AIMONGER
WHITE PAPER

AI Discovery

The Cost of Privacy: Sovereign AI as an Operating and Capital Decision

The privacy question in AI is no longer a policy slogan. It is a unit-economics and control decision: who owns the compute, the weights, the data stack, and the operating advantage that sits on top. The CFO pays whichever default you choose. Boards and shareholders care because the wrong path either burns cash on unmanaged tokens or quietly transfers institutional knowledge to someone else's platform.

David Saliba . 2026-07-23 . aimonger.com/whitepapers/cost-of-privacy-sovereign-ai/

The problem in one sentence

If you cannot say what privacy costs for each material AI workflow, you cannot tell shareholders whether you are buying control or renting a leaky utility.

Technology leaders are being asked two incompatible things at once. Ship AI into production. Keep regulated data, industrial know-how, and competitive process logic indoors. The market default is metered tokens on someone else's stack. The sovereignty default is dedicated capacity you operate. Both have a price. Only one of those prices is usually modelled before the first pilot goes live. The person who pays either way is finance.

Eurostat reports that 20.0 per cent of EU enterprises with 10 or more employees used AI in 2025, up from 13.5 per cent in 2024. Adoption is rising. Ownership of the stack is not automatic. The FinOps Foundation's State of FinOps 2026 finds that 98 per cent of surveyed practitioners now manage AI spend. Finance can see the line item. Architecture still has to answer who captures the value that line item produces.

This paper is written for the CFO who must fund and constrain the path, the CIO and CTO who must make it runnable, and the board that will approve whichever stack is chosen. It sits next to, but is not the same as, a pure placement matrix or a pure model-routing brief. The subject here is the privacy premium itself: what you pay to keep alpha indoors, and what you lose if you refuse to pay it.

> **Cost of privacy, plainly:** the extra money and operating load required to process sensitive AI work on infrastructure you control, instead of through multi-tenant token APIs that meter your questions about your own business. **In practice:** a cross-border manufacturer keeps supplier contracts and process recipes on a dedicated GPU tray in-region, and accepts a monthly always-on bill so that prompts never leave the private boundary.

> **Sovereign AI (operational), in short:** control over where inference runs, who can see prompts and outputs, who holds model weights, and whether you can exit without losing the workflow. **Worked case:** open weights on dedicated EU-region GPUs with your keys and your evaluation suite, not a slogan on a slide deck.

What boards and shareholders are actually buying

Shareholders do not buy "AI." They buy either a controllable operating asset or a recurring extraction bill with weak exit rights.

The seat-based SaaS era trained boards to expect a predictable licence line. Token-metered AI broke that habit. Industry cost commentary now labels the unmanaged pattern **tokenmaxxing**: maximising token throughput because usage feels free, leaderboards celebrate spend, and agent loops multiply calls until the invoice arrives as a discovery rather than a budget. The FinOps shift to nearly universal AI spend management is the market admitting that the old predictability is gone.

McKinsey's State of AI 2025 still shows the absorption gap: widespread regular use, far thinner enterprise EBIT impact, and a small high-performer cohort attributing more than 5 per cent of EBIT to AI. Unpriced privacy is one reason programmes stall. Security blocks the cheap path. Engineering cannot staff the private path. The workflow waits, or staff route around policy into shadow tools that destroy the control story the board thought it had bought.

Stakeholder	What they need priced	Failure if ignored
Board	Privacy premium vs leakage risk by data class	Approves a slogan, not a control
Shareholders	Capex/OpEx shape and exit optionality over 24 months	Surprise dilution of margins or of competitive position
CFO	Cash shape, ceilings, cost per successful case, idle GPU and labour in the tray	Discovers the bill after engineering has already chosen the meter or the hardware
CIO	Systems boundary, subprocessors, audit evidence	Policy without a runnable path
CTO	Weights, serve topology, eval parity, upgrade burden	Sovereignty cosplay on under-patched hardware

Plain read: if only engineering can explain the stack, finance has not yet made a capital decision.

The commercial argument: extraction versus subscription

Love him or hate him, Alex Karp has a load-bearing commercial point.

In July 2026 CNBC coverage of a Squawk Box interview, the Palantir CEO attacked the token commercial model used by frontier labs as something that had "gone completely wrong," arguing that enterprises risk wasting spend on tokens while transferring IP. He framed what serious customers want as control over compute, models, their data stack, and their **alpha**. Palantir's own social manifesto around the same moment criticised **tokenmaxxing** as a business model and pushed AI sovereignty language. SiliconANGLE's write-up of the same interview captured the sharper colour: token charges framed as a kind of wealth tax on businesses that get little durable value back.

You do not need to like the messenger. The mechanism is what boards should test:

1. **Extraction model.** Frontier labs can monetise every attempt a company makes to get answers from its own documents and systems, instead of selling a flat software subscription with known annual cost. Metering turns institutional curiosity into vendor revenue. The more deeply the firm wires agents into core workflows, the more the meter compounds.

2. **Loss of control.** If prompts, embeddings, logs, and sometimes fine-tunes live on a platform you do not operate, you have rented cognition. Ownership of computing resources, data stacks, and model weights is what makes the capability an asset rather than a dependency. Karp's line about owning the means of production is theatrical; the underlying ask is ordinary for any regulated or IP-heavy firm.

> **Extraction model (commercial) means:** a pricing and architecture pattern where the vendor earns more as you ask more questions of your own data, while retaining platform advantage over how those questions are processed. **On Monday:** an agent that re-reads the same contract corpus every night grows the vendor invoice and may leave retention or training-use residue the board never priced.

> **Tokenmaxxing, put simply:** treating token volume as a success metric, through culture, unconstrained agents, or flagship-by-default routing, until OpEx discovers the bill. **Operating case:** every draft, every triage, every overnight crew hits the frontier endpoint with no task tier and no monthly ceiling.

Argue back where due. Karp sells an application-layer and sovereignty narrative that benefits his company. Open-weight plus private compute is not free, and "sovereign" without patching, evaluation, and operators is worse than a hardened private managed endpoint. Polarising public performance does not cancel the invoice maths or the IP-transfer question.

Lessons from private capacity economics (generalised)

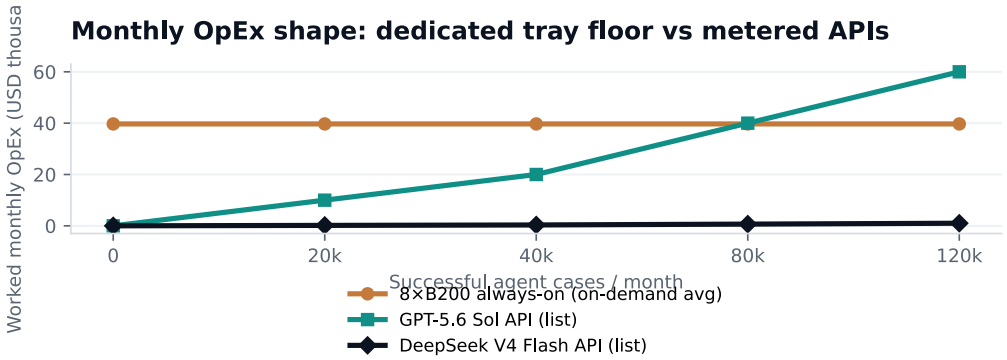
Dedicated inference teaches the same pattern every time it is priced honestly. No client name is needed for the operating lessons.

1. **Always-on trays dominate the privacy premium.** The expensive part of private agentic AI is keeping a capable model loaded on dedicated GPUs around the clock, plus storage, network, VAT where applicable, and the people who patch and evaluate it. Comparing that monthly tray to a light month of API tokens will always make privacy look "too expensive," because the comparison mixes an insurance premium with a variable utility bill.
2. **Concurrency is not a hardware multiplier.** A small number of concurrent agentic workers is usually a sequence and session budget on one serve node, not a reason to buy one full cluster per agent. Boards that hear "five agents" and approve five times the hardware are buying theatre. Boards that hear "five agents" and fund one correctly sized tray with headroom are buying a real posture.
3. **Hosted API unit prices are the wrong comparator for lockdown work.** When the requirement is that proprietary knowledge must not leave a controlled boundary, multi-tenant SaaS APIs are a different product class. The honest alternatives are private managed endpoints with contractual locks, or dedicated/sovereign trays. Foiling private GPU quotes against public token menus is how programmes get stuck in false comfort.
4. **Serve recipes beat wishful POC sizing.** Published recipes for large open-weight or frontier-class models often assume fuller GPU topologies than a thin "two-plus-two" POC budget. Under-sizing the tray to win a procurement slide creates a privacy label without privacy capacity. The board then discovers either quality failure or a mid-programme hardware uplift.
5. **Dedicated control and vendor SaaS are different risk products.** Knowledge lockdown defaults to dedicated or tightly bounded private capacity under your operational control. Vendor APIs can still be right for low-sensitivity work. Mixing the two without a data-class map is how shadow routing starts.

Market GPU rental bands put a cash number on that floor. getdeploying's B200 aggregator (retrieved 24 July 2026) shows on-demand listings averaging **$6.89 per GPU-hour**, spot averaging **$3.90**, and a full observed range of about **$3.35-$16.11**. An always-on eight-GPU tray at the on-demand average is **$39.7k per month** before storage, egress, VAT, and operators (8 × $6.89 × 720 hours). That is the privacy premium in cash terms: visible, monthly, and governable.

Figure 1 plots that floor against metered API OpEx under a transparent token assumption (40k input + 10k output per successful agent case) using July 2026 list prices: GPT-5.6 Sol at $5/$30 per 1M tokens

and DeepSeek V4 Flash at $0.14/$0.28. Under that mix, Sol costs about **$0.50 per case**, so it crosses the $39.7k tray near **80k successful cases per month**. Flash stays near the axis at these volumes. The chart is worked arithmetic from cited unit rates, not a survey and not a quote for any named deployment.



Worked arithmetic from cited unit rates (not a survey or forecast). Tray floor: 8×B200 × getdeploying on-demand avg $6.89/GPU-hr × 720h (https://getdeploying.com/gpus/nvidia-b200, retrieved 24 Jul 2026). Token lines: OpenAI GPT-5.6 Sol (5/30 per 1M in/out) and DeepSeek V4 Flash (0.14/0.28) with assumed 40k input + 10k output tokens per case. List prices and GPU listings churn; confirm vendor pages before budgeting. AIMonger.

Figure 1. Monthly OpEx shape from cited unit rates (worked example, not a forecast). Plain read: the dedicated tray is a flat ~$39.7k floor; flagship API spend rises with every agent case and crosses that floor near 80k cases under the stated token mix; cheap open-weight API spend stays trivial until volume is extreme. In practice: replace the assumption with your measured tokens-per-case and your real tray quote. Sources: getdeploying B200 on-demand avg $6.89/GPU-hr (retrieved 24 Jul 2026); OpenAI GPT-5.6 Sol and DeepSeek V4 Flash list prices (Jul 2026). List prices and GPU listings churn.

Tokenmaxxing as a business model, not a meme

The Reddit and industry thread under “tokenmaxxing” is messy, but the reusable ideas are sharp enough for a board pack.

- **Volume as vanity.** Internal cultures that celebrate who spent the most tokens are celebrating vendor revenue, not shareholder return. Without outcome tags, a leaderboard is an extraction accelerator wearing a productivity costume.
- **Invoice as discovery.** Seat SaaS let finance set a number. Usage metering lets finance discover a number after the fact. Practitioner guides such as BRM’s token-spend note for finance leaders document that vocabulary: the commercial model changed from predictable seats to consumption that must be allocated by workflow. That is market language, not independent proof that metering extracts proprietary advantage.
- **Agent loops multiply the meter.** Agentic workflows re-read, tool-call, and retry. A human chat session is a poor forecast for overnight crews. Industry commentary such as Verax’s tokenmaxxxing essay records the same practitioner worry: metered pricing plus agents ends the era when firms could pretend AI cost would rise “somewhat predictably” with headcount. Use those pieces for vocabulary and concern, not as econometric evidence.
- **SaaS vendors pass the meter downstream.** As incumbent suites add AI features, metered add-ons appear inside products boards already pay for by seat. Token risk is no longer only the frontier API line; it is embedded in collaboration, CRM, and ITSM estates.
- **Caps without routing are blunt.** Hard per-employee ceilings stop bleeding. They do not create value. Outcome-aware routing, cheaper tiers for extraction, frontier tiers for judgement, and evaluation gates are what turn a cap into an operating system.

Alpha (plain English): the proprietary judgements, workflows, data relationships, and operating logic that make your firm hard to copy. **On the P&L:** how a claims team escalates exceptions, how a plant reconciles mass balance, how counsel phrases a recurring clause; the stuff a frontier model should help execute, not quietly absorb.

Open-weight leverage: private trays, alpha, and Chinese-lab models

By mid-2026 the open-weight market is no longer a hobbyist footnote. DeepSeek-V4 (Pro and Flash, April 2026 preview with open weights and 1M-token context), Zhipu's GLM-5.2 (MIT open weights from June 2026, long-horizon and agentic coding positioning), and Alibaba's Qwen3.x family (hosted Max preview in July 2026 with open weights promised on a different cadence) give mid-market firms a real alternative to sending every alpha-bearing prompt to a US closed flagship.

CNBC's June–July 2026 journalism on Zhipu and DeepSeek framed the commercial pressure as of those publication dates: Chinese open models gained ground as US proprietary token prices stayed high and as US government limits temporarily restricted or staggered access to some Anthropic and OpenAI rollouts. The same journalism reported GLM-5.2 near Claude Opus 4.8 on a watched agentic benchmark at roughly one-fifth the token cost, and named firms such as Coinbase as downloading GLM-5.2 for self-hosted use. Treat those lines as journalistic reporting, not primary proof of quality, price, or adoption. For model facts that must hold in a board pack, prefer the vendor pages: DeepSeek's 24 April 2026 V4 preview and Z.ai's 16 June 2026 GLM-5.2 post (see References). The board question still stands: if intelligence-per-euro and revoke-resistance matter, why is every sensitive workflow still hard-coded to one closed API?

The alpha point is mechanical. If you download MIT- or Apache-licensed weights and serve them on a tray you operate, the prompt stays inside your VPC. That is a different product from calling a Chinese-hosted API where live prompts and completions may leave the EU. Where journalism reports enterprise downloads for self-hosting, use that as colour; require your own inventory of who may run which weights. The board should treat the hosted-versus-self-hosted distinction as load-bearing:

Path	What you buy	Alpha / data path	Geopolitical residual
US closed flagship API (GPT-5.x, Claude Opus-class, Gemini 3.x)	Peak quality on many tasks; vendor cadence	Prompts on vendor boundary; contractual training-use and retention	US access / export policy can change availability; meter extracts OpEx
Chinese-lab hosted API	Cheap tokens; fast experiments	Live data may leave your jurisdiction	Supplier jurisdiction + telemetry risk; procurement bans may apply
Self-hosted open weights (DeepSeek-V4, GLM-5.2, Qwen open family)	Weight possession; fixed tray OpEx; fine-tune optionality	Prompts stay on your infra if you operate the serve path	Origin of the lab still matters for policy, Entity List adjacency, and board optics; you own patching and eval

Open-weight leverage, briefly: using downloadable competitive models on private compute so alpha-bearing prompts are not metered and retained by a frontier platform you do not control. **In practice:** litigation notes and pricing logic hit a DeepSeek-V4 or GLM-5.2 tray you patch and evaluate; marketing drafts may still use a managed EU-zone Claude or GPT endpoint.

Chinese influence is a real board topic, not a culture-war slogan. Three control questions travel:

1. **Origin inventory.** NIST-style inventory should record model origin, licence, download hash, serve location, and whether any telemetry phones home. "We use open source" is not an inventory line. Zhipu and peer labs sit in a US–China technology contest; some buyers (especially US federal

and defence-adjacent) will refuse Chinese-origin weights regardless of MIT licence. EU mid-market firms still need a written origin policy for regulated and alpha-heavy workflows.

2. **Hosted versus self-hosted.** Calling a Hangzhou or Beijing API for cost savings can recreate the extraction problem Karp attacks, only with a different jurisdiction. Self-hosting open weights is the alpha-preserving path; cheap foreign APIs are often just another meter with a different flag.

3. **Revoke resistance versus quality theatre.** US access limits on closed models made "a model no agency can revoke" attractive in June 2026 reporting. That is a real resilience argument for open weights. It is not a licence to skip evaluation parity, malware/supply-chain checks on downloaded checkpoints, or the serve-recipe sizing that large MoE models (GLM-5.2 class) demand. Under-sized trays with Chinese open weights are still under-sized trays.

For shareholders, the sober synthesis is hybrid: closed US or EU-managed endpoints where contractual controls and quality justify the meter; self-hosted DeepSeek-V4 / GLM-5.2 / Qwen-class weights where alpha and residency dominate; never a silent mix. Chinese open models are leverage for privacy economics and exit optionality. They are not a free pass on supplier risk.

Sovereign AI for Europe and peer countries

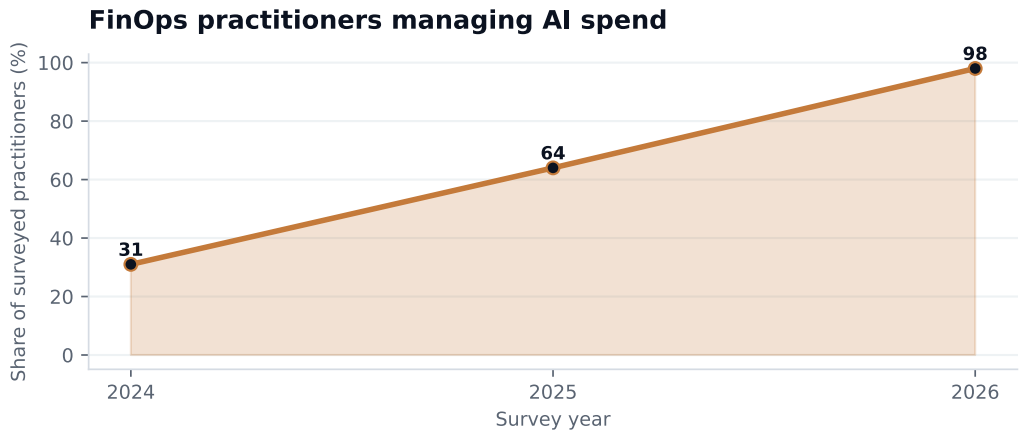
Sovereignty language is spreading because jurisdictions disagree about who may see data and who may depend on a foreign platform. OECD writing on government AI adoption notes that administrations combine commercial GenAI products with self-hosted open-weight systems when they want stronger control over data handling, customisation, and reduced single-vendor dependence. That pattern is not EU-only. It travels to any country that treats domestic compute, regulated sectors, or national security adjacency as board-relevant. Mid-2026 open weights from DeepSeek, Zhipu, and Alibaba make the "self-host" half of that OECD pattern operationally real for firms that can staff MLOps; they also import an origin decision the OECD chapter does not make for you.

Under EU rules, GDPR transfer constraints and the AI Act's role-based obligations raise the cost of vague placement. They do not, by themselves, tell a CTO which GPU tray to buy. Residency is not sovereignty. A model running "in Frankfurt" on a multi-tenant endpoint with training-use residue and weak exit terms can still fail a shareholder test. Conversely, a dedicated tray in a peer jurisdiction with your keys, your logs, and your evaluation evidence can be operationally sovereign even when the silicon brand is global.

What travels across EU markets and other sovereign AI programmes:

Control question	Metered frontier API	Private managed endpoint	Dedicated / sovereign tray
Who holds weights?	Vendor	Vendor (usually)	You (open weights) or contracted exclusive
Who sees prompts by default?	Vendor ops boundary	Narrower contract + region	Your operators / your VPC
Cost shape	Variable, success-taxed	Hybrid	Floor OpEx + ops labour
Exit optionality	Weak without dual-run	Medium	Strong if eval and connectors are yours
Best fit	Low-sensitivity volume	Sensitive enterprise with weak MLOps	High-sensitivity alpha and regulated workloads

Plain read: sovereignty is a control stack, not a flag on a purchase order.



FinOps Foundation State of FinOps (survey years as published). AIMonger redraw.

Figure 2. Share of FinOps practitioners managing AI spend across recent survey years. In short: AI cost moved from niche to nearly universal FinOps scope. In practice: the board can no longer treat tokens as an engineering experiment outside financial control. Source: FinOps Foundation State of FinOps 2026 and Linux Foundation press release on the same survey (1 Feb 2026) - see References 6 and 7. AIMonger redraw.

Operating implications for CFO, CIO, and CTO

For the CFO

1. **Refuse token volume as the success metric.** Require cost per successful case by workflow, tagged so AP, support, or discovery spend can be reconciled monthly. Token charts celebrate vendor revenue. Unit economics tell you whether privacy or metering is earning its keep.
2. **Sign the twenty-four-month tray before hardware lands.** Idle GPU, MLOps labour, evaluation labour, and upgrade cadence belong in the same sheet as the accelerator sticker. Capex slides without idle and people are how privacy programmes over-promise and under-cash.
3. **Set ceilings with routing, not panic alone.** Pair monthly budgets with task-tier routing so cost control does not become a quality cliff that drives shadow spend. A hard cap without a cheaper official tier is how staff buy consumer tools on personal cards.
4. **Own the compare: meter versus tray versus private managed.** Ask for the quiet-month and the high-volume-agent month, including the value of non-leakage and exit optionality. The cheap API wins on quiet months. Alpha workflows often lose that comparison once volume and retention risk are honest.
5. **Put a finance owner next to the technical owner on every production path.** Someone must answer for breaches of ceiling, untagged cloud lines, and shadow AI cards. Aggregate “innovation budget” ownership is how month-end surprises arrive without a name attached.

> **On the P&L:** if FinOps peers already manage AI spend and you cannot tag by workflow, you are behind peer practice, not ahead of an engineering experiment.

For the CIO

1. **Data classes before tooling.** Map confidential commercial, regulated personal, and public-content classes to allowed inference paths so teams cannot silently widen access. Without the

map, every pilot invents its own privacy story and the audit file becomes a collage. A mid-market insurer that writes three classes and two allowed placements stops arguing case by case in the security queue.

2. **Contractual residue is architecture.** Retention, training-use, subprocessors, and log access belong in the architecture review, not only in legal afterthought. Enterprise SKU names are not controls. The CIO who cannot answer where prompts are cached cannot defend the control narrative to the board.

3. **Shadow AI is a placement failure.** When the official private path is too slow or too narrow, staff buy consumer tools with personal accounts. That path often leaks more than the managed API security refused. Funding a usable private path is cheaper than pretending prohibition is a control.

4. **Evidence packs for shareholders.** Boards will ask for proof: inventory of models, data classes, owners, and incident paths. NIST AI 600-1 style inventory thinking helps even when the firm is not a US federal supplier. Evidence is how privacy spend becomes an asset narrative rather than a cost complaint.

For the CTO

1. **Price trays with finance in the room.** Build a twenty-four-month model for dedicated options: GPU hours, idle capacity, storage, MLOps labour, evaluation labour, and upgrade cadence. Sticker price on hardware is the slide that fools committees. Idle and labour are the lines the CFO will challenge first, and they decide whether private capacity beats a private managed endpoint.

2. **Keep evaluation parity across paths.** The same proposition holdouts must run on metered, private managed, and dedicated paths. Otherwise "sovereign" becomes an excuse for worse quality, and users route back to the frontier tool. Parity is what makes hybrid estates governable.

3. **Design for model switching.** If the application layer can swap open-weight and commercial models behind the same tools and permissions, the firm can negotiate. If every workflow hard-codes one vendor SDK, the extraction model has already won. Agnostic routing is a control, not a nice-to-have.

4. **Staff the patching burden.** Unpatched local CUDA, stale weights, and missing malware/supply-chain checks are how ENISA-style local risk shows up indoors. A dedicated tray without an on-call owner is concentration of risk, not reduction of risk.

Counter-position

Another board might argue that frontier APIs with strong contractual zero-training clauses, regional endpoints, and private networking are enough, and that dedicated GPUs are vanity for firms without MLOps depth. That position can be right for low and medium sensitivity workloads, especially where quality gaps between open weights and frontier models still matter commercially.

The counter fails when the firm's alpha is the prompt. If agents repeatedly expose process logic, pricing logic, clinical or industrial detail, or litigation strategy to a platform whose long-term incentive is to meter and learn from usage, the shareholder risk is not next month's invoice. It is five years of thinner differentiation. Hybrid is the adult answer: metered where alpha is thin, private where alpha is the product.

Decision criteria for CFO, CIO, and CTO (board-ready)

1. **Name the alpha at risk for each top workflow.** If leakage would change competitive position or regulatory exposure, the workflow is a privacy-premium candidate, not a token sandbox.

2. **Publish a three-path placement rule.** Metered API, private managed endpoint, dedicated/sovereign tray, with examples the board can understand in one page.
3. **Show twenty-four-month cash for the private path, signed by finance.** Include idle GPU, operators, evaluation, and upgrade labour. Refuse hardware-only slides.
4. **Report cost per successful case, not tokens alone.** Finance and engineering share one monthly pack: tier mix, ceiling breaches, and shadow spend outside the router.
5. **Set token ceilings with routing, not panic caps alone.** Pair monthly budgets with task-tier routing so cost control does not become a quality cliff.
6. **Require dual-run exit evidence.** Before concentrating a material workflow on one vendor, prove you can move it in a fixed number of weeks with quality parity.
7. **Assign named owners, including a finance owner for spend.** Every production inference path has a human accountable for cost, patching, and incidents. Aggregate "cloud team" ownership is not ownership.
8. **Report to shareholders in control language.** Inventory, data classes, spend by path, and residual training-use risk. Avoid both panic and marketing fog.

FAQ

Is private AI always more expensive than APIs? Per quiet month, often yes. Per year of high-volume agentic use against sensitive corpora, not always, and the comparison is incomplete without the value of non-leakage and exit optionality. The CFO should demand both months on one sheet before approving either path.

What does finance refuse at the steering gate? Hardware-only sovereignty slides, token volume as a KPI, untagged cloud AI lines, and programmes with no cost-per-successful-case owner. If those four appear, the privacy premium is not yet a capital decision.

Do open-weight models solve sovereignty? They solve weight possession only if you can run, secure, evaluate, and update them. As of July 2026 that includes competitive options such as DeepSeek-V4, Zhipu GLM-5.2, and Alibaba Qwen3.x (hosted preview versus downloadable weights differ by release). Weights without operators are a download, not a programme. Chinese-lab origin still needs an explicit board policy even when inference is fully local.

Should every country build national AI stacks? National programmes can change supply. Enterprise boards still decide workflow by workflow. Borrow the control questions; do not wait for a flagship national model to make the first placement decision.

Where does Karp overreach? Wherever the argument implies that only one vendor's application layer can make models safe. The durable point is control of compute, data, weights, and alpha, not loyalty to any single ontology brand.

References

1. CNBC, "Palantir's Karp bashes token-based AI model as 'completely wrong'" (1 July 2026). https://www.cnbc.com/2026/07/01/palantir-karp-open-ai-anthropic-tokens.html
2. CNBC Television, "Palantir CEO Alex Karp says 'something has gone completely wrong' with how AI is sold" (YouTube). https://www.youtube.com/watch?v=0A3sGymV6kY
3. SiliconANGLE, "Palantir CEO Alex Karp doesn't hold back in interview as he rails against AI industry" (1 July 2026). https://siliconangle.com/2026/07/01/palantir-ceo-alex-karp-doesnt-hold-back-interview-rails-ai-industry/

4. Eurostat, “20% of EU enterprises use AI technologies” (11 December 2025). https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20251211-2

5. Eurostat, “The use of artificial intelligence (AI) technologies in the European Union” (KS-01-26-009, 2026). https://ec.europa.eu/eurostat/documents/7870049/23260410/KS-01-26-009-EN-N.pdf

6. FinOps Foundation, State of FinOps 2026. https://www.finops.org/insights/state-of-finops/

7. Linux Foundation / FinOps Foundation, “State of FinOps Survey: AI Value and Skills Top Priorities” (1 February 2026; 98% manage AI spend). https://www.linuxfoundation.org/press/state-of-finops-survey-ai-value-and-skills-top-priorities-as-finops-matures-across-technology-value-98-manage-ai-90-saas-64-licensing-48-data-center-1

8. McKinsey & Company / QuantumBlack, “The State of AI: Global Survey 2025.” https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

9. Deloitte AI Institute, “The State of AI in the Enterprise: The Untapped Edge” (2026 edition). https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html

10. Stanford HAI, “The 2026 AI Index Report.” https://hai.stanford.edu/ai-index/2026-ai-index-report

11. OECD, “Digital Government Outlook 2026,” chapter “Adopting and governing AI in government.” https://www.oecd.org/en/publications/2026/06/digital-government-outlook_4585678e/full-report/adopting-and-governing-ai-in-government_7ef312a9.html

12. OECD AI Principles. https://oecd.ai/en/ai-principles

13. Regulation (EU) 2024/1689 (AI Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

14. GDPR Chapter V (transfers). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504

15. 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

16. ENISA, “Artificial Intelligence Cybersecurity Challenges” (publication). https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

17. OWASP GenAI Security Project, “OWASP Top 10 for Large Language Model Applications.” https://owasp.org/www-project-top-10-for-large-language-model-applications/

18. getdeploying, NVIDIA B200 cloud pricing comparison (commercial aggregator; on-demand avg $6.89/GPU-hr, spot avg $3.90, range ~$3.35-$16.11; retrieved 24 July 2026; not a forecast). https://getdeploying.com/gpus/nvidia-b200

19. OpenAI, API pricing (GPT-5.6 Sol $5/$30 per 1M input/output tokens; retrieved July 2026; list prices churn). https://developers.openai.com/api/docs/pricing

20. DeepSeek, Models and pricing (V4 Flash $0.14/$0.28 per 1M input/output; retrieved July 2026; list prices churn). https://api-docs.deepseek.com/quick_start/pricing/

21. 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

22. Verax, “The End of Tokenmaxxxing (Or How I Learned to Innovate Without Endless AI Spend)” (industry commentary on metered AI spend culture; not independent economic

evidence). https://www.verax.ai/blog/the-end-of-tokenmaxxxing-or-how-i-learned-to-innovate-without-endless-ai-spend

23. BRM, "AI Token Spend 101: A Finance Leader's Guide" (practitioner guidance on invoice-as-discovery vocabulary; not proof of extraction). https://www.brm.ai/blog/how-to-manage-ai-token-spend

24. CNBC, "China's Zhipu is booming with Anthropic and OpenAI held back" (26 June 2026; journalism, not primary model proof). https://www.cnbc.com/2026/06/26/china-zhipu-z-ai-open-source-anthropic-openai.html

25. CNBC, "Chinese AI models gain ground with U.S. companies as costs surge" (7 July 2026; journalism, not primary model proof). https://www.cnbc.com/2026/07/07/chinese-ai-models-costs-us-openai-anthropic.html

26. DeepSeek, "DeepSeek V4 Preview Release" (24 April 2026; as of July 2026). https://api-docs.deepseek.com/news/news260424/

27. Z.ai / Zhipu, "GLM-5.2: Built for Long-Horizon Tasks" (16 June 2026; as of July 2026). https://z.ai/blog/glm-5.2

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