wherever the data sits; FISA §702 adds programmatic intelligence collection. In June 2025, Microsoft’s counsel told the French Senate under oath that the company could not guarantee EU-held data against US access. The EU–US Data Privacy Framework survived its first court test in September 2025 (*Latombe*), but noyb announced a fresh challenge in June 2026, and the framework still rests on a revocable US executive order (CJEU/General Court; noyb, 2026).
- **The market is responding on both sides of the Atlantic.** AWS launched its European Sovereign Cloud in Brandenburg in January 2026 with a €7.8 billion commitment; Microsoft and Google now sell dedicated sovereign lines. European providers, meanwhile, still hold about 15% of their home market, down from 29% in 2017, while the top three US hyperscalers hold about 70% (Synergy Research, 2025).
- **Sovereignty carries a measurable premium and measurable momentum.** Hyperscaler sovereign offerings price up to 30% above standard regions (BCG, 2025); Delos Cloud is reported around 15% above Azure. Adoption is accelerating anyway: roughly 40% of European organizations used some sovereign cloud solution in 2025 and another 31% plan to (IDC, 2026). In Germany, 82% of firms say they no longer want US-cloud dependence, yet 65% will not accept functional limitations to escape it (Bitkom, 2025).
- **The build case clears only at high utilization.** Our TCO model (Section 5) puts a 64-GPU H100-class cluster in Europe at about $3.86 per GPU-hour all-in at 70% utilization. It beats hyperscaler on-demand pricing above roughly 38% utilization, beats committed

pricing only above 66%, and never beats specialist GPU clouds under base-case assumptions.

- **Token deflation quietly undermines capex cases.** The price of a fixed level of model capability has fallen 9x–900x per year depending on the task (Epoch AI, 2025). What an enterprise actually pays falls more slowly but still relentlessly: our companion briefing projects blended enterprise inference declining from $4.75 to about $1.57 per million tokens by 2030, roughly −24% per year, with frontier prices holding a persistent premium (StratoFactory, AI Token Cost Forecast 2026–2030, May 2026). Owned hardware has to outrun that curve to pay back.
- **The model gap re-widened but stays small; the platform gap is bigger.** The best closed model led the best open-weight model by 0.5% in August 2024 and by 3.3% in March 2026 (Stanford HAI AI Index, 2026). For bounded enterprise workloads, open-weight models are already good enough; for frontier reasoning and agentic work they are not. And hyperscalers sell more than inference: mature, continuously developing environments for building business AI, from data platforms and security integration to MLOps and agent tooling, which any sovereign alternative must replicate or forgo.

Headline predictions (2026–2029). (1) European providers will still hold under 20% of the European cloud market in 2029; the sovereign wave will be captured mostly by the hyperscalers' own sovereign lines. (2) The Data Privacy Framework will not survive to 2029 in its current form, and its failure will trigger the first board-mandated repatriation programs at scale. (3) Open-weight models will become the default for new GenAI deployments in EU regulated sectors. (4) Rented GPU prices will keep falling and at least one European enterprise AI cluster will be publicly written down. (5) The EU gigafactory program will deliver less than announced, later than announced. (6) CADA will pass in diluted form, leaving US hyperscalers access to most EU public cloud spend. Section 9 states each call in falsifiable form, with the evidence that would prove it wrong. Beneath all six sits one working estimate: the workloads that genuinely need EU-controlled infrastructure amount to perhaps 10–20% of enterprise AI spend, not the majority that sovereignty marketing implies.

2. What "Sovereignty" Actually Means — a Taxonomy of Five Layers

The first step is to stop using "sovereignty" as one word. In enterprise procurement it bundles at least five distinct properties. They differ in what they protect against, what they cost, and who needs them. Vendors profit from the confusion: a data-residency guarantee is routinely marketed with the emotional charge of jurisdictional immunity, which it does not provide.

- **Layer 1 — Data residency.** Your data is physically stored and processed in the EU. Cheap and nearly universal; every hyperscaler offers EU regions and an EU data boundary. What residency does not do is shield data from the legal reach of a non-EU parent company. Geography is not jurisdiction.

- **Layer 2 — Jurisdictional immunity.** No non-EU authority can lawfully compel access, which in practice requires that no non-EU entity controls the operator. This is the layer France's SecNumCloud 3.2 actually enforces (non-EU ownership capped at 24% individually, 39% collectively; ANSSI). On the prevailing reading of the CLOUD Act, no US-parented offering can deliver this layer regardless of how the subsidiary is structured; the point is contested, and Section 3 examines it.
- **Layer 3 — Operational autonomy.** The service keeps running if the foreign vendor is sanctioned, pressured, or simply switches you off: EU-resident staff, local support, technical continuity, no remote "kill switch." The ICC email episode of May 2025 made this concrete, and the proposed EU Cloud and AI Development Act now names it explicitly.
- **Layer 4 — Model and IP control.** You control the model weights, the fine-tuning data, and what leaves the building in prompts and telemetry. Achievable today with open-weight models, at a capability and pace-of-innovation cost that Section 6 quantifies.
- **Layer 5 — Compute and supply-chain independence.** The chips, interconnects, and fabs. At meaningful scale this layer is not for sale in Europe at any price this decade. The accelerator supply chain is US-designed (Nvidia, AMD) and Asia-fabricated (TSMC), with ASML the lone European chokepoint, and even Europe's flagship EuroHPC machines run on American silicon. Sensible sovereignty planning treats Layer 5 as a constraint to manage, not an option to buy.

Layer	What it guarantees	Typical cost of the layer	Who actually needs it
1. Data residency	EU storage and processing; EU data boundary	≈ zero — standard on every hyperscaler	Everyone; table stakes, not sovereignty
2. Jurisdictional immunity	No non-EU authority can lawfully compel access (EU-controlled operator)	+15–30% vs hyperscaler pricing; thinner service catalog	Sensitive public sector; SecNumCloud-scoped suppliers; top tier of regulated data
3. Operational autonomy	Service survives foreign sanctions or vendor coercion; no remote kill switch	Overlaps Layer 2; continuity engineering and exit rehearsal	Critical infrastructure; institutions; defense-adjacent enterprises
4. Model & IP control	Weights, fine-tuning data, prompts and telemetry stay in-house	Capability gap (Sec. 6) plus MLOps payroll (Sec. 5)	IP-intensive R&D; regulated decisioning; high-volume stable workloads
5. Compute / supply-chain independence	Non-US-controlled chips, interconnects, fabs at scale	Not purchasable in Europe this decade	No enterprise — an EU policy aspiration (EuroHPC, gigafactories)

Table 1. Five layers hiding inside one word. Each layer protects against a different threat at a very different price; every sovereignty claim in a vendor pitch should be forced to name its layer. Our taxonomy; cost indications per Sections 4–6.

Our analysis: most European enterprises, for most workloads, need Layer 1 plus contractual and technical controls, which they already have. A regulated minority genuinely needs Layers 2–3, for a minority of workloads. Layer 4 is a strategic choice that gets cheaper every quarter.

Layer 5 is a fact of geopolitics. Every euro of sovereignty premium should be traceable to a named layer and a named threat; “sovereignty” without a layer number is marketing.

3. The Regulatory Landscape: What Is Required vs. What Is Posture

European boards tend to experience regulation as a single rising tide. The instruments differ sharply in what they actually compel, and the distinction matters throughout this section: some rules *require* something specific, with dates, duties, and penalties, while others merely *reward* sovereignty postures without mandating them for the private sector.

The compliance wall arrives in layers, 2024–2028

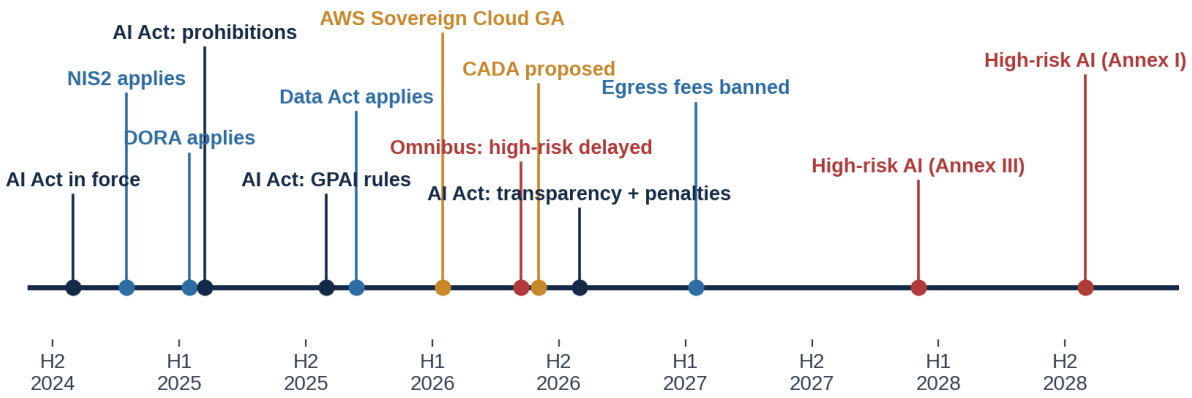

Figure 1. The compliance wall arrives in layers, 2024–2028: EU AI Act milestones (navy), data and resilience rules (blue), sovereignty-market events (gold), and the May 2026 Omnibus deferrals of high-risk obligations (red). Dates: European Commission, EuroHPC, AWS; Omnibus agreement per Gibson Dunn / Latham & Watkins (2026). CADA is a proposal, not law.

The AI Act: real dates, moving dates

The AI Act entered into force on 1 August 2024. Prohibited practices and AI-literacy duties applied from February 2025. Obligations for general-purpose AI (GPAI) models — technical documentation, copyright policy, training-data summaries, and, for systemic-risk models, evaluations and incident reporting — applied from 2 August 2025. On 2 August 2026 the governance and penalty regime becomes applicable, with fines of up to 7% of global turnover for prohibited practices, alongside the Article 50 transparency duties. The headline change of 2026 came from the “Digital Omnibus” political agreement of 7 May: obligations for stand-alone high-risk systems (Annex III — recruitment, credit scoring, essential services) are deferred to 2 December 2027, and for AI embedded in regulated products (Annex I) to 2 August 2028 (Gibson Dunn; Travers Smith, 2026). The European Parliament endorsed the package on 16 June 2026, the Council gave final approval on 29 June, and publication in the Official Journal, the last formal step, was imminent as this briefing went to press.

What the AI Act does *not* do matters just as much. It does not require European infrastructure, European models, or data localization. An enterprise can be fully AI Act-compliant running OpenAI on Azure in Virginia. The Act shapes how you deploy AI, not where, and conflating AI Act compliance with sovereign infrastructure is the most common mistake in current vendor marketing.

Data transfers: lawful today, fragile by design

GDPR itself is location-agnostic; transfers to the US are lawful under the EU–US Data Privacy Framework (DPF) adequacy decision of July 2023. The fragility is structural. The CJEU has already struck down two predecessor frameworks, Safe Harbor in 2015 and Privacy Shield in 2020 (*Schrems II*). The General Court upheld DPF adequacy in *Latombe v Commission* in September 2025, but the deeper objections were never resolved: the DPF rests on US Executive Order 14086, which any administration can revoke without Congress, and on oversight bodies whose independence has been eroding. In June 2026 noyb announced a fresh challenge arguing exactly that, and a CJEU ruling is plausibly 18–30 months out. So the legal position is: transfers are lawful today; the mechanism has failed twice before; and every large enterprise ran this fire drill in 2020 and remembers what it cost.

US extraterritorial law: the part residency cannot fix

The US CLOUD Act (2018) obliges any provider subject to US jurisdiction to produce data in its "possession, custody, or control" regardless of where the data is stored. Jurisdiction attaches to the corporate group, not the server. FISA §702 adds programmatic foreign-intelligence collection through providers. Two events made this concrete for European boards in 2025. First, Microsoft's chief legal officer for France, testifying under oath before the French Senate in June, declined to guarantee that French data could never be handed to US authorities. Second, after US sanctions against the International Criminal Court, the ICC chief prosecutor's Microsoft-hosted email account stopped working in May 2025. Microsoft denies suspending services; the ICC moved off Microsoft Office anyway, and the episode triggered formal dependence reviews across Dutch and other EU public institutions. Whether a warrant against a US parent could actually reach data held in a structure like AWS's European Sovereign Cloud is contested and untested — no CLOUD Act order against a hyperscaler's EU sovereign entity has been litigated to judgment. What is settled: GDPR Article 48 makes a US court order alone an invalid transfer basis, which leaves US-parented providers in a standing conflict of laws that no data center location can resolve.

The operational layer: Data Act, DORA, NIS2

Three instruments change the switching calculus more than any sovereignty speech. The **Data Act** became applicable on 12 September 2025: cloud switching rights apply to new contracts now, extend to legacy contracts from September 2026, and from 12 January 2027 egress and switching charges are banned outright, which materially lowers the cost of the exit option Section 5 prices. **DORA** (applicable January 2025) requires financial entities to maintain exit

strategies for critical ICT providers; in November 2025 the European Supervisory Authorities designated the first critical third-party providers, the hyperscalers among them, for direct EU oversight. Concentration risk is now official doctrine rather than commentary. **NIS2** (enforceable since October 2024, with national transposition still uneven in 2026) extends hardened security and incident-reporting duties across essential sectors. None of the three mandates sovereign infrastructure. All three make dependence auditable and exit mandatory to plan.

National schemes and the coming EU framework

The rules that genuinely *require* sovereignty live at national level and bind mostly the public sector and its suppliers. France's **SecNumCloud 3.2** qualification (ANSSI) demands EU control, with non-EU ownership below the 24%/39% thresholds, and is mandatory under the "Cloud au centre" doctrine for sensitive state data; roughly nine providers are qualified and about a dozen applications are pending (2026). Germany's **BSI C5** is an attestation standard without ownership tests — Microsoft, AWS, and Google all hold C5 attestations — while sensitive federal workloads route toward Delos Cloud, a Microsoft-technology cloud operated by SAP/Arvato. At EU level, the certification scheme **EUCS** has been stuck for four years on precisely this question: whether "high assurance" should include immunity from non-EU law. France, Italy, and Spain say yes; the Netherlands, Ireland, and the Nordics say no (cep, 2025). That deadlock explains the June 2026 move: the Commission's proposed **Cloud and AI Development Act (CADA)**, adopted as a proposal on 3 June 2026, would create a four-level sovereignty framework for public procurement, running from EU data processing (Level 1, around 70% of public workloads) to full EU ownership and control (Levels 3–4, with an explicit anti-kill-switch requirement, reserved for a single-digit share of the most sensitive workloads). CADA now enters the legislative grinder. Its final shape will determine how much of the European public market US-parented sovereign offerings can actually address.

Instrument	Status (early July 2026)	What it actually compels
EU AI Act	In force since Aug 2024; GPAI rules live since Aug 2025; penalties + transparency from 2 Aug 2026; high-risk deferred to Dec 2027 / Aug 2028 (Omnibus approved by Parliament 16 June and Council 29 June 2026; OJ publication imminent)	Risk-tiered process compliance wherever hosted. No localization, no European infrastructure requirement
GDPR + EU–US DPF	DPF valid; upheld in Latombe (Sept 2025); fresh noyb challenge announced June 2026; CJEU ruling plausibly 2027–28	A lawful transfer mechanism; Art. 48 invalidates direct foreign court orders as a transfer basis
US CLOUD Act / FISA §702	Settled US law; conflict with GDPR unresolved; sovereign-subsidiary structures untested in court	US-jurisdiction providers must produce data in their control wherever stored

Instrument	Status (early July 2026)	What it actually compels
Data Act	Applicable since 12 Sept 2025; legacy contracts covered from Sept 2026; egress/switching fees banned from 12 Jan 2027	Switching rights, contract terms, egress-fee ban — cheapens the exit option
DORA / NIS2	DORA live since Jan 2025; hyperscalers designated critical ICT providers Nov 2025. NIS2 enforceable; transposition still uneven	Exit strategies, resilience testing, oversight, incident reporting — dependence becomes auditable
SecNumCloud (FR) / C5 (DE)	SecNumCloud 3.2: ~9 qualified providers, ~12 pending. C5: attestation held by US hyperscalers too	FR: EU control (≤24%/39% non-EU ownership) mandatory for sensitive state data. DE: security baseline only
EUCS / CADA	EUCS stalled since 2022 over sovereignty tiers; CADA proposed 3 June 2026 (4-level framework, anti-kill-switch)	Nothing yet — but will define how much EU public procurement US-parented clouds can address

Table 2. The regulatory stack, separated into what is settled, pending, and contested. Sources: European Commission; ANSSI; BSI; ESAs; Gibson Dunn / Latham & Watkins / Travers Smith / DLA Piper (2026); cep (2025). Status descriptions current to early July 2026.

The bottom line on "required." For a private-sector enterprise in 2026, hard law requires AI Act process compliance wherever the workload is hosted, lawful transfer mechanisms, DORA/NIS2 resilience and exit planning, and sector rules for health and public-sector data in several member states. Hard law does not require sovereign infrastructure. The pull toward sovereignty is posture — sometimes prudent, occasionally existential, but posture — plus the anticipation that public procurement (SecNumCloud today, CADA tomorrow) will make sovereignty a market-access condition for anyone selling into European governments. Enterprises should write down which of the two is driving each decision, because the two price very differently.

4. The Market Response: What Is Real, What Is Announced

The sovereignty debate has finally produced a supply side. The task now is to separate operating reality from press releases, on both sides of the Atlantic. The starting point is uncomfortable for Europe, because the market structure has not moved in years.

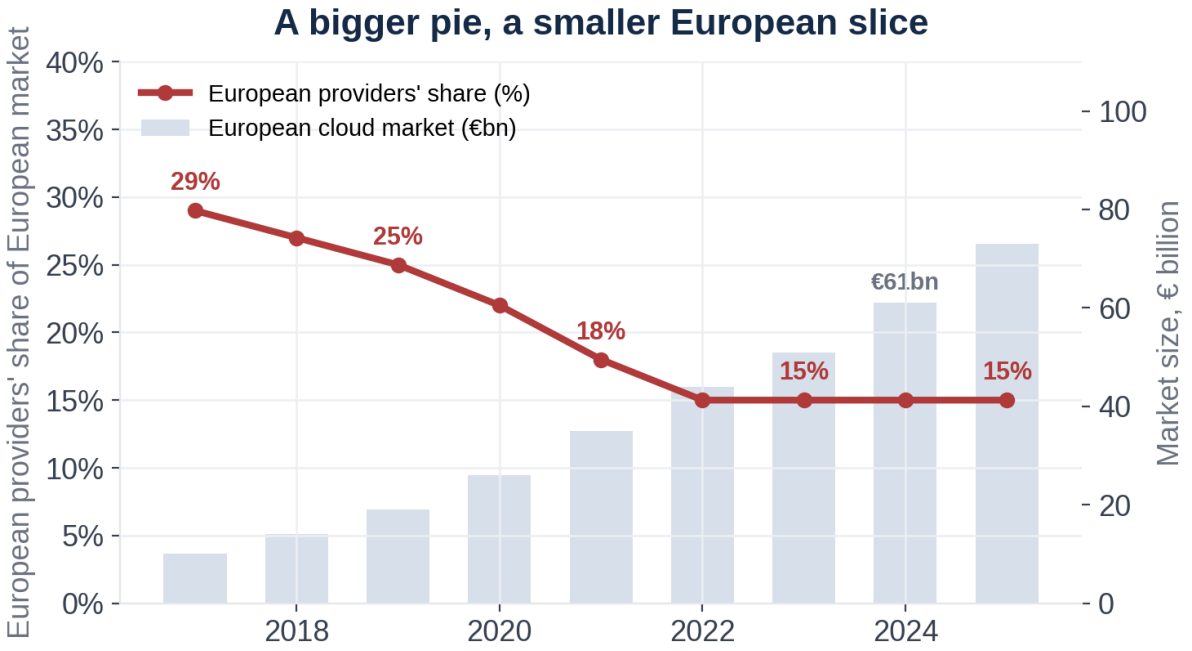

Figure 2. A bigger pie, a smaller European slice: the European cloud market grew roughly six-fold since 2017 (bars, € billion; 2025 estimated), while European providers' share fell from 29% to ~15% and has been flat since 2022. The top three US hyperscalers hold ~70% of the European market. Source: Synergy Research Group (2024–2025); market size 2024 = €61bn.

The hyperscalers' sovereign turn

- **AWS European Sovereign Cloud is the most substantive move.** Generally available since 15 January 2026, launch region Brandenburg; €7.8 billion committed; a legally separate German-incorporated operation with EU-resident leadership (Stéphane Israël, appointed October 2025), EU-only operational personnel, and further regions planned in the Netherlands, Belgium, and Portugal (AWS disclosures, 2026 — vendor claims). It is a real, physically and organizationally distinct cloud. Its parent remains Amazon.com, Inc., which is where the CLOUD Act attaches. AWS asserts independence; no court has tested it (InfoQ; KuppingerCole, 2026).
- **Microsoft treats sovereignty as a product line.** Announced June 2025: a "Sovereign Public Cloud" (EU data boundary, EU-staffed operations across all EU regions), a "Sovereign Private Cloud" (Azure Local on customer premises), and national partner clouds — Bleu in France (Capgemini/Orange), Delos in Germany (SAP/Arvato) — that run Microsoft technology under nominally independent operators. Delos reportedly prices about 15% above standard Azure (E-3 Magazine, 2025). The engineering investment is

genuine. The legal core is unresolved, and Microsoft's own Senate testimony sits awkwardly beside its marketing.

- **Google goes partner-led.** S3NS with Thales in France, pursuing SecNumCloud qualification; a sovereign hub in Germany; air-gapped and dedicated deployment options (2025–2026). The analysis is structurally similar: the partner layer adds real protection at the operational level, while the jurisdictional question follows the technology and the parent.

What do the hyperscaler sovereign offerings actually deliver? Layers 1 and much of 3 from Section 2's taxonomy: residency, EU staffing, operational continuity engineering, all backed by genuinely large investments. They also deliver the asset that rarely gets named in sovereignty debates — a mature and continuously developing environment for building business AI solutions, spanning managed data platforms, security and identity integration, MLOps and agent tooling, model catalogs, marketplaces, and certification portfolios. That platform depth, more than raw compute, is what the Europeans' 15% market share reflects, and it is why enterprises stay. What the sovereign offerings cannot demonstrably deliver is Layer 2, jurisdictional immunity. That claim will be settled by a court or a crisis, not by an architecture diagram, and enterprises paying the sovereign premium for immunity should understand they are buying a legal hypothesis.

The European providers: real, profitable, small

OVHcloud crossed €1 billion in annual revenue in FY2025 (€1,084.6m, up 9.3% year on year, at a 40.4% adjusted EBITDA margin, per company disclosure). Deutsche Telekom/T-Systems and SAP hold roughly 2% of the European market each; OVHcloud, Telecom Italia, and Orange follow (Synergy, 2025). Schwarz Group's StackIT is spending billions to become a European hyperscaler anchored in its own retail workloads. IONOS is listed and profitable. Scaleway (Iliad) is chasing AI workloads hard, and nine providers hold SecNumCloud qualification. The category is commercially real and strategically undersized: the largest European provider is roughly one-thirtieth the size of AWS in Europe. For a CIO the practical questions are capability coverage — managed services, AI tooling, global reach — and counterparty durability. Section 7 returns to both.

The EU programs: compute is arriving, coordination is not

The public-investment story improved markedly in 2024–2026. **JUPITER**, Europe's first exascale supercomputer, was inaugurated at Jülich in September 2025: a €500 million EU/Germany investment, ranked #4 on the TOP500 (EuroHPC, 2025). Nineteen **AI Factories** were selected across 2024–2025 to wrap EuroHPC compute in startup- and enterprise-facing services. **InvestAI** (February 2025) targets €200 billion of mobilized investment, of which €50 billion is public money, including €20 billion for four to five **AI gigafactories** of around 100,000 accelerators each. The informal call drew 77 expressions of interest, and the formal EuroHPC call is expected in summer 2026 (European Commission; EIB). The €200 billion should be read as a mobilization headline, not a budget line. The cautionary counter-example is **Gaia-X**, whose

federation vision dissipated into committee work and member exits; its successor project **IPCEI-CIS** (reported at €2.6–3.5 billion in public support, depending on the source — the range itself is telling) is real but slow. The pattern for planners: Europe has become good at funding compute hardware. It remains unproven at building competitive cloud services on top.

The European model builders

Mistral AI is the load-bearing asset of European model sovereignty. It carried a €11.7 billion valuation after its September 2025 ASML-led round; in June 2026 it was reported in talks to raise about €3 billion at around €20 billion (Bloomberg); in March 2026 it took an $830 million debt facility to buy 13,800 Nvidia chips for a Paris-region data center; and it has a stated 1 GW compute target for 2030. The Stanford AI Index 2026 reports Mistral models as the leading choice in EU enterprise deployments, a striking figure we treat with caution given definitional ambiguity, though it matches procurement behavior we observe. **Aleph Alpha** tells the other half of the story: having pivoted in 2024 from frontier-model competition to an enterprise AI operating system, it merged with Canada's Cohere in April 2026 in a transaction valued around $20 billion, backed by €500 million from Schwarz Group and blessed by both governments. The merger is an admission that even well-funded national champions cannot fund frontier training alone; it also created the largest sovereign-AI vendor outside the US and China. The deeper asset is the **open-weight ecosystem** (Mistral, Llama, Qwen, DeepSeek, Gemma), which turns model sovereignty from a heroic national project into a deployment decision any enterprise can make. Section 6 prices the capability cost of that decision.

Offering	Structure	What it demonstrably delivers	The open question
AWS European Sovereign Cloud	Legally separate German-incorporated operation; EU-resident leadership and staff; parent: Amazon.com, Inc.	GA since Jan 2026 (Brandenburg); €7.8bn committed; physical + organizational separation (vendor claims)	Whether separation defeats a CLOUD Act order at the parent — asserted, never litigated
Microsoft Sovereign Public / Private Cloud; Bleu (FR), Delos (DE)	EU data boundary + EU operations; national partner operators running Microsoft technology	Broad service catalog with EU controls; national certification tracks; Delos reported ~+15% vs Azure	Own Senate testimony conflicts with the immunity narrative; partners depend on Microsoft tech and licensing
Google sovereign lines / S3NS (Thales)	Partner JV in France pursuing SecNumCloud; air-gapped and dedicated options	Credible operational-autonomy engineering for the French market	Catalog depth and scale; the same parent-jurisdiction question
OVHcloud, T-Systems, StackIT, IONOS, Scaleway	EU-owned, EU-operated	Layer-2 jurisdictional immunity by construction; SecNumCloud qualification (subset); OVHcloud >€1bn revenue FY2025	Scale (~1/30 of AWS in Europe), managed-service and AI-tooling depth, counterparty durability

Offering	Structure	What it demonstrably delivers	The open question
EuroHPC, AI Factories, gigafactories	Public compute programs (JUPITER exascale #4 TOP500; 19 AI Factories; €20bn gigafactory envelope)	Real public compute for research and startups; call for gigafactories expected summer 2026	Enterprise access model, services layer, delivery timelines — and all of it runs on US silicon
Mistral; Cohere–Aleph Alpha; open-weight ecosystem	EU and transatlantic model suppliers; downloadable open-weight models	Layer-4 model control today; Mistral at €11.7bn valuation raising toward ~€20bn (reported June 2026)	Whether independent frontier pace is fundable in Europe at all — the Aleph Alpha pivot says no

Table 3. The sovereign supply side, mid-2026: structures as disclosed by vendors (flagged), open questions ours. Sources: AWS, Microsoft, Google, OVHcloud disclosures [vendor]; Synergy Research (2025); EuroHPC JU; Bloomberg, TechCrunch (2026).

5. The Economics: An Original Build-vs-Buy Model

Strip away the politics and the sovereignty decision is a capital-allocation problem under three uncertainties: how fully you will load the hardware, how fast rented intelligence gets cheaper, and how long the silicon stays competitive. This section builds the case from market prices upward. All model outputs are our own analysis; the assumptions are stated in full in Table 4 and stress-tested in Figure 5.

5.1 What a GPU-hour costs, depending on who sells it

The most under-appreciated fact in this market: the identical H100-class GPU-hour trades at a spread of roughly 12x in mid-2026, from about $1 on spot marketplaces to over $12 on hyperscaler on-demand list pricing (public price trackers, 2026). Specialist GPU clouds, the “neoclouds,” undercut hyperscaler GPU instances by 60–70% (Computer Weekly, 2025). Sovereign options price in the other direction, with hyperscaler sovereign lines carrying premiums of up to about 30% (BCG, 2025). Any TCO claim built on a single “cloud price” — and vendor pitches in both directions are built exactly that way — can be discarded on sight.

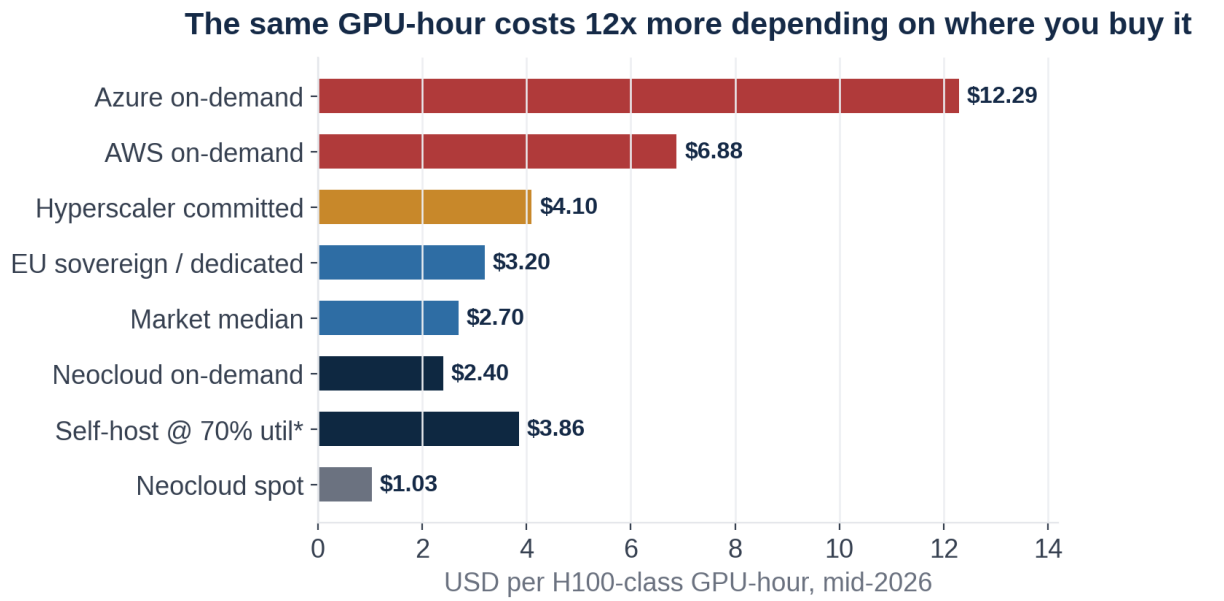

*Figure 3. The same GPU-hour costs 12x more depending on where you buy it: indicative H100-class per-GPU-hour pricing, mid-2026. Hyperscaler on-demand list prices (red) vs committed rates (gold), EU sovereign/dedicated indicative pricing (blue), neocloud on-demand/spot (navy/gray), and our modeled self-hosting cost at 70% utilization (*Section 5.2). Sources: public GPU price trackers and provider list prices (2026); self-host figure is our model output.*

5.2 The break-even model

We model a 64-GPU H100-class cluster operated in Europe, the scale at which a large enterprise runs serious inference and fine-tuning without pretending to train frontier models. Capital cost: $2.3 million including fabric, storage, and integration, straight-lined over three years. Operating cost: European colocation, electricity at €0.15/kWh with PUE 1.3, hardware maintenance, and a three-person MLOps team — the minimum credible staffing, and a cost line that build pitches routinely leave out. Table 4 lists every assumption. The model deliberately leans in the build's favor: it excludes cost of capital, software licensing, and downtime risk.

Parameter	Base case	Rationale / source
Cluster	64 × H100-class GPUs (8 nodes)	Enterprise inference and fine-tuning scale; no frontier-training pretensions
Capital cost	$30k per GPU + 20% fabric, storage, integration = $2.30m	Mid-2026 street pricing (public trackers); H100-class incl. node share
Depreciation	3-year straight line	Nvidia annual product cadence; stress-tested at 2 and 4 years (Fig. 5)
Power draw	0.7 kW per GPU (IT) × PUE 1.3	Typical H100 node loading; European colo PUE
Electricity	€0.15/kWh (swept €0.08–0.22)	European industrial range (IEA 2026; Eurostat 2025)
Colocation	€150 per kW-month	Indicative European metro colo
Operations	3 MLOps FTE at €150k loaded (€486k/yr with FX)	Minimum credible 24/7-capable team; routinely omitted in build pitches

Parameter	Base case	Rationale / source
Maintenance	5% of capex per year	Industry norm for accelerated hardware
Comparator prices	On-demand $6.90; committed $4.10; EU sovereign dedicated $3.20; neocloud $2.40 per GPU-hour	Public price trackers and list prices, mid-2026 (indicative)
Token-price planning curve	Blended $4.75/M tokens (2026) → ~$1.57/M (2030), −24%/yr central case; frontier −12.5%/yr, low-tier −25%/yr; Shock scenario +15% in 2027	StratoFactory AI Token Cost Forecast 2026–2030 (May 2026)
Deliberately excluded	Cost of capital, software licensing, downtime risk (worsen build case); resale value, tax treatment (improve it)	Model kept conservative in the build's favor on balance

Table 4. Every assumption behind Figures 4–5 and the token arithmetic in Section 5.4. Currency converted at €1 = $1.08. The model is our own analysis; it is illustrative and must not be quoted as measured market data.

The result is a utilization cliff. At 30% sustained utilization, which is optimistic for a first-year enterprise cluster, the machine costs about $8.81 per GPU-hour, worse than hyperscaler on-demand. At 50% it reaches $5.34; at 70%, $3.86; at 90%, $3.03. The break-even thresholds follow. Self-hosting beats hyperscaler on-demand pricing above roughly 38% utilization, a low bar that explains why naïve TCO decks always favor building. It beats committed pricing, which is what a competent CIO actually pays, only above about 66%. It beats an EU sovereign provider's dedicated pricing above 85%, and under these assumptions it never beats neocloud rates. Note the symmetry in that last result: the same arithmetic that says "don't build" also says the cheapest sovereign-leaning capacity is often a European specialist provider rather than a hyperscaler sovereign region at a premium.

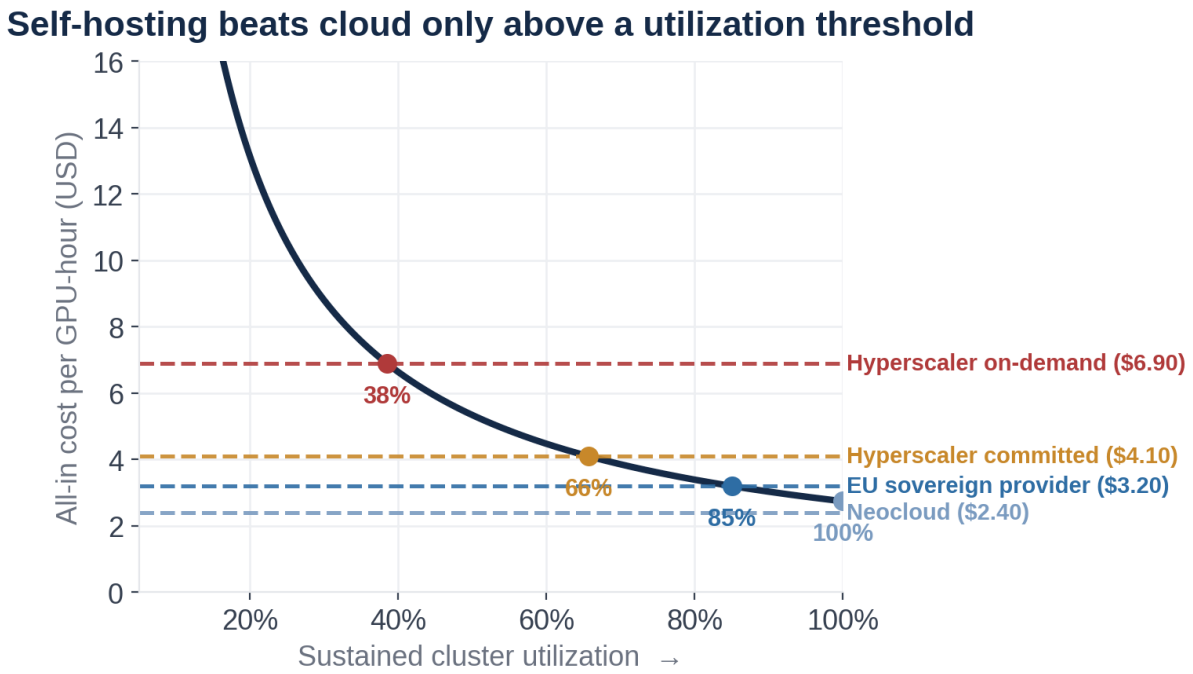

Figure 4 (original analysis). All-in self-hosting cost per GPU-hour versus sustained utilization (navy curve), against market price lines: hyperscaler on-demand ($6.90), hyperscaler committed ($4.10), EU sovereign dedicated ($3.20), neocloud ($2.40). Break-even utilizations marked: 38%, 66%, 85%, ~100%. Our model; assumptions in Table 4.

5.3 Sensitivity: what actually moves the answer

Figure 5 stress-tests the 66% break-even against committed cloud pricing, and the least intuitive finding comes first: Europe's energy premium is second-order for AI economics. Sweeping electricity from €0.08 to €0.22/kWh, roughly the Stockholm-to-Frankfurt range, moves the break-even by barely two points (65% to 67%), because power is about $0.15 of a $4 all-in GPU-hour. What dominates is the capital side. Shorten GPU life from four years to two — a live risk given Nvidia's annual cadence, which has already reshaped H100 pricing — and the break-even jumps from 57% to 83%. The cadence behind that risk now has product names: Blackwell shipped in 2025, Rubin arrives from late 2026, Rubin Ultra follows, which puts the frontier-competitive life of a fleet at two to three years. Hyperscalers absorb this aging through a value cascade, re-deploying yesterday's accelerators into cheaper workload tiers across millions of pooled customers. A dedicated enterprise inference fleet has far fewer places for old silicon to go, so the residual-value risk stays with the owner. A 30% fall in cloud prices, roughly one year of recent trend, pushes it to 95%, effectively out of reach. The build case is therefore a bet on utilization discipline and depreciation, not on cheap power, and it degrades on its own as cloud prices fall. Europe's energy disadvantage matters a great deal to operators siting gigawatt campuses: a 100 MW campus pays around €44m a year in Sweden against €149m in Dublin (IEA/market data, 2026). For an enterprise deciding whether to own 64 GPUs, it is a rounding error.

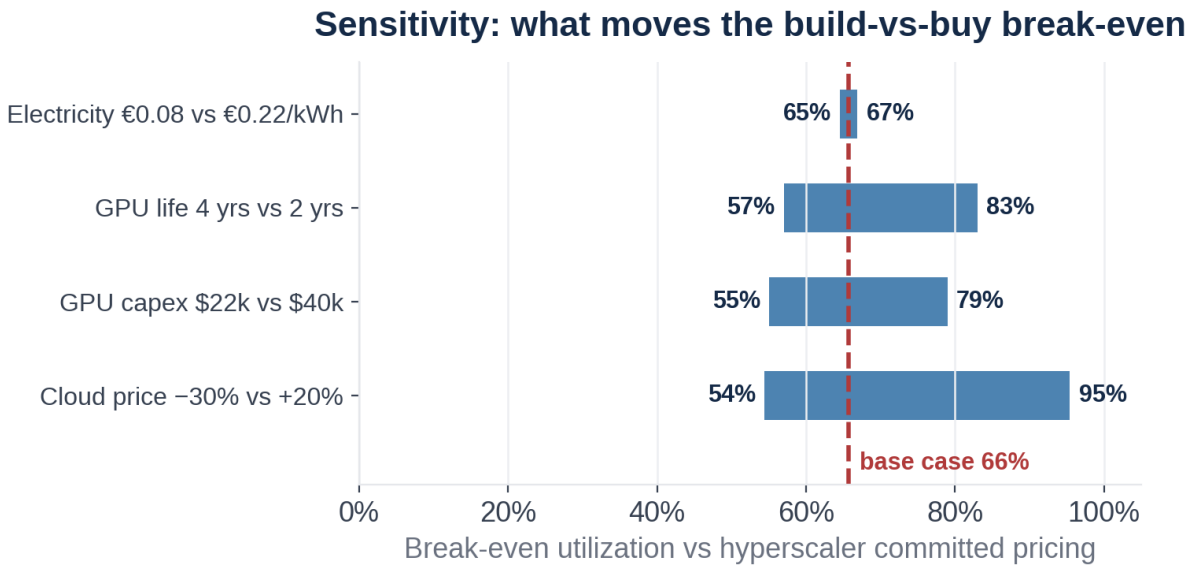

Figure 5 (original analysis). Sensitivity of the build-vs-buy break-even (base case 66%, red line, vs hyperscaler committed pricing): ranges for electricity price, GPU depreciation life, GPU capex, and cloud price trajectory. Depreciation and cloud-price deflation dominate; electricity barely registers. Our model.

5.4 Token deflation and the option value of waiting

The strongest argument against building is not in the cluster math at all. It is the price trajectory of the alternative, and here two curves matter that are routinely conflated. The *fixed-capability* curve is the spectacular one: Epoch AI finds inference prices at a fixed performance level falling 9x–900x per year depending on the task, and GPT-3.5-class capability fell from $20 to $0.07 per million tokens in under two years, and GPT-4-level capability has followed the same slope, from roughly $20 per million to around $0.40 in under three years (Epoch AI, 2025–2026; Stanford HAI, 2026; public price lists). The *realized blended* curve, what an enterprise actually pays as it keeps upgrading models and mixing workload tiers, declines far more slowly. For planning we adopt the assumptions of our companion briefing, the **StratoFactory AI Token Cost Forecast 2026–2030** (May 2026): blended enterprise inference falls from **$4.75 per million tokens in 2026 to about $1.57 by 2030 (−24% per year, central case)**, from tier anchors of $1.50 (low-tier, 60% of workload), $8.50 (mid), and $13.00 (frontier, 10%). The decline is tier-differentiated, roughly −25% per year at the commoditized end against −12.5% at the frontier, whose pricing premium persists. And it is not guaranteed: under the forecast's Shock scenario (GPU capacity squeeze, agentic mix-shift, frontier moat), blended prices rise about 15% in 2027 and remain 2–3x above the central case in 2030.

Run these assumptions through our cluster. At 70% utilization an 8-GPU node costs about $31 an hour; serving a 70B-class open-weight model at a conservative 2,500 tokens per second, that works out to $3.40 per million tokens — below today's mid-tier and frontier API anchors, but more than double the $1.50 low-tier price already available in 2026, and above the $1.57 blended central case for 2030. At an optimized 10,000 tokens per second the figure reaches $0.86, competitive with the 2030 blended price but still above where the low tier is heading. The

structural problem for the build case falls straight out of the tier math: self-hosted open-weight models compete in the commoditized tier, which is where prices fall fastest, while the tier whose premium endures — the frontier — cannot be self-hosted at all. There is also a plainer way to put what the curve means for ownership: if market token prices fall faster than you amortize capex, power, and operations, then cost-per-token leadership on owned hardware is temporary by construction. The question is not whether the rented alternative catches up, but whether it does so before payback. The forecast's own conclusion points the same way. Routing discipline across tiers and vendors captures three to five times more savings than negotiated discounts, an argument for architectural flexibility rather than for owning silicon. Committing capex today locks in 2026 economics against a planning curve falling 24% a year, with fixed-capability deflation running far faster beneath it, and the Data Act's egress-fee ban (January 2027) makes the stay-flexible strategy cheaper still. One caveat cuts the other way: the Shock scenario is the one world in which owned capacity outperforms. The sensible hedge for that world is contract structure and a pre-authorized contingency, not speculative capex.

5.5 Talent, lock-in, and the costs both sides omit

Two lines complete the ledger. **Talent:** our three-FTE assumption (about €486k a year loaded) is a floor, and the people are scarce. Only 20% of EU enterprises used any AI in 2025 (Eurostat), the MLOps population is thin outside a few capitals, and a sovereignty pitch that assumes the team into existence is selling the cluster without the payroll. **Lock-in, which runs in both directions:** hyperscaler dependence carries pricing-power risk, and the Broadcom/VMware episode is the textbook case, with reported increases of 8–15x after the acquisition and 74% of IT leaders exploring exits (CIO Dive, 2025). But a self-built stack is locked to its own depreciation schedule and to Nvidia's roadmap, and a small sovereign provider can itself be acquired, repriced, or fail. From January 2027 the Data Act removes egress fees, historically the strongest lock-in lever, which makes a rehearsed exit plan the cheapest sovereignty instrument on the market. The section's verdict: build only with sustained utilization credibly above two-thirds for three years, workloads stable enough to survive the deflation curve, or a legal mandate (Section 3) that removes the choice. Otherwise rent, and spend a fraction of the savings on encryption, exit rehearsal, and contract discipline. One exception scales down rather than up: for narrow, stable tasks, a fine-tuned small model on one or two GPUs avoids the cluster economics altogether — Section 6 takes this up.

6. The Capability Gap: How Much Model Do You Give Up?

Sovereign postures lean on open-weight models (Layer 4) or on European API providers. What does that cost in capability? The answer changed direction recently: after nearly closing in 2024, the frontier gap re-opened, modestly.

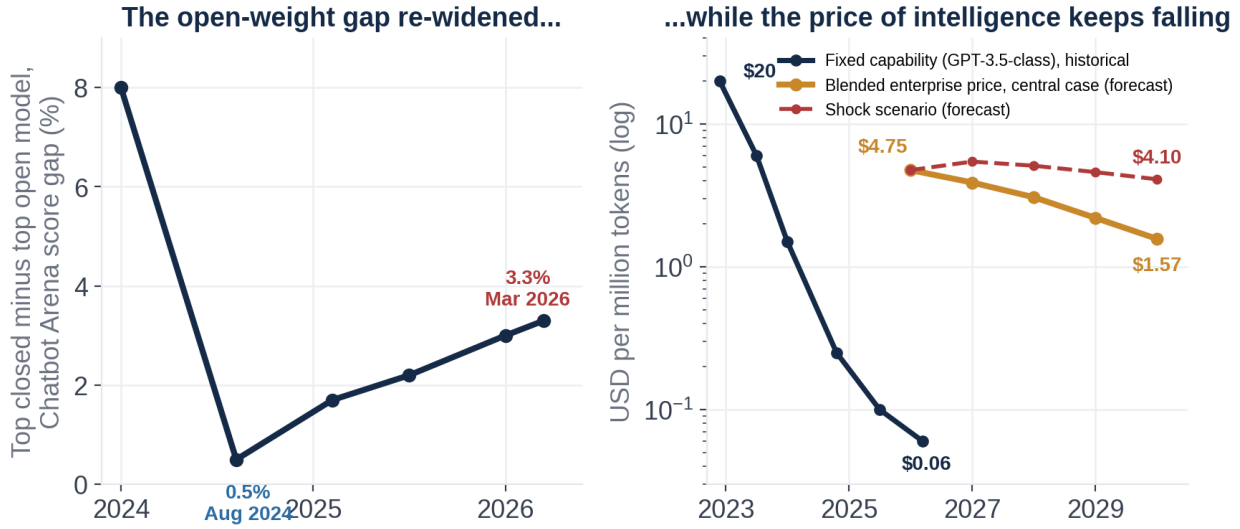

Figure 6. Left: the gap between the best closed and best open-weight model on Chatbot Arena narrowed to 0.5% in August 2024, then re-widened to 3.3% by March 2026 as frontier labs pulled ahead on reasoning; six of the top ten models are closed. Right: the price of a fixed capability (GPT-3.5-class) collapsed from $20 to ~$0.06 per million tokens (navy, historical), while the blended price enterprises actually pay is forecast to fall more slowly — $4.75 to ~$1.57 per million by 2030, central case (gold), with a Shock scenario staying 2–3x higher (red). Sources: Stanford HAI AI Index (2026); Epoch AI (2025–2026); StratoFactory AI Token Cost Forecast 2026–2030 (May 2026). Curves interpolated between reported points.

Where open-weight is already good enough. Document extraction and summarization, RAG over enterprise corpora, classification, translation, internal copilots, structured-output workflows: the bounded workloads that carry most actual enterprise volume. Here the 2026 generation of open models (Llama 4, Qwen 3.5, DeepSeek V4, Gemma 4, Mistral Medium 3.5) sits within a few points of the frontier on standard benchmarks, and the deployment questions that matter are latency, integration, and governance rather than raw capability. This is why open-weight defaults are spreading through EU regulated sectors, and why Mistral's enterprise traction (Section 4) is believable.

Small models change the sovereignty math at the low end. Not every sovereign workload needs a frontier-scale model, or the 64-GPU cluster of Section 5. Small language models — roughly the 1–15B class (Gemma, Ministral, Phi, the small Qwen variants) — have improved faster than any other tier: the smallest model clearing a fixed benchmark bar shrank 142-fold in two years (Stanford HAI AI Index, 2025), and a small model fine-tuned on a narrow task often matches a general model many times its size. The economics change completely. A fine-tuned SLM serves on one or two GPUs, or on existing edge hardware on a plant floor, which sidesteps

the utilization cliff entirely and makes it the cheapest working implementation of Layer 4 — frequently the only sovereignty investment whose payback survives the Section 5 math at low volumes. It is also where the token forecast's routing logic lands: the low tier expected to carry about 73% of workload by 2030 is, in self-hosted form, small-model territory. Two caveats apply. "Open weight" and "open source" are not synonyms — most downloadable models publish weights under licenses with commercial or usage restrictions, while fully open models with training data and code (OLMo, and the EU-funded OpenEuroLLM effort) remain rare, a distinction that matters for audits and under the AI Act, which applies lighter GPAI obligations to free and open-source models unless they carry systemic risk. And a fleet of small models still carries the platform burden described below: evaluation, guardrails, and lifecycle management stay with the owner.

Where the frontier still earns its price. Long-horizon agentic work, complex multi-step reasoning, cutting-edge coding, the newest modalities. The 3.3-point Arena gap understates the difference on hard tasks, where leaderboard compression hides real separation, and the frontier labs are also shipping the agentic scaffolding — tool use, computer use, long-context orchestration — fastest. An enterprise betting core product differentiation on AI capability will feel the self-hosting penalty in quarters, not years. A frozen stack's pace-of-innovation penalty is often larger than its price advantage.

The gap is wider than the model. Hyperscalers do not sell inference; they sell an environment for turning models into business solutions, and that environment is mature and developing continuously. Around every hosted model sits a managed scaffold: data lakes and vector stores, identity and security integration, guardrails and evaluation harnesses, LLMOps pipelines, agent orchestration frameworks, prebuilt industry solutions, marketplaces, compliance certifications, and a partner and skills ecosystem measured in hundreds of thousands of practitioners. The scaffolding improves weekly, without a change request. A sovereignty decision that prices only the GPU-hour or the token is implicitly assuming all of this is free to replicate. It is not; rebuilding even a serviceable fraction on a sovereign stack is a multi-year engineering program. It also sets the real benchmark for the build decision: hyperscalers pair frontier models with fine-tuning, evaluation, and agent tooling at global utilization, so an open model on owned nodes has to win on total solution cost, not per-node efficiency. This, more than any benchmark, explains the Bitkom paradox from Section 1 — 82% of German firms want out of US-cloud dependence, 65% will not accept functional limitations to get out. The model gap is 3.3 points and narrowing at the commodity end. The platform gap is larger, harder to measure, and currently widening, and a sovereign business case that omits it is not a business case.

Reliability and scale cut both ways. Hyperscaler-hosted frontier APIs deliver elastic scale and SLAs that a self-hosted cluster cannot match without over-provisioning, which wrecks Section 5's utilization math. On the other side, self-hosted inference offers latency determinism, data control, and freedom from provider-side model deprecations, all real operational virtues in regulated production. The practical rule is the same one running through this whole report:

match the capability tier to the workload, and let the small set of workloads that truly need Layers 2–4 absorb the penalty knowingly.

7. Risk, Both Ways: The Two Uncomfortable Columns

Most sovereignty commentary argues one column. Both are real, and the discipline is holding them in view at the same time.

The risks of hyperscaler dependence

- **Extraterritorial access is not hypothetical law.** The CLOUD Act/FISA regime, the standing GDPR Article 48 conflict, and a DPF that is one court ruling from its third collapse (Section 3). Weighted by probability, the scenario that hurts is not a surveillance headline. It is a transfer-mechanism failure that forces re-architecture on a court's timetable instead of yours.
- **Coercion now has a precedent.** Whatever the precise mechanics, and despite Microsoft's denials, the ICC episode showed that US sanctions can degrade a specific European institution's services. For most enterprises the direct risk is low. For anyone adjacent to geopolitics — defense, critical infrastructure, international institutions — it is a documented scenario rather than a thought experiment.
- **Concentration is officially systemic.** EU financial supervisors designated the hyperscalers as critical third-party providers under DORA in November 2025. Once your regulator has formally labeled your vendor a systemic risk, "everyone uses them" no longer works as a risk assessment.
- **Pricing power compounds.** Broadcom/VMware, with reported increases of 8–15x, demonstrated the pattern: dominant infrastructure vendors harvest locked-in bases. Three vendors holding 70% of the European market is the precondition. The Data Act's egress ban trims one lever; it does not touch the deeper switching costs of managed-service entanglement.

The risks of the sovereign path

- **Stranded capex is the base case, not the tail.** Section 5 shows why: a cluster bought in 2026 must beat a rental curve on a −24%-per-year planning trajectory, far steeper at fixed capability, while Nvidia's annual cadence compresses resale values. The first prominent European enterprise GPU write-down will be a news story within the forecast window (Section 9).
- **Capability lag is a compounding tax, and the platform lag compounds faster.** The 3.3-point model gap (Section 6) is small. The environment gap — the continuously developing hyperscaler scaffolding of data, security, MLOps, and agent tooling around the models — is larger, and on a sovereign stack it must be rebuilt or foregone. A sovereignty program that freezes its stack pays both taxes every year.

- **Counterparty fragility.** European providers run at a thirtieth of hyperscaler scale; national champions pivot (Aleph Alpha) or merge across the Atlantic (Cohere–Aleph Alpha); Gaia-X showed the limits of institutional stamina. A sovereign vendor that fails, gets acquired, or raises prices is the same lock-in risk under a European flag.
- **False sovereignty is the quietest risk of all.** Paying a 15–30% premium for an offering that delivers residency plus branding, while the jurisdictional core stays contested (Section 4), combines hyperscaler exposure with sovereign prices. Demand the layer number (Section 2) and the legal opinion in writing.

Our net read: for most enterprises, the probability-weighted cost of the hyperscaler risks remains lower than the certain, immediate costs of the sovereign path. Three exceptions flip the sign: a legal mandate (SecNumCloud-scoped public work, and CADA tiers to come), existential exposure (institutions adjacent to geopolitics), and genuinely high-utilization, stable AI workloads where the Section 5 numbers clear. Note that each of these is a statement about a workload, not about a company. That is Section 8’s framework.

8. Original Analysis: A Decision Framework by Archetype and Workload

This section is our own analysis rather than sourced fact. It builds on everything above: sovereignty is a property of workloads, not of companies. A bank does not need a sovereign marketing chatbot. A machine-tool maker may need sovereign handling of exactly one dataset, its process know-how. We therefore map workloads on two axes — regulatory and legal exposure (what the law and coming procurement rules attach to the data) and strategic damage if access is lost or the data exposed (the continuity axis the ICC episode made concrete). Four postures fall out of the quadrants.

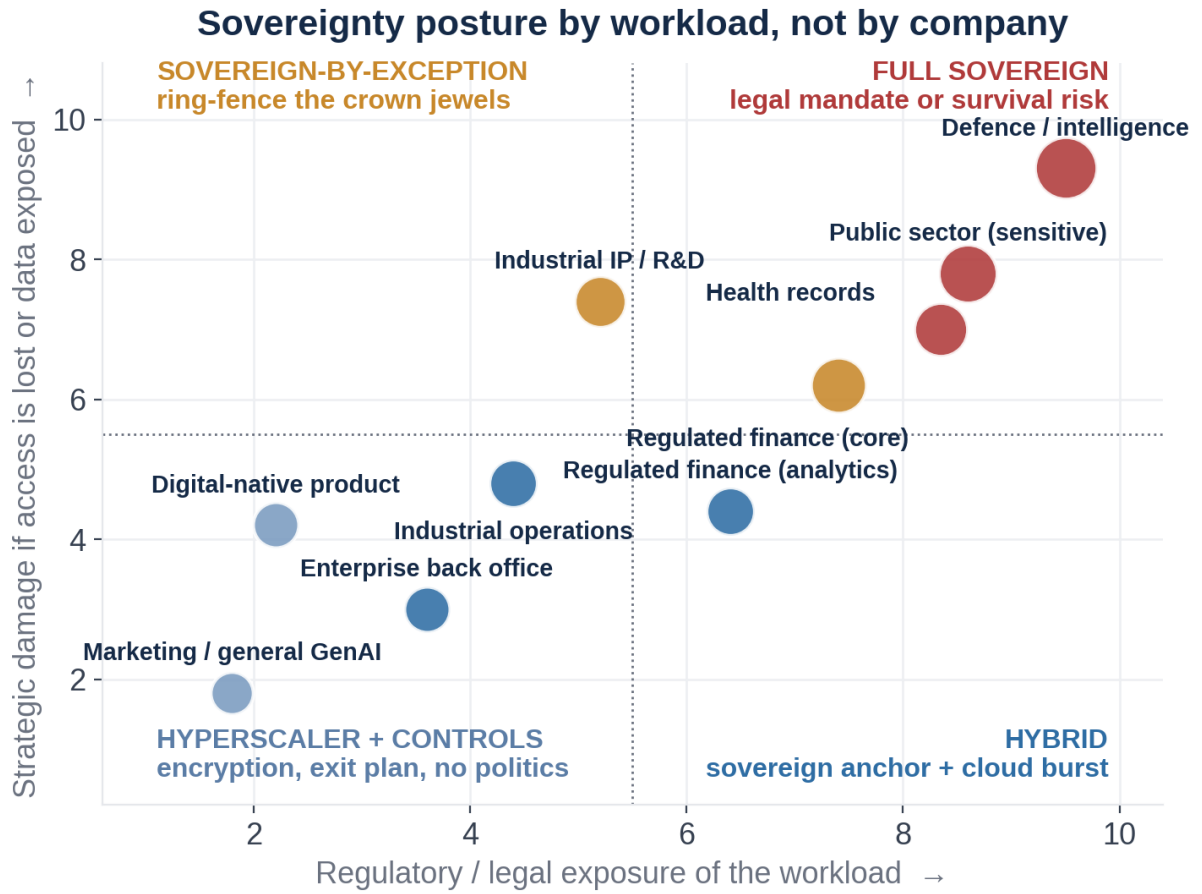

Figure 7 (original analysis). Sovereignty posture by workload, not by company: regulatory/legal exposure (x) versus strategic damage from loss of access or exposure (y). Bubble size indicates typical share of enterprise AI spend. Quadrants map to the four postures: hyperscaler-with-controls, hybrid, sovereign-by-exception, and full sovereign. Placements are our judgment, calibrated to Sections 3–7.

- **Full sovereign** (high exposure, high damage): EU-controlled infrastructure — SecNumCloud-class or on-premises — open-weight or EU-provider models, and no US-parented operator anywhere in the chain. The cost premium is certain and immediate; what it buys is Layers 2–3. This is mandatory territory for sensitive public sector and

defense-adjacent work, and, as CADA-style tiers harden, for the top slice of regulated finance and health.

- **Sovereign-by-exception** (lower legal exposure, high strategic damage): the enterprise stays on hyperscalers by default, but named crown-jewel workloads — R&D corpora, process know-how, M&A rooms, board material — run on EU-controlled or on-premises stacks with open-weight models. This is usually the cheapest defensible form of sovereignty, because the premium applies to 5–15% of spend — and fine-tuned small models on modest hardware are usually the right implementation (Section 6).
- **Hybrid** (high legal exposure, moderate damage): a sovereign anchor for regulated data and models, hyperscaler burst for scale and frontier capability, and classification discipline at the boundary. The natural landing zone for most regulated finance and health analytics, and the posture DORA exit-planning pushes firms toward in any case.
- **Hyperscaler-with-controls** (lower exposure, lower damage): the right answer for most workloads at most enterprises, and it should be stated without embarrassment — on price, and because the hyperscalers' continuously developing AI platform environment (Section 6) is itself a competitive asset the sovereign alternatives cannot yet match. The controls that matter: EU regions and data boundary, customer-held keys (HYOK where available), contractual transfer safeguards, Data-Act-era exit rehearsal, and a standing repatriation design for the minority of workloads that might ever need to move. Politics may dislike this answer. The TCO, capability, and legal-probability analysis above supports it.

Archetype	Default posture	Sovereign scope (the exceptions)	The honest note
Regulated finance	Hybrid	Core banking data and models, under DORA exit discipline; DPF-failure contingency for customer data	Analytics, CX, and productivity AI stay on hyperscalers; the premium is for the core only
Health & public sector	Full sovereign → hybrid	Patient and citizen records; anything SecNumCloud- or future CADA-scoped	Research and imaging compute often lands legitimately on hyperscalers — with controls, not slogans
Industrial & manufacturing	Sovereign-by-exception	Process know-how, R&D corpora, M&A material on EU-controlled or on-prem stacks with open-weight models	ERP, back-office, and general analytics AI: hyperscaler, despite the politics
Digital-native	Hyperscaler-with-controls	Almost none; exit rehearsal and key custody only	Frontier capability pace is the business model; sovereignty premium is pure drag

Table 5. Four enterprise archetypes, four default postures — with the exceptions named explicitly. Our analysis (original), calibrated to Sections 3–7.

Where the answer is "hyperscaler, despite the politics." Digital-native firms whose product cadence depends on frontier capability. Industrial enterprises' general analytics and back-office

AI. Any workload whose data is neither special-category nor strategically identifying. And, uncomfortably for the sovereignty narrative, most early-stage GenAI experimentation, since committing sovereign capex before workloads stabilize is Section 5's stranded-capex scenario in miniature. The same directness applies in reverse. Sensitive public-sector buyers should stop pretending that hyperscaler sovereign branding settles the CLOUD Act question; it demonstrably does not. And regulated firms should stop treating "we use the EU region" as an answer to transfer risk — that is Layer 1 wearing Layer 2's clothes.

The 90-day version for a CIO. First, classify AI workloads on Figure 7's two axes; this is a week of workshops, not a consulting year. Second, price the premium per workload against Table 1's layers, and demand the layer number from every vendor. Third, buy option value everywhere: key custody, exit rehearsal, a second-source open-weight fallback for the highest-volume workloads. Fourth, revisit annually — the deflation curve, the DPF docket, and CADA's passage each move the frontier between quadrants.

9. Predictions, 2026–2029

These are concrete, falsifiable calls. They are our judgment, not sourced fact, and each is paired with the evidence that would prove it wrong. Together they extend the through-line of this report: regulation tightens slowly, hyperscalers absorb the sovereignty wave, and economics beats politics wherever the two conflict.

#	Prediction (2026–2029)	What would prove it wrong
1	European providers still hold under 20% of the European cloud market at end-2029; the sovereignty wave is captured mostly by hyperscalers' own sovereign lines	Synergy quarterly share ≥20%; or independent EU providers taking >50% of sovereign-segment revenue
2	The EU–US Data Privacy Framework does not survive to end-2029 in its current form (CJEU strike-down, material conditioning, or US executive-order change) — and its failure triggers the first board-mandated repatriation programs at scale	DPF intact and unmodified through 2029, with no major transfer disruption
3	Open-weight models become the default — more than half of new GenAI deployments — in EU regulated sectors by end-2028	Deployment surveys showing closed APIs retaining majority share in EU finance, health, and public sector
4	The median rented H100-class GPU-hour falls below $1.50 by end-2027, and at least one European enterprise AI-cluster write-down or fire-sale is publicly disclosed by end-2028	Rental prices holding above $2; no disclosed impairment or distressed resale
5	Fewer than half of the funded EU AI gigafactories operate at ≥50,000 accelerators by end-2029	Three or more gigafactories at target scale, on schedule
6	CADA passes by 2028 in diluted form: Levels 3–4 reserved to a single-digit share of public workloads, and US hyperscalers retain access to ≥70% of EU public cloud spend	A final text excluding US-parented providers from the majority of EU public procurement

Table 6. Six testable calls and their falsifiers — our judgment, not sourced fact. Calls 1, 3, and 4 are trackable against Synergy, sector surveys, and public price trackers respectively.

The through-line, stated once more: over this horizon, the binding constraint on European AI sovereignty is not regulation, ambition, or even capital. It is the compound interest of capability pace and cost deflation, running through infrastructure someone else owns. Sovereignty strategies that respect that arithmetic — workload triage, open-weight defaults, spending on option value — will do quietly well. Strategies that fight it, through company-wide repatriation, residency-as-immunity at a premium, or capex ahead of workload stability, will produce the write-downs that prediction 4 anticipates.

10. Appendix: Methodology, Conflicts, Glossary, and Sources

10.1 Methodology

This briefing synthesizes public regulatory texts and official timelines (European Commission, EuroHPC JU, ANSSI, BSI, ESAs), market research (Synergy Research, IDC, BCG), academic and independent trackers (Stanford HAI AI Index 2026, Epoch AI), industry-association surveys (Bitkom, Cigref), vendor disclosures (flagged as such), and reporting by Bloomberg, CNBC, TechCrunch, Computer Weekly, and specialist legal analyses (Gibson Dunn, Latham & Watkins, Travers Smith), current to early July 2026. Token-price planning assumptions in Section 5.4 are adopted from the companion StratoFactory briefing *AI Token Cost Forecast 2026–2030* (May 2026). The TCO model in Section 5 is original; every assumption appears in Table 4, and the sensitivity ranges in Figure 5 bound the conclusions we rely on. Where sources conflict, we show the conflict rather than average it away (10.2). Figures 4, 5, and 7 are original analysis and must not be quoted as measured market data.

10.2 Cross-referencing and conflicts (a feature, not a bug)

- **Sovereign-cloud market sizes diverge by more than 5x.** Global 2026 estimates span roughly $139bn (Precedence) to $195bn (Fortune Business Insights) at CAGRs of 19–27%, and a European figure of €12.4bn for 2026 coexists with a ~€20bn “today” estimate (Roland Berger-aligned). The spread reflects definitional chaos: “sovereign” variously includes any EU-region revenue, only EU-controlled providers, or anything certified. We therefore cite adoption surveys (IDC) and structural shares (Synergy) rather than any single market-size number.
- **Whether hyperscaler sovereign structures survive CLOUD Act scrutiny is contested and untested.** Vendor legal positions (AWS, Microsoft) assert effective protection. Independent analyses (KuppingerCole, InfoQ, several EU law firms) point to parent-level attachment, and Microsoft’s own French Senate testimony leans against the vendor position. No litigation has resolved it. We treat the jurisdictional immunity of US-parented sovereign offerings as an open legal hypothesis.
- **Sovereignty premiums range from about 15% (Delos vs Azure, reported) to about 30% (BCG, hyperscaler sovereign lines), while neoclouds price 60–70% below hyperscaler GPU rates.** The apparent paradox dissolves once “sovereign premium” is read correctly: it is a premium over the hyperscaler’s own pricing, which already sits at the top of the market.
- **Adoption statistics measure different things.** Eurostat’s 20% (EU enterprises using any AI, 2025) coexists with Stanford HAI’s 88% (global organizations reporting AI use) and IDC’s 40% (European organizations using sovereign cloud solutions). The sampling frames and definitions differ. None is wrong, and none should be quoted against another without labels.

- **Public-program figures are mobilization targets, not spend.** InvestAI's €200bn includes €150bn of hoped-for private capital; IPCEI-CIS support is reported at both €2.6bn and €3.5bn depending on scope. We flag both rather than choosing.
- **Token-price forecasts diverge by an order of magnitude because they measure different curves.** Fixed-capability declines are extreme (Epoch AI: cost at fixed performance halving roughly every two months; a16z: ~10x per year). Realized-blended declines, which are what a budget experiences, are far slower: the advisory-consensus view is about −37% per year, and the StratoFactory central case is −24% per year with a Shock scenario in which prices rise. Quoting a fixed-capability rate in a budget discussion, or the reverse, is the most common analytical error in this space. We use fixed-capability rates for the capability-price narrative (Fig. 6) and the StratoFactory blended central case for all planning arithmetic (Section 5.4).

10.3 Caveats and limitations

Cloud and GPU pricing moves monthly and varies heavily with negotiation; our price points are indicative mid-2026 market observations, not quotes. The TCO model is deliberately simple. It carries no cost of capital, software licensing, or downtime and failure costs, all of which would worsen the build case, and no resale value or tax treatment, which would improve it modestly. Legal characterizations reflect the state of pending matters in early July 2026 and are analysis, not legal advice; the Digital Omnibus had been approved by both co-legislators in June 2026 but not yet published in the Official Journal at press time, and CADA is a proposal whose final text may differ materially. Token-price projections adopted from the StratoFactory AI Token Cost Forecast are themselves forecasts with a wide scenario band, including a Shock scenario in which prices rise, and inherit that briefing's caveats. Survey data (Bitkom, IDC) reports intent, and Section 4's market-share data shows that intent does not reliably convert into procurement.

10.4 Glossary

CLOUD Act — US law (2018) compelling US-jurisdiction providers to produce data regardless of storage location. **FISA §702** — US foreign-intelligence collection authority operating through providers. **DPF** — EU–US Data Privacy Framework, the adequacy mechanism for transatlantic transfers. **Schrems II** — 2020 CJEU ruling invalidating Privacy Shield. **GPAI** — general-purpose AI models under the AI Act. **SecNumCloud** — ANSSI qualification requiring EU control (France). **C5** — German BSI cloud security attestation (no ownership test). **EUCS** — draft EU cloud certification scheme, stalled over sovereignty tiers. **CADA** — proposed EU Cloud and AI Development Act (June 2026). **DORA / NIS2** — EU financial-sector ICT resilience regulation / cybersecurity directive. **Data Act** — EU regulation banning cloud egress/switching fees from January 2027. **Egress** — charges for moving data out of a cloud. **Neocloud** — specialist GPU cloud provider. **Open-weight model** — model whose weights are downloadable and self-hostable. **PUE** — power usage effectiveness, data-center energy overhead ratio. **Utilization** — share of available GPU-hours actually used. **Blended token price** — average price per million tokens across workload tiers, as actually paid; distinct from fixed-capability price.

10.5 Source list

- European Commission — AI Act texts, implementation timeline, Digital Omnibus materials; Data Act; Cloud and AI Development Act proposal (June 2026); InvestAI announcements. 2024–2026.
- Gibson Dunn; Latham & Watkins; Travers Smith; DLA Piper; Kennedys — analyses of the AI Act Digital Omnibus agreement, revised high-risk deadlines, and June 2026 adoption steps. 2026.
- Court of Justice of the EU / General Court — *Schrems II* (C-311/18); *Latombe v Commission* (September 2025); noyb announcements on DPF challenges. 2020–2026.
- French Senate hearing testimony of Microsoft France legal representatives (June 2025); reporting by heise, Computer Weekly on the ICC/Microsoft episode and Dutch government reviews. 2025.
- ANSSI — SecNumCloud 3.2 referential and qualified-provider list; BSI — C5 catalogue; cep — “EU Cloud Certification at an Impasse” (April 2025); EU ISS on EUCS politics. 2024–2026.
- European Supervisory Authorities — designation of critical ICT third-party providers under DORA (November 2025).
- Synergy Research Group — European cloud market size and provider shares (European providers ~15%; top-3 US ~70%; €61bn market 2024). 2024–2025.
- IDC — Worldwide Digital Sovereignty Survey (June 2026); public cloud spending forecasts. 2025–2026.
- BCG — sovereign cloud premium analysis (2025). E-3 Magazine — Delos pricing report (2025). Computer Weekly — neocloud vs hyperscaler GPU pricing (2025).
- Bitkom — Cloud Report 2025 and 2026 position papers on cloud sovereignty criteria. Cigref/Asterès — technological-dependence economic assessment (2025).
- Stanford HAI — AI Index Report 2026 (open-vs-closed gap; adoption; model economics). Epoch AI — LLM inference price trends (2025–2026).
- StratoFactory — AI Token Cost Forecast 2026–2030: A multi-cloud, multi-vendor pricing outlook (May 2026) — source of the blended token-price planning assumptions, tier anchors, and Shock scenario used in Section 5.4 [companion briefing, same publisher].
- EuroHPC JU — JUPITER launch materials (September–November 2025); AI Factories selections (2024–2025); gigafactory call status (2026). EIB — InvestAI co-financing (2025).
- AWS — European Sovereign Cloud launch materials (January 2026) [vendor]. Microsoft — Sovereign Cloud announcements (June/November 2025) [vendor]. Google Cloud — sovereign partnerships and air-gapped offerings [vendor]. OVHcloud — FY2025 results [company disclosure].
- Bloomberg; TechCrunch; CNBC — Mistral funding and valuation reporting (June 2026); Cohere–Aleph Alpha merger (April 2026).

- IEA — Electricity 2026 (EU industrial electricity ~2x US); Eurostat — electricity price statistics and AI-use-in-enterprises series (2025); market reporting on European data-center energy costs. 2025–2026.
- Public GPU price trackers and provider list prices (getdeploying, gpuperhour, provider pages) for H100/H200/B200 rental rates [market data, indicative]. CIO Dive; Register — Broadcom/VMware pricing reporting. 2025–2026.

11. Disclaimer and Notices

This briefing is published for public distribution (including via LinkedIn and similar platforms). The following notices apply to all readers.

11.1 Author and nature of this document

This briefing reflects the personal professional analysis of Piotr Jurowiec and is published by StratoFactory.com. Views expressed are those of the author and do not necessarily represent the views, positions, or official forecasts of any employer, client, partner, or affiliated organization. This document is independent market research prepared for general informational and educational purposes. It is not investment, legal, tax, accounting, procurement, or engineering-management advice; it is not a recommendation to purchase from, contract with, or invest in any specific vendor, model, platform, or technology; and it does not constitute an offer or solicitation of any kind. Regulatory and legal matters discussed herein — including the EU AI Act, GDPR transfer mechanisms, the US CLOUD Act and FISA, and pending proposals and litigation — are presented as analysis, not legal advice; readers should consult qualified counsel before acting. Readers should not rely on this document as the sole basis for any financial, procurement, hiring, or strategic decision.

11.2 Forward-looking statements and uncertainty

All forecasts, predictions, scenarios, models, illustrative figures, and recommendations contained herein are estimates based on publicly available information and analyst judgment as of the publication date. The economics, capabilities, and regulation of AI have historically moved several multiples faster than any 12-month-old plan, and legal, architectural, or geopolitical shocks — including court rulings on transfer mechanisms and the final texts of pending EU legislation — could shift the trajectory materially. Actual outcomes may differ, in either direction, from any prediction presented. The original models and estimates in Sections 5 and 8 are illustrative and must not be quoted as measured data.

11.3 Third-party sources, trademarks, and non-endorsement

This briefing cites publicly available research, press releases, vendor disclosures, regulatory and corporate statements, and analyst commentary from a range of organizations including (but not limited to) the European Commission, EuroHPC JU, ANSSI, BSI, ENISA, the European Supervisory Authorities, Synergy Research Group, IDC, Boston Consulting Group, Stanford

HAI, Epoch AI, Bitkom, Cigref, the IEA, Eurostat, Amazon Web Services, Microsoft, Google, OVHcloud, Deutsche Telekom/T-Systems, Schwarz Group/StackIT, IONOS, Scaleway, Mistral AI, Cohere, Aleph Alpha, Nvidia, Broadcom/VMware, Bloomberg, CNBC, and TechCrunch. All company names, product names, logos, and trademarks are the property of their respective owners and are used here solely for identification, reference, comparison, commentary, and analysis (nominative fair use) under the fair-use doctrine (US) and fair-dealing principles (UK / EU / other applicable jurisdictions). No association, sponsorship, affiliation, partnership, or endorsement by any trademark owner is intended or implied, and no trademark owner has reviewed, approved, or contributed to this briefing.

The statistics, survey results, and other figures attributed to third parties are individual facts reproduced under quotation and fair-use principles for purposes of commentary and analysis, with sources identified throughout. Any errors in transcription or interpretation are the author's own.

No third-party chart, graphic, photograph, table, or passage of text is reproduced in this briefing. Figures 1–7 are original visualizations created for this document; where they plot third-party figures, they use individual data points as facts, with sources credited in each caption. Direct quotations are limited to short identifying phrases: statutory language (such as the CLOUD Act's "possession, custody, or control") is taken from official legal texts, which are not subject to copyright as US government works and, for EU instruments, are reusable with attribution under the European Commission's reuse policy (Decision 2011/833/EU); publication titles and product names appear for identification only. The document's design, template, charts, and logo are the property of StratoFactory.com, and the companion AI Token Cost Forecast briefing relied on in Section 5.4 is by the same publisher and used with its consent.

11.4 Data currency and verification

Statistics cited reflect figures published as of early July 2026 and drawn from the sources listed in Section 10.5. Cloud pricing, GPU economics, regulatory timelines, and litigation status move quickly; readers should verify current figures — in particular the Official Journal publication of the AI Act Digital Omnibus, the legislative progress of the proposed Cloud and AI Development Act, and the docket of challenges to the EU–US Data Privacy Framework — before relying on them.

11.5 No warranty; no liability

This document is provided strictly "as is" without warranty of any kind, express or implied, including but not limited to warranties of accuracy, completeness, merchantability, or fitness for a particular purpose. To the maximum extent permitted by applicable law, the author and StratoFactory.com disclaim all liability for any loss or damage, direct or indirect, arising from reliance on this document or its contents.

11.6 Currency of information

The information, predictions, and recommendations in this briefing reflect publicly available data and the author's judgment as of July 2026 and may be revised without notice as new information becomes available. No obligation is assumed to update this document.

11.7 Sharing, attribution, and corrections

Readers are welcome to share this briefing, quote brief portions with attribution, and reference the analysis with a link to the original. Corrections of factual errors are welcome and will be acknowledged in any revised edition; please direct them to the author via StratoFactory.com.

11.8 Conflicts of interest and independence

The author and StratoFactory.com received no compensation from any vendor, cloud provider, model developer, advisory firm, or other organization mentioned in this briefing, hold no undisclosed positions in any company analyzed, and prepared this document independently. No vendor reviewed or approved any portion of the content prior to publication.