Research Article

Journal of Human Resource Sustainability and Organizational Studies

Tokens, Watts, and Geography
A Four-Tier Framework for Pricing the AI Inference Trade, 2026–2030

Djellal Djouad*

CrossVol Research, France

***Corresponding Author**
Djellal Djouad, CrossVol Research, France.

Submitted: 2026, Jun 19; **Accepted:** 2026, Jul 17; **Published:** 2026, Jul 28

Citation: Djouad, D. (2026). Tokens, Watts, and Geography A Four-Tier Framework for Pricing the AI Inference Trade, 2026–2030. *J Hum Res Sus Org Stud, 1*(2), 01-09.

Abstract

The sell-side consensus on AI infrastructure prices capex in absolute terms — six hundred billion here, seven hundred billion there — and prices inference output in narrative terms — "the price of intelligence is collapsing." Both framings miss the same thing. Inference output is a manufactured good. Its marginal cost is dominated by electricity. And electricity is not priced globally, it is priced by interconnect. This paper develops a four-tier framework for the geographic pricing of AI inference between 2026 and 2030. ***Tier 1 (Sovereign Capable)*** *— China, the Gulf, the Nordic countries, France — combines low marginal electricity cost with state-backed capacity expansion.* ***Tier 2 (Emerging Stretched)*** *— India, South Korea, Japan — has committed hundreds of billions to sovereign compute on grids that are not yet stable enough to clear it.* ***Tier 3 (Power Constrained Wealthy)*** *— the United States in the PJM interconnect, Germany, the United Kingdom — concentrates capex without the physical electricity to dispatch it; PJM has now cleared its annual capacity auction at or near the regulatory cap three years in a row, and the 2027/2028 auction came in short of its reliability target for the first time in the RTO's history.* ***Tier 4 (Arbitrage Suppliers)*** *— Russian Siberia, Canadian hydro, the Brazilian Norte/Nordeste, Australian Pilbara — owns the cheapest electrons on earth and is now exporting them, increasingly, to Tier 1 workloads via cross-border power agreements and dual-resident cloud structures.*

We derive a simple unit-economics identity for token output — marginal cost equals energy per token times utilization-adjusted delivered electricity price, plus capex amortization — and quantify the spread across tiers using 2024–2026 data. We then show that Shanghai Futures Exchange's preliminary AI-token futures, which most coverage treats as a domestic Chinese compute-hedging tool, will function instead as the world's first tradable instrument on the ***cross-tier electricity risk premium of AI inference****. We close with five testable predictions for Q3 2026 to Q4 2027. We are not arguing that one tier wins. We are arguing that pricing models which ignore the floor are mispriced by an order of magnitude that the market has not yet been forced to confront.*

1. The Question the Sell-Side is not Asking

Read any major bank's AI-infrastructure framework in the first half of 2026 and you will find a remarkably uniform picture. The four large US hyperscalers will spend somewhere between US$ 700 and US$ 725 billion on capex in calendar 2026, a year-on-year increase north of 70%. AWS turned over US$ 44.2 billion of that in the first quarter alone. Microsoft added US$ 30.88 billion in its fiscal third quarter, with AI revenue passing a US$ 37 billion annual run rate. Meta lifted its full-year guide to a US$ 125–145 billion range. Alphabet doubled its Q1 spend year-on-year while its Cloud backlog crossed US$ 460 billion. [1-3]. These numbers travel as a single narrative: scarcity of GPUs, scarcity of trained model capacity, an arms race for compute. The narrative is real. It is also incomplete. The narrative is incomplete because it counts the dollars and the chips and stops there. It does not count the watts. Compute is the visible bottleneck. Electricity is the structural one. And electricity in the United States, where most of that capex is being deployed, has now run into a wall that the same banks' utility-sector teams have been documenting in a different report all along.

PJM Interconnection — the regional transmission operator for thirteen states from Illinois to North Carolina, the largest power market in North America by served population — held its

2025/2026 Base Residual Auction in July 2024. Capacity cleared at US$ 269.92 per megawatt-day across most of the footprint, with the Baltimore Gas & Electric zone clearing at US$ 466.35 and Dominion at US$ 444.26 [4]. One year later, the 2026/2027 auction cleared at US$ 329.17 per megawatt-day across the entire PJM footprint, a FERC-approved record [5,6]. In July 2025 PJM's board called for a backstop reliability auction because the system was running short [7]. The 2027/2028 auction, cleared in December 2025, hit US$ 333.44 per megawatt-day — at or near the cap — and **PJM procured 6,623 megawatts less than its installed-reserve-margin target.** That is the first time the entire RTO, including the Fixed Resource Requirement areas, fell short of the reliability requirement [8,9].

Run the arithmetic on the capacity charge alone. At US$ 329 per megawatt-day, a 100-megawatt data-center campus is paying US$ 12 million per year just to reserve the right to draw power, before a single kilowatt-hour is consumed. New data-center load drove roughly 70 percent of PJM's demand-side increase in this last auction cycle, and PJM expects peak demand to grow by 32 gigawatts between 2024 and 2030 with 30 of those gigawatts coming from data centers [9,10]. Meanwhile, SynMax's satellite imagery of US data-center sites suggests that more than 60 percent of campuses pencilled into the 2027 commissioning schedule have not yet broken ground.

So the question the sell-side capex slide is not asking is the obvious one. **What is the price of a token of LLM output if you can't get the megawatt to produce it?**
This paper answers that question.

2. The Unit Economics of Inference

2.1. A Simple Identity

Let E_t be the energy, in joules, required to produce one output token from a given model running at a given utilization on a given hardware stack. Let η be the data-center power-usage-effectiveness ratio (PUE), so that 1 J of compute energy delivered draws η J at the meter. Let p_e be the all-in delivered electricity price, in US dollars per joule, faced by the data-center operator in a given region. Let C_a be the per-token allocated amortization of GPU, networking, cooling, and building capex. Then the marginal cost of one output token, ignoring margin, is

$$\mathrm{MC}_{\mathrm{token}} = E_t \cdot \eta \cdot p_{\mathrm{e}} + C_a$$

This identity is trivial. Its consequences are not. Two of the three terms are essentially **geographic**: η
depends on climate and design, pe depends on the regional power market. The third, C_a, is global in the

sense that an H100 costs roughly the same to buy in Texas and in Chongqing, but it is local in the sense that financing cost, tax treatment, and utilization rate vary by jurisdiction. A token is a tradable, fungible unit of intelligence. Its cost is not.

2.2. Calibrating the Terms

We need real numbers.

Energy per token, E_t. Recent measurements on H100-class hardware show that a well-batched Llama-3-70B inference in FP8 precision can be delivered at approximately 0.39 joules per output token, an order of magnitude better than the 3–4 J/token figures reported for first-generation 65B-parameter models in 2023 [11,12]. Mixture-of-experts architectures — DeepSeek-V3's 671 B parameters with 37 B activated per token, for example — give a roughly comparable per-token energy at scale because only the routed experts are loaded into active memory [13]. Frontier closed-weight systems (the GPT-5 family, the Claude 4 family, Gemini 3) are larger and operate at lower per-card utilization in production; we use a working range of $E_t \in [0.3,1.5]$ J/output token to span this distribution, with 0.5 J/token as our central case for a modern dense or active-MoE workload.

UE, η. Hyperscaler facilities now report fleet-wide annual averages between 1.09 (Google, 2025) and 1.16 (Microsoft, 2024). AWS reports a 1.15 average. The industry as a whole is closer to 1.56 (Uptime Institute 2024 global survey), but the entrants we care about — net new AI-scale builds — sit at 1.10 ± 0.05 [14-16]. Nordic free-cooling sites achieve below 1.10. Russian Siberian pilots near Norilsk and Yakutsk have demonstrated PUE under 1.15 thanks to nine months of free-air cooling [17]. We use η = 1.15 as a portable central case.

Delivered electricity price, p_e. This is where geography splits the market. We do not use posted residential or industrial tariffs because they bear no relation to the true delivered cost faced by a marginal new data-center load. We construct an **effective all-in delivered price** that includes wholesale energy, capacity payment, transmission, and grid-firming services. Our 2026 calibration:

Region (representative interconnect)	pe, all-in effective (US$/MWh)	Notes
United States, PJM zone (new DC load)	110–150	Wholesale ≈ $50 + capacity ≈ $60 at cleared auction + transmission/firming.
Germany	100–130	Industrial tariff post-nuclear-exit, gas-pegged.
United Kingdom	130–160	Highest in Europe; gas import dependence.
Japan (Kansai/Kanto)	110–140	Post-Fukushima mix with partial restarts.
South Korea (Yongin cluster)	80–110	KEPCO industrial; new 3 GW transmission committed.
France	55–80	Nuclear-heavy wholesale.
Nordic (NO/SE)	30–60	Hydro + Swedish nuclear; free-cooling boost.
China (Eastern coast, AI-zones)	40–55	Coal + nuclear baseload; state-set industrial tariffs.
Gulf (UAE/Saudi, AI campuses)	30–50	Gas + nascent nuclear (Barakah).
Russia/Siberia (hydro-fed clusters)	20–35	Bratsk-Irkutsk hydro; ruble-denominated.
Canada (Quebec hydro)	35–55	Hydro abundance, BTC-occupied.
Brazil (Norte/Nordeste hydro)	45–70	Hydro 65% but seasonal.
Australia (NSW pipeline)	90–130	Coal legacy + permitting bottleneck.
(Sources for each row are listed in §References and discussed regionally in §4. The wholesale and capacity components for PJM are derived directly from cleared auction prices and ISO bulletins)[4,5,8].		

Amortization, C_a. This is the noisiest term and, today, the largest. An H100 SXM acquired at roughly US$ 28–32 thousand, with a five-year useful life and 70 percent annual utilization, generates an amortization charge near US$ 1.0 per H100-hour, before financing, networking, cooling capex, building, and software. In realistic latency-bounded inference deployments — not batch-saturated benchmarks — a single H100 sustains roughly 200–500 output tokens per second, or 7×10^8 to 2×10^9 tokens per hour. The per-token amortization works out to between US$ 5×10^{-10} and US$ 1.4×10^{-9}, or roughly **US$ 0.5–1.4 per million tokens** at full real-world utilization. This is, today, roughly an order of magnitude larger than the energy term. We will return to this in §4 and §7. The crucial property of Ca is that it is **not** structurally geographic: an H100 costs roughly the same to acquire in Texas and Hong Kong. Chinese providers running domestic accelerators (Ascend 910C) acquire silicon at a lower dollar cost but at lower memory bandwidth and a worse software stack, which is approximately a wash. The geographic divergence of the marginal cost of a token, therefore, runs almost entirely through the **energy term** — and only the energy term has the property that it grows in relative weight as utilization climbs and capex matures.

2.3. The Energy Floor

Putting the central case together at 0.5 J/output token and $\eta = 1.15$, one million output tokens requires
$1\,000\,000 \times 0.5 \times 1.15 = 575\,000$ J = 0.16 kWh.

Multiplying by region-specific p_e the energy-only floor on the marginal cost of producing one million output tokens is:

Region	Energy floor, US$/M output tokens
US PJM zone (new DC load)	0.018 – 0.024
Germany	0.016 – 0.021
UK	0.021 – 0.026
Japan	0.018 – 0.022
South Korea	0.013 – 0.018
France	0.009 – 0.013
Nordic	0.005 – 0.010
China (AI-zones)	0.006 – 0.009
Gulf	0.005 – 0.008
Russia/Siberia	0.003 – 0.006
Canada (Quebec)	0.006 – 0.009

These are the **lowest** possible numbers, ignoring amortization (US$ 0.5–1.4 per million tokens today), R&D allocation, networking, gateway, cooling capex, and margin. The total realistic marginal cost of one million output tokens for a US PJM-zone operator is therefore on the order of US$ 0.7–1.5, of which the energy share is roughly 2–4 percent. For a Chinese AI-zone operator, the total is roughly US$ 0.5–1.0, of which the energy share is roughly 1–2 percent.

Why, then, does the energy term matter?

It matters for two reasons. First, it is the only term in the identity that is **structurally geographic**. Amortization globalizes. Energy doesn't. Second, the energy term's share rises monotonically as utilization climbs and the capex stack is amortized down. A 2026 inference call is dominated by amortization. A 2029 inference call on a 2026 hardware stack is much less so. By the time the consensus terminal-margin models clear (2027–2029 is the assumed maturity window in most sell-side decks), the energy term is the dominant remaining cost, and its 3:1 to 5:1 cross-tier spread is the dominant remaining source of regional margin divergence.

For comparison: DeepSeek-V3.2 is currently priced at US$ 0.14 per million input tokens and US$ 0.28 per million output tokens — barely above an estimated marginal cost of US$ 0.5–1.0 per million tokens, consistent with a market-share strategy at near-zero margin. Claude Sonnet 4 sits at US$ 3.00 / US$ 15.00, a healthy gross-margin band over its own marginal cost. GPT-5 frontier is at US$ 10 / US$ 30. The headline 20–100× gap between Chinese and US frontier API pricing is partly distillation efficiency, partly market-share land-grab, partly brand premium, and partly Chinese willingness to sell at marginal cost. The energy floor is the only component of that gap that does not compress over time.

3. The Four-Tier Framework

We classify the world's AI-relevant geographies on two axes: **delivered electricity cost and capacity reliability** (the ability of the grid to dispatch new datacenter load without rationing existing demand).

3.1. Tier 1: Sovereign Capable

Cheap electricity, reliable dispatch, state alignment.

3.1.1.China. Sixty-two operating reactors as of April 2026 delivering 61.215 GW of net installed nuclear capacity, second globally only to the United States. Thirty-nine units under construction adding another 37.3 GW, ranked first for the 19th consecutive year. The 15th Five-Year Plan (March 2026) targets 110 GW of nuclear by 2030 [18,19]. Daily token consumption hit 140 trillion by March 2026, a thousand-fold rise on three years prior. DeepSeek, ByteDance's Doubao, and Zhipu AI all rationed API access during the March 2026 compute crunch; DeepSeek suffered a near-twelve-hour outage [20,21].

3.1.2. The Gulf (UAE, Saudi Arabia). Stargate UAE, a 5 GW AI campus partnered between G42, Oracle, NVIDIA, Cisco, and SoftBank, with an initial 200 MW phase due online before end of 2026 [22]. HUMAIN (Saudi Arabia's PIF-backed AI vehicle) plans 1.9 GW of AI-specific compute by 2030, scaling to 6.6 GW by 2034 [23]. The Barakah nuclear facility (4 × 1.4 GW) is already running. Microsoft and G42 are adding 200 MW of capacity by end-2026, a 50%+ increase in one year [24].

3.1.3. Nordic (Norway, Sweden, Denmark). Microsoft's US$ 6.2 billion Norwegian data-center commitment, plus expansions in Sweden and Denmark. NVIDIA's programme manager observes that "nearly every big AI company is in Norway." Free-cooling brings PUE below 1.10 on a sustained basis [25.26].

3.1.4. France. Nuclear delivers roughly 70 percent of generation; wholesale prices structurally below the rest of Western Europe.

The Tier 1 economics: p_e in the US$ 30–80 / MWh band, reliable dispatch, and — crucially — **state policy that treats AI inference as a strategic export industry**. China's industrial-grid tariff is set by policy. The Gulf's gas pricing is internal. Sweden's nuclear permits are political. None of these governments will allow data-center load to be rationed in favor of residential demand the way the PJM operator is now publicly debating.

3.2. Tier 2: Emerging Stretched

Big sovereign-AI commitments on grids that cannot yet absorb them.

3.2.1. India. Reliance has pledged Rs 10 trillion (US$ 110 billion) over seven years for AI build-out, beginning with a Jamnagar campus whose first 120 MW comes online in the second half of 2026, supported by up to 10 GW of solar surplus in Gujarat and Andhra Pradesh [27,28]. Adani Group committed US$ 100 billion through 2035 and operates a 2 GW AdaniConnex base targeting 5 GW. Tata is building roughly 100 MW with OpenAI, scaling to 1 GW. India's current installed data-center capacity is approximately 1.5 GW and is projected to reach 8–10 GW by 2030 [29,30]. The constraint is not capex commitment, which is ample; it is grid stability and last-mile transmission, which is not yet reliable enough to support continuous gigawatt-scale inference loads outside the dedicated enclaves Reliance is constructing.

3.2.2. South Korea. A US$ 65 billion sovereign-AI commitment through 2027; Gyeonggi Province and KEPCO have agreed to build new transmission to deliver an additional 3 GW into the Yongin semiconductor cluster. Naver's GAK Sejong is the largest single-company data center in Asia (294,000 m², up to 600,000 rack units, six-phase expansion through 2026). Samsung SDS is expected to win the National AI Computing Center bid [31-33]. But Korea is a chronic net energy importer with a 1.5–2× cost disadvantage versus China.

3.2.3. Japan. SoftBank is the largest company in Japan by market capitalization (US$ 308 billion), having passed Toyota; Masayoshi Son has stated publicly that he is "all in" on AI. The Portsmouth nuclear consortium (Hitachi, Mitsubishi Electric, Toshiba, TDK, SMBC) is funding cleanup of a US nuclear site for an SB Energy-backed data-center deployment. Domestic Japanese industrial electricity remains expensive (US$ 110–140 / MWh effective) [34].Tier2's pattern is consistent: aggressive sovereign-AI announcements, plausible 2028+ inference capacity, but a **2026–2028 transition risk** that is not in any current sell-side model.

3.3. Tier 3: Power Constrained Wealthy

Capex without dispatch capacity.

3.3.1. United States (PJM zone). US$ 725 billion of 2026 hyperscaler capex, the deepest software talent pool on earth, and a grid operator that has cleared its capacity auction at or near the cap in three consecutive years, missed its reliability target for the first time in the 2027/2028 auction, and is publicly debating an emergency backstop procurement. Peak demand is projected to grow by 32 GW from 2024 to 2030, with 30 GW from data centers. New gas-turbine lead times are 24+ months; transmission lines are 5–10 years of permitting [3-5,7-10]. Amazon's free cash flow collapsed from US$ 148.5 billion of operating cash flow into a US$ 1.2 billion FCF as US$ 59.3 billion of incremental infrastructure spend hit the books in a single year.

3.3.2. Germany. Industrial electricity is 2× US prices. The nuclear exit is complete; gas is imported. Data-center growth is fleeing toward the Nordic interconnect.

3.3.3. United Kingdom. Highest electricity prices in Europe, retarded SMR pipeline, structural marginalization of the British AI buildout.

The Tier 3 paradox: the regions with the most capex, the deepest capital markets, and the strongest software ecosystems have the highest structural floor on inference cost. The sell-side narrative — "hyperscalers will spend whatever it takes" — is true at the cash-flow level and irrelevant at the marginal-cost level. You cannot capex your way out of a power-market clearing price.

3.4. Tier 4: Arbitrage Suppliers

The cheapest electrons on earth, exported.

3.4.1. Russia (Siberia). Wholesale power in Moscow oblast rose 12–16 percent in 2025, but Siberian hydro and Rosatom-adjacent zones still clear at US$ 20–35 / MWh. Cold-climate datacenters in Norilsk and Yakutsk achieve PUE under 1.15 nine months per year. Sber's chairperson has publicly stated a need for 200–400 MW of dedicated DC capacity. Yandex Cloud committed RUB 12 billion (US$ 131 million) to lift GPU capacity by 50 percent in 2026 and launched a sovereign AI suite. Mandatory data-residency under Federal Law 242-FZ accelerated repatriation from foreign hyperscalers; SberCloud, Yandex Cloud, and VK Cloud each report enterprise customer growth above 35 percent year over year [35,36]. What is most consequential, however, is the cross-border export: open-source Chinese cloud workloads are increasingly routed to Siberian inference clusters where electricity costs are below the Chinese floor, and the resulting tokens are sold globally. There is no clean public number for this flow yet; the LinkedIn note from Sandeep De and a handful of trade press dispatches mark its earliest visibility [37].

3.4.2. Canada (Quebec hydro). Abundant, but BTC-occupied and politically constrained on incremental allocation.

3.4.3. Brazil. Hydro at 65 percent of mix but seasonal, and political instability adds a discount rate.

3.4.4. Australia (NSW pipeline). 11.4 GW in development, but legacy coal phase-out and a broken permitting regime gate the rate of arrival [38].

Tier 4 is the **silent subsidy layer** to Tier 1. The cheapest tokens that reach the global market in 2026 are, in part, the product of Siberian electricity sold to Chinese workloads.

4. Cross-Tier Spread Quantification

We assemble the central case from §2 with the regional bands from §3 to produce a structural-floor estimate for the marginal cost of one million output tokens. Energy term only:

Tier	Region	Energy floor (US$/M tok)	Indicative API output price 2026	Implied margin band
1	China (Eastern)	0.006 – 0.009	0.28 (DeepSeek V3.2)	≈ 95–98 %
1	Gulf	0.005 – 0.008	n/a (no public commercial endpoint at scale)	—
1	Nordic	0.005 – 0.010	0.28 (DeepSeek V3.2)	—
1	France	0.009 – 0.013	n/a (Mistral, EU-internal)	—
3	US PJM	0.018 – 0.024	15.00 (Claude Sonnet 4); 30.00 (GPT-5)	≈ 99.9 %
3	Germany	0.016 – 0.021	n/a	—
4	Russia/Siberia	0.003 – 0.006	sold into Tier 1	—

Two observations matter.

First, the absolute floor is so low in every tier that current commercial pricing is essentially unconstrained at the energy term. Today's spread between DeepSeek and OpenAI is a strategic pricing decision, not a structural one. A price war can compress US prices toward US$ 0.20 per million input tokens without breaking marginal economics, and a Chinese leader could in principle move further down.

Second, this changes when utilization climbs. The energy term is a per-token *floor;* the per-token amortization term is a function *of utilization*. As model deployment scales and capex amortizes, the realized cost of an inference call approaches the energy floor asymptotically. At low utilization, electricity is a small fraction of total cost. At high utilization — which is where the industry is heading — electricity becomes the dominant term. The crossover region is exactly the 2027–2029 window in which most current sell-side models are pricing terminal-state margins.

This means the geographic spread does not show up *today.* It shows up at the **maturity** point of the build-out, which is precisely when the buy-side terminal values are computed. A 3:1 structural energy ratio at maturity rounds-trips into a margin spread of multiple hundred basis points on returns on invested capital. That number does not exist in current consensus models because the consensus models assume the energy floor is a uniform global constant.

5. Shanghai's AI Token Futures as an Electric-Spread Derivative

In late May 2026, the Shanghai Futures Exchange disclosed that it had begun preliminary design of AI-token futures contracts — taking, as the underlying asset, the per-token output price of a defined LLM service [39-42]. The contract is still preliminary; no Chinese regulatory approval has been secured; the working window for launch is three to five years per Baocheng Futures' market note. Two parallel US efforts at CME and ICE are designing GPU compute-capacity futures with different underlyings.

The trade-press framing has been domestic: SHFE is building a hedging instrument to let Chinese LLM operators lock in inference costs as the domestic price collapses. That framing is correct and incomplete.

The deeper structure is this. A futures contract on the price of an AI token, with delivery in a specific Chinese cloud, is by construction a contract on the marginal cost of producing a token in that cloud. The marginal cost of producing a token in that cloud is dominated by the local electricity floor. The local electricity floor is dominated by China's nuclear-and-coal baseload mix. **Therefore the AI-token futures price will track Chinese electricity, mediated by model-efficiency improvements.**

But the same contract, traded against a US-dollar reference and cleared internationally, is also a tradable instrument on the **cross-tier electricity spread**. A non-Chinese counterparty who shorts the SHFE AI-token contract is implicitly long the Chinese energy floor and short the global one. A non-Chinese inference buyer who is long the contract — i.e., locks in supply at the Chinese price — has implicitly purchased a hedge against US power-market dislocations. SHFE has accidentally, or deliberately, created the first derivative of the eight-way cross-tier electricity spread of the global AI inference market.

This is not a domestic hedging tool. It is a geographic risk premium swap. The fact that it is denominated in tokens, not megawatts, does not change the underlying.

Three structural consequences follow:

- **a) Pricing convergence**. As the SHFE contract gains liquidity (assume 2028–2029 in the published Baocheng window), the implied volatility of the AI-token forward curve will price the electricity spread between China and the rest of the world. Chinese inference providers gain a hedging tool. Their non-Chinese customers gain a hedging tool. The US hyperscalers' shareholders, who have no equivalent instrument, do not.
- **b) Cross-border arbitrage formalization.** Today's grey-market routing of US, EU, and Japanese workloads to Chinese (and increasingly Siberian) inference is unhedged. A liquid SHFE contract gives it a financial backbone. This is the same evolution that took crude oil from a spot-cargo trade to a structured commodity in the 1980s.
- **c) A new comparable for utility-scale energy infrastructure.** A SHFE token contract that tracks the Chinese energy floor is implicitly an equity-like claim on the marginal-cost advantage of Tier 1 grids. This creates a tradable proxy for the regional electricity-cost spread that long-cycle utility investors do not currently have.
- **d) The contract will not launch tomorrow.** But it is being designed, today, by an exchange that already runs the world's deepest copper and steel-rebar contracts. SHFE understands what it is doing.

6. Five testable predictions, Q3 2026 – Q4 2027

We commit to five predictions, each falsifiable within the next eighteen months, against which this framework can be checked.

- a) **PJM 2028/2029 capacity auction (cleared mid-2026) will again clear at or above the FERC-approved cap,** and the cleared megawattage will again undershoot the installed-reserve-margin target. PJM's backstop mechanism will be triggered before year-end 2026.
- b) **At least one US hyperscaler will publicly disclose a multi-gigawatt commitment to a non-US sovereign-capable jurisdiction** (the Gulf, the Nordic countries, or France) for new inference capacity by Q2 2027. The Microsoft–G42 200 MW announcement of mid-2026 is the first instance; the next round will be measured in gigawatts.
- c) **The headline gap between Chinese and US frontier API pricing will compress, not from the US side downward, but from the Chinese side upward,** as Chinese providers begin to monetize a market that has matured. Within twelve months of SHFE's contract reaching even an indicative liquidity, DeepSeek-class endpoints will reprice their headline output token by 30–60 percent upward, and the upward reprice will track the implied curve of the SHFE contract.
- d) **At least one earnings cycle in 2027 will surface, on a major hyperscaler's earnings call, an analyst question that pairs PJM clearing prices with terminal-margin model assumptions**. The CFO answer will be evasive. The stock will move on it.
- e) **A first publicly disclosed cross-border power agreement between a Russian Siberian operator and a Chinese hyperscaler cloud will appear, denominated in token-output units rather than megawatt-hours, during the second half of 2027. The arbitrage layer will move from grey to formal.**

These predictions cost us nothing if we are wrong: we simply note where we miscalibrated. They cost the sell-side something if we are right, because the consensus terminal margins for the US AI-infrastructure stack are built on assumptions about the energy term that the data does not support.

7. Why Models that Ignore the Floor are Mispriced

Take a representative sell-side AI-infrastructure model with consensus inputs for 2027–2030: hyperscaler capex compounded at 25 percent CAGR through 2028, AI revenue growing at 60 percent CAGR, gross margin on inference services expanding from 35 to 55 percent as utilization climbs, terminal ROIC of 22 percent. The model probably has a US$ 2.5 to US$ 4 trillion cumulative AI infrastructure capex bucket and a terminal multiple in the high teens. Now ask the model what it assumed for the delivered cost of electricity to that capex stack. In our reading of the major desks' public reports through Q2 2026, the answer is either nothing — electricity is implicitly held flat at 2024 levels — or, more sophisticated, the analyst applies a uniform 3–5 percent annual cost inflator that does not differentiate between the PJM zone and the Nordic interconnect. In neither case is the structural electricity floor framed as a binding constraint on margin expansion.

Apply our four-tier framework. Hold capex flat. Hold revenue assumptions flat. Replace the uniform-electricity assumption with a tier-weighted average:

- 60 percent of US new inference capacity in PJM zone (effective \$130/MWh by 2028 if the auction stays at the cap and capacity charges roll into delivered rates),
- 20 percent in other US RTOs (effective \$80–100/MWh),
- 20 percent in Tier 1 jurisdictions via cross-border arrangements (effective \$50/MWh).

The shifted terminal-margin range, holding everything else constant, falls to 35–42 percent rather than 55 percent. The terminal ROIC drops from 22 percent to 15–17 percent. On a 17× multiple, this is a 30–50 percent miss on terminal value for the US-anchored leg of the trade.

We are not arguing that the entire AI infrastructure trade is mispriced by this amount. We are arguing that the *US-anchored portion* of the trade is. And we are arguing that there exists, today, a counter-trade — long Tier 1 sovereign-capable proxies, long the SHFE AI-token forward when it lists, short the most PJM-exposed names — that the market has not yet been forced to put on. It will be forced to put it on by the same chain of disclosures that triggered the dispersion-trade unwind of February 2026 and the silver flash collapse of October 2025: a single high-profile earnings miss that names the energy term, followed by a 90-day repricing cycle.

We have been here before. The market is slow to incorporate a constraint that has been quantifiable in plain sight for two years. By the time it stops being controversial, the trade is gone.

8. Limitations and Caveats

A few honest disclaimers.

The energy-per-token term used in this paper is a working central case. Frontier closed-weight models — Anthropic's, OpenAI's, Google's largest — are larger than the public benchmark data on which the 0.3–1.5 J/token range is calibrated, and the operators do not publish marginal energy figures. Our band is therefore likely conservative on the high end. A factor-of-two error here moves the absolute floor but does not change the cross-tier ratio. The delivered electricity prices used in §3 reflect the marginal effective price faced by new large-load entrants, not posted average industrial tariffs. In some jurisdictions (China most notably) the gap between posted and marginal is policy-managed. Our calibration is sourced from public auction data where available (PJM in particular) and from published trade-press estimates elsewhere. We invite correction. The SHFE contract is, as of this writing, preliminary. The argument that it will function as a cross-tier electricity-spread derivative depends on it actually listing with sufficient liquidity in a 2028–2029 window and being cleared internationally rather than purely on a domestic Chinese basis. If it lists as a purely onshore Chinese contract with no foreign access, the cross-border arbitrage argument is weakened, though not eliminated — the spread can be replicated synthetically with cross-listed equities and power-curve forwards.

Tier 4 (the arbitrage suppliers) is the weakest-data leg of the framework. The Siberia-to-China power flow into AI workloads is real and growing, but the public quantification is poor. We have flagged where this matters; we would benefit from primary-source data here. Finally, the framework treats inference as the dominant share of AI infrastructure spend over the next four years. This is consistent with all major sell-side and IEA forecasts but not certain — a substantial shift back toward training-dominated workloads (driven by, say, a foundation-model unlock in synthetic data or test-time-compute scaling) would change the energy-per-token-to-energy-per-training-run ratio and complicate the picture. We treat this as the principal scenario risk.

9. Conclusion

The market is pricing AI infrastructure as a capex-and-talent trade. It is in fact a capex-and-electricity trade, and the electricity term is not priced globally. The geographic distribution of cheap electrons no longer follows the distribution of capex commitments. The United States has the capex without the electrons; the Gulf, China, the Nordic countries, and France have the electrons; India and South Korea have the ambition without yet the grid; Russia is selling the cheapest electrons on earth to whoever can route the workload. The Shanghai Futures Exchange's AI-token futures, when they list, will be the first instrument that gives this asymmetry a price. The headline volatility of that contract will track Chinese electricity. Its term structure will price the cross-tier electricity spread. The hyperscalers' shareholders, whose terminal models assume a flat global electricity floor, do not yet have a hedging instrument for the structural cost gap that the framework documents.

The sell-side consensus on AI infrastructure has been correctly diagnosed as a capex bubble by no one and as a power-market constraint by very few. It is both. The bubble is in dollars and in chips. The constraint is in megawatts and in hours of grid-firming and in the political willingness to allow industrial load to take precedence over residential. We have written this paper to argue that the constraint is binding and that there exists, beginning now, a tradable expression of it.

Acknowledgements

The author has spent seventeen years on derivative trading desks, the last five focused on cross-asset macro signals into volatility. Conversations with practitioners on the buy-side and in utility

regulation informed the framework but bear no responsibility for its claims. The paper builds on a thesis first articulated in *The China AI Disruption Thesis* (CrossVol Research, May 2026; DOI 10.5281/zenodo.20509815), of which this is the geographic and energy-floor extension.

References

1. Tom Tunguz, "The $112 Billion Quarter: Hyperscalers Bet the Farm on AI," April 29, 2026.
2. Tom's Hardware, "Google, Microsoft, Meta, and Amazon capex spending to hit $725 billion in 2026," 2026.
3. UncoverAlpha, "Amazon, Google, Microsoft, Meta Q1 earnings: AI profits are here," 2026.
4. S&P Global Commodity Insights, "PJM power capacity auction clears at record high price of $269.92/MW-day for most of footprint," July 30, 2024.
5. PJM Inside Lines, "PJM Auction Procures 134,311 MW of Generation Resources; Supply Responds to Price Signal," July 22, 2025.
6. Enel North America, "PJM 2026/2027 Capacity Auction Results,"2025.
7. Utility Dive, "PJM accelerates backstop reliability auction amid uncertainty over data center cost allocation," 2025.
8. PJM Inside Lines, "PJM Auction Procures 134,479 MW of Generation Resources," December 2025.
9. Utility Dive, "Data centers were 40% of PJM capacity costs in last auction: market monitor," 2025.
10. IEEFA, "Projected data center growth spurs PJM capacity prices by factor of 10," 2025.
11. Joule (Cell Press), "Energy use of AI inference, efficiency pathways, and test-time scaling," 2026.
12. arXiv, "TokenPowerBench: Benchmarking the Power Consumption of LLM Inference," December 2025.
13. DeepSeek-AI, "DeepSeek-V3 Technical Report," arXiv 2412.19437, December 2024.
14. Google Data Centers, "Power Usage Effectiveness."
15. Statista, "Data center average annual P U E worldwide 2025."
16. Uptime Institute, "Global Data Center Survey Report 2024."
17. Mordor Intelligence, "Russia Data Center Market Outlook to 2031."
18. World Nuclear Association, "Nuclear Power in China."
19. Carbon Credits, "2026: The Year Nuclear Power Reclaims Relevance."
20. Xinhua, "Chinese AI models harness token economy, expand global footprint," April 3, 2026.
21. CNN Business, "China's AI upstart DeepSeek drops new model," April 24, 2026.
22. Data Center Dynamics, "US and UAE plan to build 5GW AI data center campus, run by G42 and American hyperscalers."
23. Introl, "The Middle East's trillion-dollar bet on AI infrastructure."
24. Middle East Institute Switzerland, "The AI-Energy Nexus in the Gulf Region."
25. CNBC, "Why Europe's electricity prices threaten its AI ambitions," May 18, 2026.
26. Euronews, "Europe is hungry for AI data centres — but its energy grid cannot feed them," May 5, 2026.
27. TechCrunch, "Reliance unveils $110B AI investment plan as India ramps up tech ambitions," February 19, 2026.
28. DD News, "Reliance and Jio to invest Rs 10 lakh crore over seven years for India's AI transformation."
29. Light Reading, "Reliance and Adani commit $210B to make India an AI-infra hub." infra-hub
30. IBTimes India, "India's AI awakening: Ambani, Tata, and Adani launch $310B+ sovereign intelligence revolution."
31. Introl, "South Korea $65B AI: Samsung & SK Hynix Lead."
32. Data Center Dynamics, "Samsung SDS set to win bid to build South Korea's National AI Computing Center."
33. Korea Herald, on South Korea growth fund AI sector designation, 2026.
34. The Register, "SoftBank builds AI mega-datacenter on nuke site," March 23, 2026.
35. Data Center Dynamics, "Yandex to launch availability zone next year at data center in Vladimir, Russia."
36. Computer Weekly, "Putting Russia's datacentre market under the microscope."
37. Sandeep De, public LinkedIn note on Siberian electricity and Chinese cloud workloads, 2025.
38. Climate Council Australia, "Seizing the opportunity to do data centres right."
39. AI Weekly Alerts, "Shanghai Futures Exchange designs AI token futures."
40. TechCrunch, "Just like gold and oil, we'll soon be able to trade AI token futures," May 28, 2026.
41. CryptoBriefing, "Shanghai Futures Exchange studies AI token futures to hedge rising AI costs."
42. Invezz, "China explores AI token futures market amid growing compute demand," May 28, 2026.
43. The Standard (Hong Kong), "China works on AI token futures market, sources say."
44. IEA, "Energy and AI: Energy supply for AI."
45. IEA,"Executive Summary — Electricity 2026."
46. Brookings, "Global energy demands within the AI regulatory landscape."
47. Yahoo Finance, "Hyperscalers Hit $700 Billion in 2026 AI Spending Plans."
48. AI Automation Global, "AI Inference Cost Crisis 2026."
49. Silicon Data, "Understanding LLM Cost Per Token: A 2026 Practical Guide."
50. CostGoat, "LLM API Pricing Comparison & Cost Guide, June 2026."
51. Djellal Djouad, *The China AI Disruption Thesis: Why The Sell-Side Is Six Months Late On AI Infrastructure.* CrossVol Research, May 2026.