tanitdata
OPEN DATA · ARTIFICIAL INTELLIGENCE · PUBLIC POLICY

POLICY BRIEF · TD-PB-2026-02

Importing Intelligence

Token dependence, chokepoints, and what production-centric AI sovereignty misses.

A policy brief on the political economy of LLM token flow

AUTHORS

Tarek Gasmi · Slim Abdelbari

PUBLISHED

June 2026

EDITORIAL STATUS

Independently produced; not externally peer reviewed

VERSION

v1.0

KEY FINDING

AI dependence is concentrating less in who produces models than in the flow layer — the gates every inference request must clear (supported-country rules, payment rails, rate limits, contracts) — which production-centric sovereignty spending does not reach.

CORE RECOMMENDATION

Token-importing states should pursue resilience over self-sufficiency: map token dependencies by flow type, build tested substitutability, and treat routing, aggregators, and compute-payment rails as critical infrastructure.

tanitdata — autonomous studio · Tunis, Tunisia
tanitdata.org · contact@tanitdata.org

Executive Summary

The geopolitics of artificial intelligence is debated almost entirely as a production problem: who makes the chips, builds the data centers, trains the frontier models. The more immediate exposure for most countries sits elsewhere, in the flow — this brief calls it the token flow problem. The 2026 Hormuz disruption restated the oil century's version of that lesson; this brief makes the equivalent analytical move for AI.

The LLM market presents a paradox. Model catalogs are diversifying and commodity-tier token prices have collapsed, roughly 600-fold since 2020 (Du, 2026). Yet enterprise API dollar flow has grown *more* concentrated: an estimated 88% lands with three U.S. providers (Menlo Ventures, 2025 — Menlo is an investor in Anthropic). The mistake is to read the first fact as fading dependence. The leverage has moved, from model production to the recurring passage of inference requests through enforceable gates: account eligibility, supported-country rules, payment rails, rate limits, routing platforms, contracts, jurisdictional permissions. The gates are documented, not hypothetical: providers publish the country lists, and payment methods from outside them are grounds for blocking.

Dependence is not one flow but three (consumer, developer, enterprise), each with distinct chokepoints. For token-importing economies, including most of MENA, the exposure concentrates on the consumption side, beyond the reach of production-centric sovereignty spending. The agenda is resilience over self-sufficiency: map token dependencies by flow type; build tested substitutability, not just redundancy; exploit the open-weight price window through governed shared serving capacity; treat the flow layer of routing, aggregators, and compute-payment rails as critical infrastructure.

KEY TERMS

Inference: running a trained model to generate outputs. **Token:** the unit of text a model processes; the unit in which inference is priced and metered. **API:** the machine-to-machine interface through which applications call hosted models. **Aggregator / router:** a service that routes requests across many models and providers. **Open weights:** model parameters that can be downloaded and served outside the original provider, subject to license and hardware. **HHI:** the Herfindahl–Hirschman Index, a market-concentration measure; higher means more concentrated. **Multi-homing:** using several providers simultaneously. **Substitutability:** the *tested* ability to switch providers quickly without unacceptable capability loss.

STATUS OF EVIDENCE

Market shares, provider country lists, payment rails, token prices, and legal regimes are time-sensitive. All such claims are current to June 12, 2026 unless otherwise noted, and stamped to their observation dates throughout. Per-cell sources and access dates for the availability and payment matrices are maintained in the brief's research repository (see *Sources and Notes*).

1. The Issue: A Production-Centric Blind Spot

In February 2026, the lesson the oil century taught repeatedly stopped being a reminder and became an event. The Strait of Hormuz — through which 20.9 million barrels of oil per day passed in the first half of 2025, about one-fifth of global petroleum liquids consumption (U.S. EIA, 2026) — was disrupted by regional conflict; as of June 2026, EIA estimates production shut-ins exceeding 11 million barrels per day, with flows assumed to resume only gradually in the second half of the year (EIA STEO, June 2026). Control over flow, the narrow points a commodity must pass through, confers leverage that ownership of production does not capture, and no producer-side statistic predicted what disrupting the strait would do. Every Hormuz figure in this paragraph is now a pre-disruption baseline. That is the point.

The AI policy debate of 2024–2026 has been, with few exceptions, a production debate. Export controls target chips; national AI strategies fund compute clusters and sovereign foundation models; "sovereign AI" offerings bundle local data centers with locally hosted models. The underlying theory of vulnerability is extraction-shaped: whoever owns the means of producing intelligence holds the power. This brief argues the leverage sits elsewhere — in the gates every hosted inference request must clear (Figure 1).

Consider one such request. A three-person startup in Tunis has built a document-processing tool for regional clients; each job sends a few hundred thousand tokens to a frontier model and streams the answer back. Before a single token returns, the request must clear a series of gates, none of which the startup controls. The account must exist, which requires the provider to serve Tunisia at all: providers publish supported-country lists, and access from outside them is stated grounds for blocking or suspension (OpenAI, 2026; Anthropic, 2026). The account must be funded, which in Tunisia is a binding constraint: dinar cards are not approved for foreign-currency transactions, and neither Stripe nor Adyen onboards businesses anywhere in North Africa (central-bank and processor documentation, as of June 2026). The request must fall within a rate-limit tier set by the provider's capacity-allocation policy, a tier that itself advances with cumulative payments. The model must be available in the startup's jurisdiction; the use must comply with terms defined under foreign law. Only then does the request travel — national gateway, subma-

rine cable, European edge node, the provider's private network — and the route itself is not guaranteed: when cables were cut in the Red Sea in September 2025, cloud traffic across the Middle East degraded within hours (Section 4). Every gate is ordinary, defensible, and open on most days. That is precisely what the oil century teaches about chokepoints: their politics are invisible until the flow matters more than usual, or the gate-keeper's incentives change. No national statistic captures this exposure, no production figure, no data-center count. The startup's dependence lives in the passage, and this brief is about mapping it.

A growing body of work complicates the production frame: full-stack AI sovereignty is structurally infeasible for almost any country (Tanner, Kerry, et al., 2026); deployment-layer control sits atop U.S.-controlled dependencies — the "sovereignty gap" (Chavez, 2026); vendors selling "sovereign AI" back to states define it on their own terms (Yew et al., 2026). These analyses agree that the dependencies that matter are not resolved by buying production capacity. What none provides is a systematic account of the *flow itself*: how inference moves, through which channels, under whose control, with what options for rerouting. That is the gap this brief addresses.

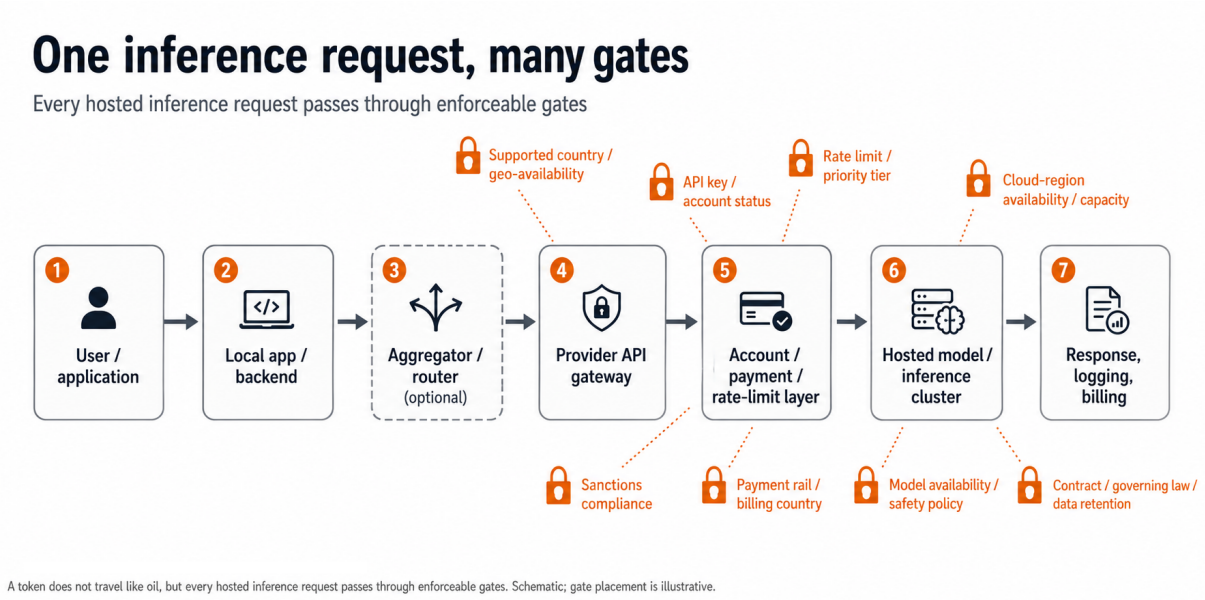


Figure 1. One inference request, many gates. A token does not travel like oil, but every hosted inference request passes through enforceable gates. Schematic; gate placement is illustrative.

2. Framework: The Form of Flow

The method comes from the political economy of energy. Timothy Mitchell's *Carbon Democracy* (2011) showed that the political possibilities of the coal and oil eras were shaped less by who owned the resource than by the material form of its flow. Coal moved through branching networks — many small veins feeding a few trunk lines — creating chokepoints where concentrated labor could interrupt the flow and extract democratic

concessions. Oil was redesigned around that vulnerability: fluid, capital-intensive, moving through networks "where there is more than one possible path and the flow of energy can switch to avoid blockages" (Mitchell, 2011: 38). The transition did not eliminate concentration; it *relocated* it — from the mine and the railhead to the corporate-state cartels governing routes and prices.

Three analytical tools carry over to information infrastructure:

- **Form of flow:** how the commodity is produced, metered, routed, and consumed — the unit of analysis is the channel, not the producer.
- **Control points:** "the narrow points of passage where control is particularly effective" (Mitchell, 2011: 7). A chokepoint is not a topological fact alone; it becomes politically consequential when actors can act on it.
- **Political affordances:** what an infrastructure's form *lets* differently positioned actors do — not what it forces them to do.

Applying this method across the history of computational infrastructure reveals three eras, each with a distinct form of flow and a new layer of concentrated power (Figure 2 and Table 1).

The internet regime (1990s–2000s). The commodity was connectivity; the flow was packets over a deliberately redundant, open-protocol mesh. Control points existed — backbones, exchange points, cables, the DNS root — but were thinly activated; the era's celebrated affordance was distribution.

The cloud regime (2010s). Computation and storage became a rented service, reorganized around hyperscale data centers owned by a handful of firms. Concentration moved from connectivity to *hosting* — who runs your workloads, under which contracts and jurisdictions (Srnicek, 2017; Narayan, 2022) — and the same hyperscalers then extended this power into AI through vertical integration of the model stack (Luitse, 2024).

The LLM regime (2022–present). The commodity is now algorithmic reasoning itself, metered and sold per token: priced per million tokens, input and output metered separately, quality tiered, prices set in a competitive spot-like market. The flow runs from an application, often through an aggregator, to a hosted model and back, every leg governed by API terms, rate limits, geographic availability rules, payment rails, and contract. And the volume sits on the serving side: as of mid-2025 a majority of surveyed organizations reported inference workloads exceeding development and training workloads (Menlo Ventures, 2025b — survey-based; no dollar-crossover figure is reported), across 2025–26 every major provider disclosed serving-capacity rationing,[1] and institutional estimates attribute between one-third and roughly two-thirds of AI compute to inference, a range reflecting definitional spread rather than doubt about direction (IEA, 2026).

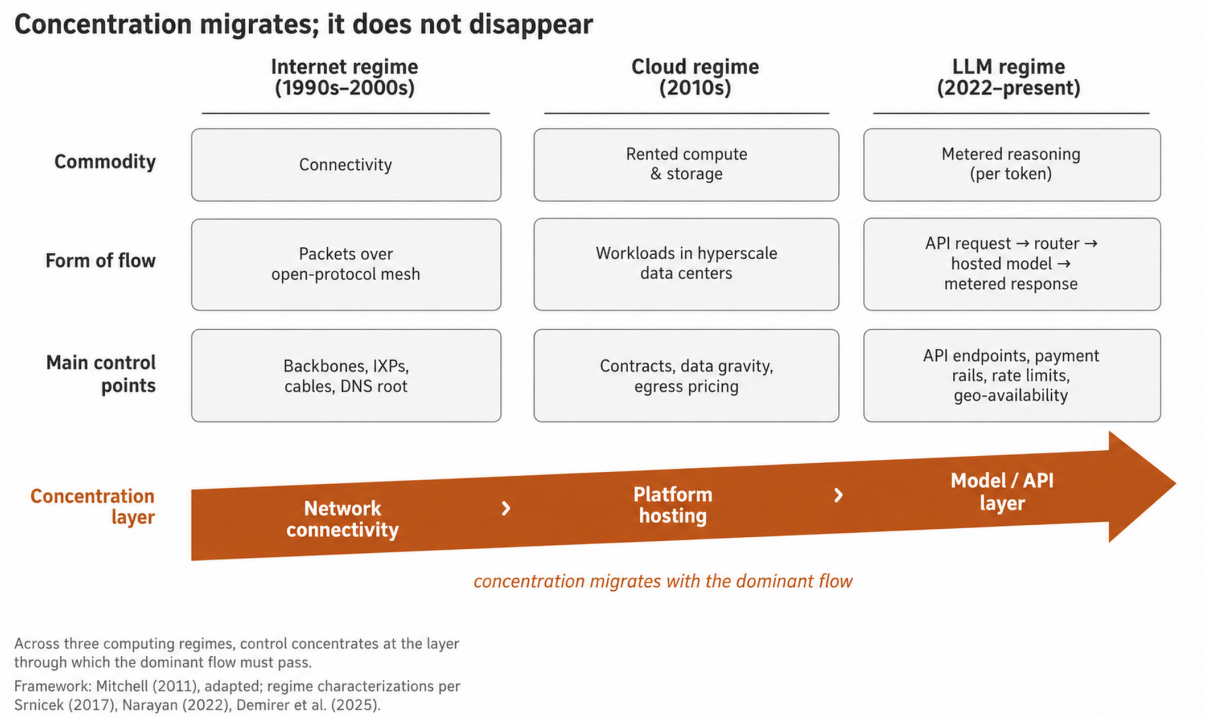


Figure 2. Concentration migrates; it does not disappear. Across three computing regimes, control concentrates at the layer through which the dominant flow must pass. Framework: Mitchell (2011), adapted; regime characterizations per Srnicek (2017), Narayan (2022), Demirer et al. (2025).

	Internet regime (1990s–2000s)	Cloud regime (2010s)	LLM regime (2022–present)
Commodity	Connectivity	Computation and storage as rented service	Algorithmic reasoning, metered per token
Form of flow	Packets over a redundant, open-protocol mesh	Workloads in hyperscale data centers	API request → (aggregator) → hosted model → metered response
Control points	Backbones, IXPs, undersea cables, DNS root	Hyperscaler platforms, contracts, data gravity, egress pricing	API endpoints, aggregator routing, payment rails, geo-availability
Concentration layer	Network connectivity	Platform hosting	Model / API layer

Table 1. Three regimes of computational flow. Framework: Mitchell (2011) adapted to information infrastructure; cloud-regime characterization per Srnicek (2017) and Narayan (2022); LLM-regime supply chain per Demirer et al. (2025).

Training remains a strategic chokepoint; chip export controls are proof that episodic production events attract chokepoint politics of their own. But training is episodic and hosted inference is recurrent: every request, every hour, must pass through an account, a gateway, a payment rail, a rate limit, and a jurisdictional rule. That recurring passage

creates an additional layer of leverage — continuous, fine-grained, contractual — that production-side policy does not capture and production-side statistics do not measure.

BOX 1 · WHERE THE OIL ANALOGY BREAKS — AND WHERE IT HOLDS

Tokens are not oil, and the differences are not details. Information is non-rivalrous: model weights can be copied at near-zero marginal cost, downloaded across any border, and served on hardware the importer controls. There is no tanker to intercept; you cannot embargo arithmetic. A small open model on a laptop completes the disanalogy — some inference now requires no flow at all.

The analogy survives because the commodity economies actually depend on is narrower than "information." It is *hosted frontier inference*: served tokens at a given quality, latency, and reliability, under contract, at scale. That service is rivalrous at the point of inference — serving capacity is finite, rationed visibly through rate limits, usage tiers, and priority contracts — and it is allocable: someone decides whose requests are served when capacity binds. Downloaded weights blunt the embargo logic only at the commodity tier; the frontier tier, where prices resist deflation (Section 3.2) and high-value applications concentrate, exists only as a hosted, gated service. Edge inference is a real resilience margin, not an exit. The right analogy is therefore not oil-the-substance but oil-the-logistics: production capacity versus passage control. What flows can be cut off not by seizing the cargo, but by closing the gates the cargo must pass through.

Mitchell's third tool, political affordances — what an infrastructure's form lets differently positioned actors *do* — can now be applied directly. Scarcity in this regime originates in production (GPUs, power, cooling) but is exercised at the flow layer, and each tier grants a distinct power. Model creators hold allocation power: when capacity binds, rate limits, usage tiers, and reserved-capacity contracts decide whose requests are served, openly and contractually (provider documentation, June 2026). Aggregators hold routing power, and with it both a resilience function and a new single point of failure. Payment processors and sanctions authorities hold exclusion power that operates without any AI-specific decision being taken. Token-importing developers hold a thin rerouting power: VPNs, resellers, fallback models, workarounds that exist precisely because the gates do. Enterprises gain leverage only to the degree they can credibly multi-home and switch, which the evidence shows few yet can (Section 3.4). The affordance map, in short, favors the gate-keepers over both producers and consumers of the flow — which is why this brief's policy agenda targets the gates.

3. Evidence: What the Token Flow Actually Looks Like

3.1 The creator tier: where the enterprise dollars land

The most direct measure of leverage is where the money lands — one measure among several; availability, routing, and contract control matter too. Menlo Ventures' end-2025 enterprise survey estimates LLM API spend shares at Anthropic 40%, OpenAI 27%, Google 21%: a combined 88% (estimated dollar-spend shares, triangulated with public financials per Menlo's methodology note), the remaining 12% spread across Meta's Llama ecosystem, Cohere, Mistral, and a long tail (Menlo Ventures, 2025a; Figure 3). Nor is this transient: summing the report's stated 2023 per-provider shares (12/50/7) yields roughly 69%, so the top three's combined share — a derived figure — rose from about 69% to 88% even as the lead changed hands within the oligopoly. Leadership churn atop durable concentration is precisely what distinguishes a structural chokepoint from a temporary market position. In coding, the largest single enterprise use case, Anthropic alone holds an estimated 54%; enterprise model-API spend reached $8.4 billion by mid-2025, from $3.5 billion in November 2024 (Menlo Ventures, 2025b).

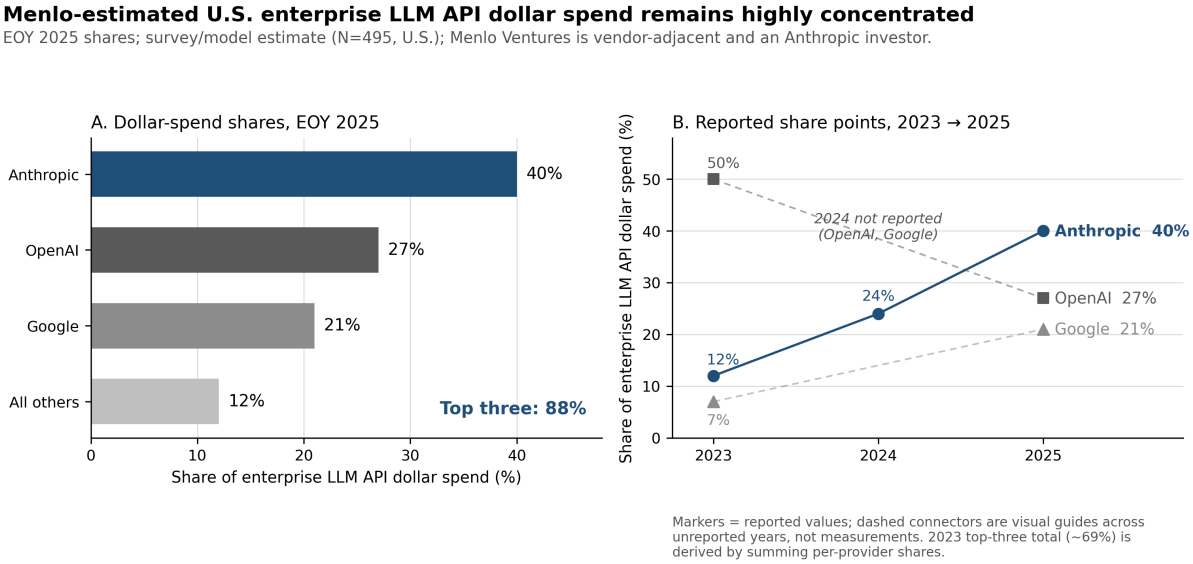


Figure 3. Provider concentration in enterprise LLM API dollar spend (Menlo Ventures estimate, EOY 2025; survey-based, N=495, U.S.; vendor-adjacent, as Menlo is an Anthropic investor). Markers are reported values; dashed connectors span years with no reported figure. The 2023 top-three total (≈69%) is derived by summing per-provider shares.

Three caveats cut in different directions. First, Menlo is an investor in LLM companies

including Anthropic; its figures are survey-based (N=495, U.S.-only, November 2025) and vendor-adjacent. Second, no independent source replicates Menlo's exact metric: 40/27/21 is the only published point estimate of that quantity, not a consensus figure. Third — constructively — the *concentration* triangulates well: a16z's January 2026 survey of Global 2000 CIOs independently places roughly 90% of enterprise model spend with the same three providers, while inverting the ordering (OpenAI ∼56% wallet share). The two surveys are vendor-adjacent in opposite directions, Menlo backing Anthropic and a16z backing OpenAI, and each finds its portfolio company ahead. The honest summary is the pattern already named above: contested ordering, agreed concentration (∼88–92%), churn atop a durable structure. (Menlo's mid-2025 report measured *usage*, not dollars — Anthropic 32%, OpenAI 25%, Google 20%, Meta 9%; the metrics are not comparable and not mixed here.)

3.2 Serving capacity and price: cheap tokens, sticky frontier

Du (2026), a preprint integrating OpenRouter pricing, the Epoch AI panel, and 62 cross-validated milestones, estimates a roughly 600-fold decline in token prices between 2020 and 2026. The decline is tiered: in Du's estimates, economy-tier prices halve every 1.10 years and mid-tier every 1.55, far faster than Moore's Law, while flagship-tier prices show no statistically detectable exponential decay, sustained by a reasoning premium Du puts at roughly 31.5×.[2] The direction is independently supported: Epoch AI's capability-constant price series and raw milestones, from $60 per million tokens (GPT-3 davinci, 2021) down to $0.10–0.28 at the commodity tier today, tell the same story (Figure 4, milestones overlaid). The frontier's resistance is real but not absolute: flagship prices oscillate in a band rather than decaying smoothly, and discrete repricings — most notably OpenAI's 80% cut to o3 in June 2025 — show the frontier is not immune to competition. "No smooth exponential trend" is the defensible claim; "prices don't fall" is not. Du's further decomposition attributing the decline almost entirely to software, with near-zero hardware contribution, is a model-dependent preprint result with no independent replication; it is cited as Du's estimate, not established fact.

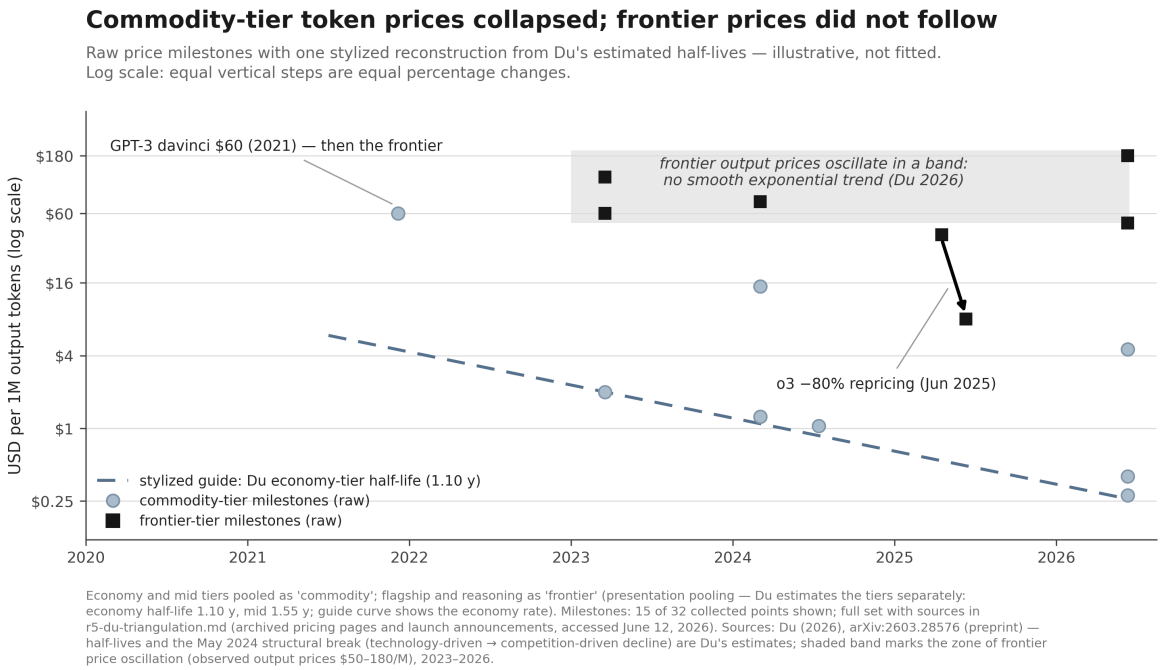


Figure 4. Commodity-tier token prices collapsed; frontier prices did not follow. Raw price milestones with one stylized reconstruction from Du's estimated half-lives — illustrative, not fitted; log scale. Economy and mid tiers pooled as "commodity," flagship and reasoning as "frontier" (presentation pooling). Source: Du (2026), arXiv:2603.28576 (preprint); 15 of 32 collected milestones shown, full set in the research repository (archived pricing pages, accessed June 2026).

The policy meaning is double-edged. Commodity-tier intelligence is becoming radically cheap — good news for token importers. But frontier reasoning capability, the tier high-value applications need, behaves like a differentiated good whose sellers — overlapping heavily with the firms capturing the enterprise dollars — have so far sustained its price. Dependence on the frontier tier is not being eroded by deflation.

3.3 The commodity is deflating, but the flow is thickening

Falling prices look like falling dependence. The opposite is closer to the truth, for a reason the energy century would recognize: when a useful input gets cheaper, economies do not pocket the savings — they embed the input deeper.

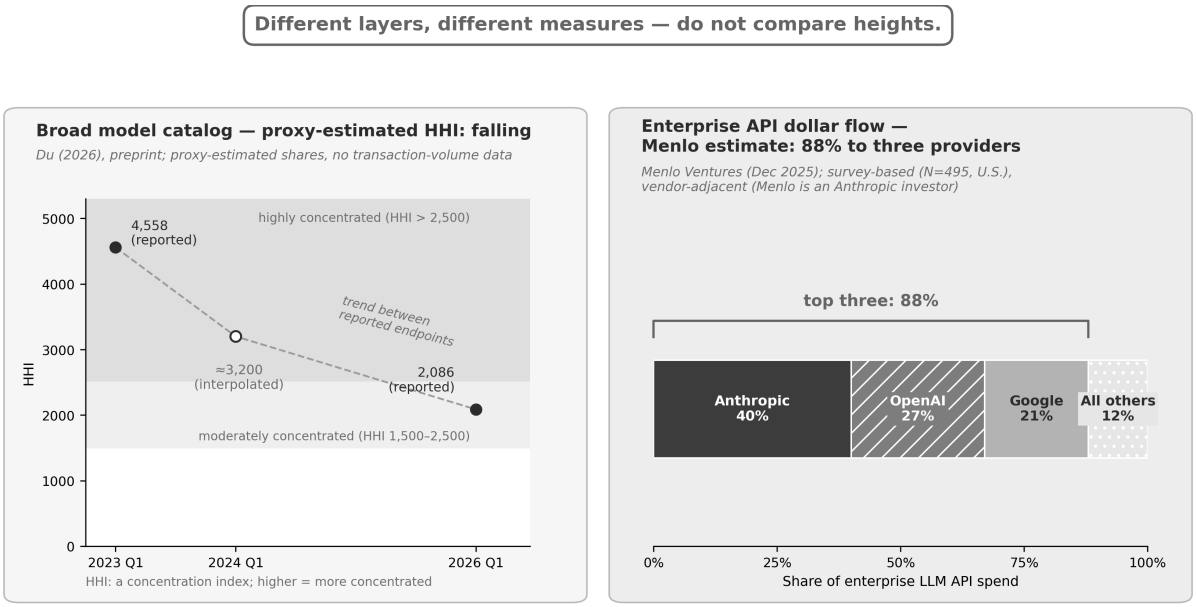
The mechanism is the shift from chat to agents. A chat exchange consumes tokens episodically: thousands per session, human-paced. An agentic workload — a coding agent iterating on a repository, a research agent traversing documents — consumes millions per task, machine-paced, in long autonomous loops. As unit prices fell, total consumption rose steeply: enterprise model-API spending more than doubled in eight months to $8.4 billion by mid-2025, and a majority of surveyed organizations reported

inference workloads now exceeding development and training workloads (Menlo Ventures, 2025b). Tokens routed through OpenRouter, a single developer-skewed aggregator, grew roughly tenfold in 2025, with reasoning-optimized models rising from a negligible share to more than half of its traffic; Google's disclosed monthly token volume rose from 9.7 trillion in April 2024 to over 1.3 quadrillion by October 2025; Microsoft reported volumes up fivefold even as its cost per token more than halved (OpenRouter/a16z, 2025; company disclosures, 2025–26). This is a Jevons-like dynamic, and the qualifier matters: Demirer et al. (2025) estimate short-run price elasticities just above one (−1.11, preferred specification) and read this as evidence *against* a short-run Jevons paradox, while leaving open the slower adoption channel the volume data show. What the evidence supports is *thickening*: cheaper tokens have made inference load-bearing.

Load-bearing changes what an interruption means: a chatbot outage is an inconvenience; an agent outage halts production workflows. The shift is a trajectory rather than an accomplished level — only 16% of enterprise deployments qualified as true agents in Menlo's end-2025 survey (27% among startups) — but the direction is what loads the flow, and agents deepen the stickiest layer of lock-in in this regime: scaffolding, prompts, and tool integrations tuned to one model's behavior, which no price comparison captures. Deflation is not eroding this dependence. It is financing its expansion.

3.4 The routing layer: diversity above, leverage below

The same market generates two headline statistics that appear contradictory (Figure 5). Du (2026) estimates that the HHI of the broad inference market fell from 4,558 to 2,086 over three years — crossing from "highly concentrated" to "moderately concentrated" on proxy-estimated shares, as open-weight entrants and price competition visibly diversified supply. Menlo's data show 88% of enterprise dollars flowing to three firms. These two metrics measure different realities. The falling index reflects a vibrant bazaar of models available to experimenters. The 88% reflects enterprise procurement: when large organizations deploy critical applications, the dollars flow almost exclusively to three U.S. providers. A falling index at the first layer says nothing about leverage at the second.



The apparent contradiction disappears once the measures are assigned to different layers of the supply chain.
Sources: Du (2026), arXiv:2603.28576, proxy-estimated HHI over OpenRouter-available models (card A); Menlo Ventures, 2025: The State of Generative AI in the Enterprise, Dec 2025 (card B).

Figure 5. Two layers, two measures. Broad-catalog concentration (Du 2026, proxy-estimated HHI) falls while enterprise dollar flow (Menlo estimate) stays concentrated; the panels measure different layers of the supply chain and are not comparable.

The layered structure follows the three-tier supply chain identified by Demirer, Fradkin, Tadelis, and Peng (2025): **model creators** at the top; **inference providers** (Azure, Cerebras, Together AI, Groq) operating the compute that serves models; and **aggregators** (OpenRouter and peers) routing demand across providers on price and latency. The creator tier is where dollars concentrate; the inference tier is intensely price-competitive; the aggregator tier is the most novel — simultaneously a resilience tool (one integration, many models, automatic failover) and a new single point of failure through which diversified demand passes. Notably, the largest aggregator imposes no country gates of its own; upstream restrictions pass through per-model (OpenRouter terms, June 2026) — the router inherits the gates rather than setting them.

Two further facts matter for policy. Demirer et al. report "open-source" models roughly 90% cheaper than closed-source at the same measured intelligence tier — what resilience operationally requires is the stricter category of *open weights*, parameters that can be downloaded and served independently. Yet open models' token share remains below 30%, a gap signaling differentiation (reliability, tooling, support) that benchmarks miss. And dependence is stickier than catalog diversity suggests: only 11% of surveyed builders switched primary vendors in the year to mid-2025 (Menlo Ventures, 2025b), and while multi-homing is rising, most firms keep a single primary model (Demirer et al., 2025). Redundancy is growing; *substitutability* — swapping providers in production at speed without capability loss — still lags. Lock-in is a gradient, with layers the cloud era lacked: prompts, fine-tuned behaviors, agent scaffolding tuned to one model's idiosyncrasies.

3.5 Not one flow but three

"U.S. dominance of AI" conflates at least three flows with different geographies and chokepoints (Figure 6):

- **Consumer flow** — end users of chat products, routed through a handful of U.S.-based applications. Chokepoints: app stores, consumer payment rails, content and compliance policy. (Circulating geographic estimates rest on unverified methodology and are excluded here; the jurisdictional point stands regardless.)
- **Developer flow** — API traffic from applications. Markedly more international: on OpenRouter's developer-skewed platform, over half of usage originates outside the United States (OpenRouter, 2025). Chokepoints: API geo-availability, developer payment rails, rate limits.
- **Enterprise flow** — contracted usage by firms, dollar-concentrated per the Menlo data. Chokepoints: contracts, data residency, cloud-region availability, compliance regimes.

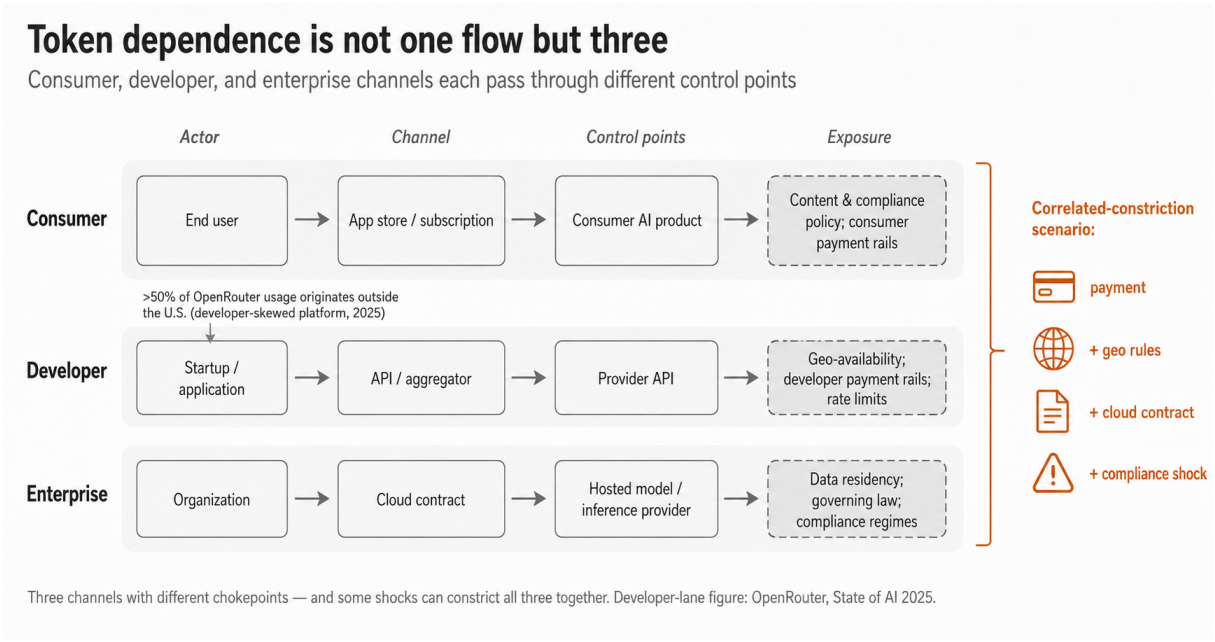


Figure 6. Token dependence is not one flow but three. Consumer, developer, and enterprise channels pass through different control points — and some shocks can constrict all three together. Developer-lane figure: OpenRouter, State of AI 2025.

Even the gates are not uniform across providers: as of June 2026, Yemen appears on the OpenAI, Google, and Microsoft country lists but not Anthropic's, and Sudan appears on the OpenAI, Anthropic, and Google lists but not Microsoft's — single-provider divergences that make "U.S. provider policy" a misleading aggregate (provider documentation, June 2026). Policy that treats the three flows as one will misdiagnose exposure. A country can host a sovereign data center (addressing enterprise-flow optics) while its developers

remain one payment-processor decision away from losing API access, and its consumers remain subject to availability decisions made elsewhere.

4. An Exposure Map: MENA as the Consumption-Side Case

This section is an exposure map: deductive in structure, anchored where the record permits in availability, payment, and incident evidence. MENA is not one AI market. It includes Gulf states investing in compute and national models (the production-side pole); middle-income developer economies that mostly import tokens (the primary concern here); comprehensively sanctioned or conflict-affected contexts where access rules bind differently; and public-sector adopters facing procurement and continuity questions.

The regional AI literature has concentrated on production governance and on ethics and bias — including the well-documented structural penalty Arabic suffers under tokenization schemes designed around English — but has left the consumption flow largely unexamined: how foreign tokens enter MENA economies, under which payment rails, through which jurisdictional envelopes, with what substitutability in a crisis.

At the provider layer, as of June 2026, fifteen of nineteen MENA jurisdictions appear on all four major U.S. positive country lists (OpenAI, Anthropic, Google, Microsoft); Yemen and Sudan each appear on three, excluded by a single provider apiece (Anthropic and Microsoft respectively — the divergences noted in Section 3.5); Iran and Syria appear on none, and Mistral, the European provider, names the same two exclusions in its terms: the EU/U.S. divergence many expect does not exist at this layer. At the payment layer the constraint is broader: Stripe onboards businesses in exactly one MENA country (the UAE), Adyen in none, and the Maghreb and Levant operate under documented central-bank restrictions on international card use — Algeria's effective bar, Libya's $2,000-a-year quota, Tunisia's exclusion of dinar cards from foreign-currency transactions, Lebanon's fresh-dollar regime (central-bank instruments, June 2026). Since every major provider gates API throughput behind cumulative-payment tiers, payment-rail access is a second-order chokepoint stacked on formal availability. Palestine is the distinctive case: rails technically functional, yet every major processor excludes West Bank and Gaza residents (7amleh, May 2026).

At the incident layer, no verified MENA-specific interruption at the payment or policy level is on record: the documented gates are standing mechanisms, not realized regional cutoffs. What is documented sits at other layers — the Red Sea cable cuts (below), Sudan's conflict-driven connectivity collapse of 2023–24, and Egypt's exclusion from ChatGPT's supported list from launch until November 2023: a gate, not an interruption.

The comprehensively sanctioned or conflict-affected category is the existence proof of

correlated constriction. Iran — the binding comprehensive closure — is simultaneously absent from every provider list, excluded from every payment rail, and barred by the Iranian Transactions and Sanctions Regulations: constriction at all three layers at once, by law. Syria is the transitional case, and an instructive one: the comprehensive U.S. sanctions program was revoked effective July 2025 and the Caesar Act repealed that December, yet as of June 2026 Syria remains absent from every major provider's supported list. The gates have outlived their stated rationale — a persistence consistent with compliance lag, residual targeted measures and the still-extant State Sponsor of Terrorism designation, or simple risk aversion; the record does not say which.

The strategic risk for non-sanctioned importers is correlation of a softer kind. Many model APIs, cloud contracts, app stores, and payment rails are governed through a small set of firms and jurisdictions; in a sanctions, export-control, fraud-control, or compliance shock, the three flows could tighten together, access interrupted as a byproduct of compliance and payment-risk decisions, with no actor intending a comprehensive cutoff. The jurisdictional instruments deserve precision rather than alarm. Under the CLOUD Act (18 U.S.C. § 2713), providers subject to U.S. jurisdiction must preserve and disclose data within their possession, custody, or control in response to lawful U.S. process, regardless of where the data is stored. The Act is narrower than its reputation (no blanket access, covered providers only, case-by-case legal process); its significance here is that exposure follows jurisdiction over the provider, not the location of the server. Export controls on inference are a foreseeable channel, not a binding regime: as of June 2026 no U.S. rule restricts inference-as-a-service to any jurisdiction — the 2025 AI Diffusion Rule was suspended before its compliance date — but the Remote Access Security Act, House-passed in January 2026 and awaiting Senate action, would create the statutory predicate under which inference access could become a licensable flow.

The flow layer also has a physical stratum, and for this region the two are co-located. On September 6, 2025, submarine cables (SMW4 and IMEWE, with FALCON GCX identified separately by Kuwaiti authorities) were cut near Jeddah, on a corridor carrying roughly 17 percent of intercontinental internet traffic; Microsoft's Azure status notice confirmed elevated latency as Middle East transit rerouted, and connectivity degraded from the Gulf to South Asia within hours (NetBlocks, 2025; archived Azure status notice; press accounts). Hosted inference rides the same cables. Nor does a gate require enforcement to move flows: when the Houthi leadership announced in February 2026 that Red Sea shipping attacks would resume, carriers began withdrawing voyages on the announcement alone — no confirmed strike on a merchant vessel had been independently verified through late March 2026 (EUNAVFOR ASPIDES, 2026) — chokepoint leverage exercised through credible threat rather than action. The point is not that tokens face a literal Hormuz; rerouting absorbed the September shock in days, as oil cannot. It is that MENA's token imports transit a narrow set of physical corridors *and* a narrow set of jurisdictional gates, governed by overlapping firms and states. Correlated passage is the region's inherited condition; the token flow is its newest layer.

The production-side response — sovereign compute — addresses little of this. A national GPU cluster does not keep a Cairo or Tunis startup's API account funded, change the governing law of a Riyadh enterprise's cloud contract, or provide a frontier-quality fallback when access tightens. The exposure is on the flow side; the spending is on the production side.

5. Policy Implications: Resilience Over Self-Sufficiency

The form-of-flow analysis converges with the "managed interdependence" position (Tanner, Kerry, et al., 2026) but sharpens it: managing interdependence requires mapping *flows*, not just suppliers. The recommendations below are ordered by feasibility, assigned to actors, and deliberately start with what a single ministry can begin within ninety days using existing powers — because the prerequisite dependency data does not yet exist. One gap deserves naming once: no incident-reporting channel exists for payment- or policy-layer interruptions of AI access — providers document their gates, but no one records when and where they bind. Visibility obligations would create that record; their absence is not evidence that incidents occur, only that no one would know.

1. Map token dependencies by flow type

LEAD ACTOR: Digital ministry / national statistics office.

FIRST 90 DAYS: Inventory public-sector AI systems by provider, contract, payment rail, jurisdiction, and criticality.

12–24 MONTHS: Annual national token-dependency report on the three-flow structure.

FEASIBILITY: High — a desk exercise within existing powers.

METRIC: Share of critical public AI systems with a documented dependency profile.

2. Build substitutability, not redundancy

LEAD ACTOR: Procurement authority + cybersecurity agency.

FIRST 90 DAYS: Insert provider-portability clauses in new public AI procurement; define per-function degradation thresholds via task-level evaluation.

12–24 MONTHS: Switching drills for critical functions, modeled on financial-sector stress tests.

FEASIBILITY: Medium — requires procurement reform.

METRIC: Time-to-switch and measured quality loss in drills.

3. Exploit the open-weight window via governed shared capacity

LEAD ACTOR: Regional organization, development bank, sovereign consortium, or public–private operator (candidate classes, not an endorsement; full design beyond scope).

FIRST 90 DAYS: Feasibility study: workload sizing, candidate models, host law.

12–24 MONTHS: Pilot regional inference facility with allocation, residency, and lawful-access rules defined before launch.

FEASIBILITY: Low–medium — governance is the hard part, stated as a condition.

METRIC: Tested fallback coverage: share of critical workloads with a drilled open-weight path.

4. Treat the flow layer as critical infrastructure

LEAD ACTOR: Critical-infrastructure agency + payments regulator.

FIRST 90 DAYS: Visibility: mandatory dependency disclosure for systemically important deployments; API-outage and cutoff incident reporting.

12–24 MONTHS: Continuity-of-service expectations for aggregators and compute-payment rails; regional mutual-aid arrangements.

FEASIBILITY: Medium — start with visibility obligations, escalate later.

METRIC: Incident-report coverage; continuity clauses in force.

On Recommendation 3, the conditions are the substance. Open-weight models can be a resilience resource only when the weights are actually held, served, tested, and integrated before a crisis; a hosted open model can still fail at the payment, cloud, or

routing layer. The target is a tested fallback stack: acceptable quality loss defined per critical function in procurement, validated in drills; serving capacity; portability rules. The 90%-cheaper finding makes this affordable; the sub-30% consumption share and 11% switching rate measure how much testing stands between the price window and an actual fallback. On Recommendation 4, the instrument ladder runs from visibility (dependency inventories, incident reporting) through continuity (service expectations, payment continuity, portability) to international arrangements: the analogue of the energy-security frameworks built after the oil shocks, covering availability assurances among token-importing states.

6. Conclusion

Each infrastructural transition of the past three decades has relocated the layer at which control over computational flow concentrates: from network connectivity, to platform hosting, to the model-API layer that now meters and routes machine reasoning itself. The current policy conversation, fixated on producing that capability, is analyzing the wells and ignoring the strait — in a year when an actual strait was disrupted and an announced threat alone rerouted shipping. The token market's economics — deflating commodity tiers, a price-resistant frontier, a diversifying catalog above a concentrated dollar flow, agents making the flow load-bearing — mean that for most countries the relevant question is not *how do we produce intelligence?* but *how do we keep it flowing on acceptable terms when conditions change?* That is a resilience question, answerable with instruments that exist today, beginning with a dependency map any ministry could start this quarter. And for token-importing regions, the 2026 Hormuz disruption removed the hypothetical: it is urgent.

About the Authors

Tarek Gasmi

FOUNDER, TANITDATA

Tarek Gasmi is Founder of Tanitdata and Head of Data and AI at datadoit, with an academic affiliation at the University of Manouba. His work focuses on AI governance, digital infrastructure, sustainability, and technology policy, with particular attention to transitional economies and the geopolitical economy of AI systems.

Slim Abdelbari

PARTNER, POLICY RESEARCH, TANITDATA

Slim Abdelbari is a Partner at Tanitdata overseeing policy research, with an academic affiliation at the University of Sousse. His work applies systemic analysis and system dynamics to geopolitics, with a focus on automation, sustainability, and policy frameworks for transitional economies.

Sources and Notes

Methods and repository

The two empirical artifacts behind Section 4 are a provider *availability matrix* (nineteen MENA jurisdictions × five providers — OpenAI, Anthropic, Google, Microsoft, Mistral) and a *payment-rail matrix* (card-acquirer and processor coverage against central-bank foreign-currency rules). Each cell records the status, the primary source (an official country list, terms page, or central-bank instrument), and the access date; provider-list cells are backed by archive snapshots for the core pages. Both matrices, the 32-milestone token-price dataset underlying Figure 4, and the figure scripts are maintained in the brief's public research repository at github.com/tanitdata/importing-intelligence. Claims in the running text are stamped to their observation dates; the repository is the per-cell system of record.

Notes

1. The familiar "up to 90% of ML compute cost is inference" claim is an AWS figure from the December 2019 launch of its Inferentia chip (Amazon press release and AWS News Blog, Dec. 3, 2019; recycled across AWS pages through 2020–2022, where it is usually encountered). It describes pre-LLM production machine learning, measures infrastructure cost, and carries an "up to" qualifier — evidence that inference dominance predates the LLM era, not a current LLM estimate. Current institutional anchors: IEA, *Energy and AI* (Apr. 2025) and the April 2026 update reporting AI-focused datacenter electricity up ∼50% in 2025; Epoch AI estimates ∼one-third of AI compute capacity serves inference (power basis), while operational-energy figures from Meta and Google reported by the IEA run 60–70% (older data); the spread is definitional (cost vs. energy vs. capacity).

2. Supporting statistics in Du (2026): flagship-tier exponential-decay fit $R^2 = 0.031$; structural break May

2024 (Chow F = 5.74, p = 0.005), which Du reads as the market's transition from technology-driven to competition-driven price declines; HHI series 4,558 (2023Q1) → 2,086 (2026Q1) on proxy-estimated shares — the paper has no transaction-volume data.

Data sources

Access dates 2026-06-12 unless noted. **Market concentration:** Menlo Ventures, *2025: The State of Generative AI in the Enterprise* (Dec. 2025; N=495, U.S.; dollar-spend shares per methodology footnote) and *2025 Mid-Year LLM Market Update* (Jul. 2025; N=150; usage shares, switching) — different definitions, not mixed; Menlo is an investor in LLM companies including Anthropic. **Triangulation:** a16z enterprise AI survey (Jan. 2026; a16z is an OpenAI investor); Ramp AI Index (2026). **Token economics:** Du, "Tiered Super-Moore's Law" (arXiv:2603.28576, 2026; preprint), triangulated against Epoch AI price analyses and archived provider price pages. **Market structure and elasticities:** Demirer, Fradkin, Tadelis, and Peng (NBER WP 34608, Dec. 2025, doi:10.3386/w34608; Microsoft/Amazon affiliations disclosed; elasticities labeled preliminary). **Platform flow:** OpenRouter, *State of AI 2025* (single-aggregator, developer-skewed, geography by billing entity) and 2026 disclosures; Alphabet and Microsoft investor communications (company-disclosed, unaudited). **Capacity and energy:** provider statements and rate-limit documentation (2025–26); Microsoft/Alphabet/Amazon earnings calls; IEA, *Energy and AI* (2025) and 2026 update; LBNL (2024). **Hormuz:** U.S. EIA, *World Oil Transit Chokepoints* (Mar. 2026) and *Short-Term Energy Outlook* (Jun. 2026). **Provider gates:** OpenAI, Anthropic (incl. Sept. 2025 ownership rule), Google, Microsoft, AWS, OpenRouter, Mistral, and Cohere official country lists and terms — accessed June 12, 2026; archive snapshots held for the core country-list pages, with the remainder recorded by access date in the research repository. **Payment rails:** Stripe, PayPal, Adyen, Checkout.com official lists; central-bank instruments (Banque d'Algérie 05-2025, CBL 10/2025, CBE Aug. 2025, BCT card rules, BdL); IMF Article IV reports; 7amleh (May 2026). **Red Sea:** NetBlocks (Sept. 2025); archived Azure status notices (the live status-history page no longer lists the incident); CNBC/BBC/Network World; EUNAVFOR ASPIDES (Mar. 2026). **Legal:** 18 U.S.C. § 2713; CRS R45173; DOJ CLOUD Act white paper (2019); BIS Diffusion Rule actions (2025); H.R. 2683 (House passage Jan. 2026); OFAC ITSR (31 CFR 560); EO 14312 (Jul. 2025); Caesar Act repeal (FY2026 NDAA).

Literature

Timothy Mitchell, *Carbon Democracy* (Verso, 2011), quotations at pp. 7, 38; Ashley Carse, Jason Cons, and Townsend Middleton, "Chokepoints" (*Ethnos*, 2020); Laleh Khalili, *Sinews of War and Trade* (Verso, 2020); Kate Crawford, *Atlas of AI* (Yale, 2021); Nick Srnicek, *Platform Capitalism* (Polity, 2017); Devika Narayan, "Platform capitalism and cloud infrastructure" (*Environment and Planning A*, 2022); Dieuwertje Luitse, "Platform power in AI" (*Internet Policy Review* 13(2), 2024); Brooke Tanner, Cameron F. Kerry, et al., "Is AI sovereignty possible?" (Brookings, Feb. 2026); Pablo Chavez, "The Sovereignty Gap in U.S. AI Statecraft" (*Lawfare*, Feb. 2026); Rui-Jie Yew, Kate Elizabeth Creasey, Taylor Lynn Curtis, and Suresh Venkatasubramanian, "The Commodification of AI Sovereignty" (arXiv:2601.11763, Jan. 2026).

Scope notes

This brief summarizes the analytical framework of a longer essay in preparation. Geographic estimates of consumer LLM traffic concentration in circulation (e.g., ∼85–90% U.S. shares of consumer web traffic) are excluded pending verification of methodology. All provider lists, payment-rail statuses, prices, and market shares are perishable and stamped to their observation dates; the full availability and payment matrices, with per-cell sources and access dates, are maintained in the brief's research repository.

Citation and disclosure

Suggested citation. Gasmi, T., & Abdelbari, S. (2026). *Importing Intelligence: Token Dependence, Choke-points, and What Production-Centric AI Sovereignty Misses.* Tanitdata Research policy brief TD-PB-2026-02, v1.0, June 2026. tanitdata.org.

Version. v1.0, June 2026 — supersedes v2, *The Flow Problem* (June 2026); the analysis is unchanged in substance, the title revised for standalone clarity.

Disclosure. Independently produced and published under the Tanitdata Research brand; no commissioning client, no client funding, and no client or personal data were used. The authors' commercial work is in data and AI advisory and digital-policy consulting; they hold no equity in, and received no compensation from, any model provider, inference platform, aggregator, or payment processor whose market position this brief assesses. Where the analysis relies on vendor-adjacent data — notably Menlo Ventures (an Anthropic investor) and a16z (an OpenAI investor) — that adjacency is disclosed at the point of use.