The Energy Shock Behind Artificial Intelligence

The Economy Research Editorial[1,2]

[1]The Economy Research, 71 Lower Baggot Street, Dublin 2, Co. Dublin, D02 P593, Ireland

[2]Swiss Institute of Artificial Intelligence, Chaltenbodenstrasse 26, 8834 Schindellegi, Schwyz, Switzerland

Abstract

This article reconceptualizes the current AI boom not just as a software-driven transition, but as an infrastructure shock. Though much public discussion of AI takes for granted a discourse centered around algorithms, model capabilities, and productivity growth, it is electricity consumption, data-center geography, the production of cutting-edge semiconductors, and competition among nations for computing power which have become the actual binding constraints. The argument is that the current price-tag of AI reflects not simply an unusual spike on the market but three synergistic and interacting pressures: a strain on power grids, scarcity in hardware components, and geopolitical rivalry among major states. Although falling prices to users for inference may begin to make AI look more economically feasible at the interface, its widespread adoption still depends on costly and hard-to-acquire physical components for many companies, regions, and governments. The data center itself is becoming politically sensitive infrastructure due to its high demand for land, power and water as well as its meager job creation beyond the construction phase. Meanwhile, constraints on availability of GPUs, high-bandwidth memory and advanced packaging are rendering semiconductor capacity a critical strategic issue for states. This analysis concludes that the scope of AI policy must be widened beyond the regulation of models alone to include the physical realities of compute as infrastructure, with a commensurate concern for power provisioning, conditional siting approvals, supply-chain resilience, and pragmatic state action.

1 Introduction - AI's Physical Bottleneck: Energy, Hardware and Geopolitics

Artificial Intelligence (AI) is frequently called the next revolution, on par with the internet and mobile technology. While advances in deep learning, natural language processing and big computing continue, the revolution will not be in the algorithms, but in the hardware, energy, data centers and geopolitics that support AI. Every past major technological revolution was not merely driven by the invention of a new idea or piece of software, but the ability of humanity to access resources and infrastructure to make that idea broadly available (e.g., the steam engine revolution driven by access to coal and iron; the electrification of society reliant upon copper wiring and reliable power grids). Now, policy and market debates often speak abstractly of algorithms, models and data for AI, but actual AI applications will not advance any further without their physical constraints being addressed. This shift from software to hardware-driven narrative is a critical inflection point to understanding where AI goes from here and now, from code to kilowatt-hours and from algorithms to raw minerals. The realm of possibility is rapidly being defined by material capacity and geopolitical constraint.

Standard narrative about AI has focused on software; discussions around its growth now center around the quality of models, demand, or regulation and it is time for this narrative to be revised. With the deployment of conversational AI across wide uses, we are facing more physical and geopolitical rather than algorithmic bottlenecks in the shape of energy provision, advanced chips and grid capacity, cooling infrastructure, choice of data center locations and ownership over supply chains. In 2024, data centers used 415 TWh of electricity, accounting for 1.5% global use and this figure could touch 945 TWh by 2030.[1] Investment could more than double to $500 billion from 2022-24 (IEA Jan 2025) and the average AI-intensive data center consumes as much as 100,000 homes and this capacity may increase by 20x.[2] Power constraints exist everywhere, from hyperscale data centers in Northern Virginia, Ireland, Singapore, to cities elsewhere where their density and power requirements lead to moratoriums on new centers, land use fights and strict regulations. Besides electricity, data centers require enormous quantities of water for cooling; a typical large facility can consume millions of gallons a day and it becomes an issue in drought-ridden regions where air-cooled systems and closed-loop mechanisms are preferred. The land use impacts range from data center concentrations and clusters fundamentally altering landscapes to increased property values. Innovations for improved energy efficiency, such as immersion cooling and heat capture, are used in a race to minimize impact, but advances cannot catch up with the exponential growth in computational demands.

The transition taking place in the United States highlights this trend. US data centers are projected to consume 176 TWh (4.4% of US electricity use) in 2023 and this is expected to grow to between 325 TWh (6.7%) and 580 TWh (12.0%) by 2028,[3] depending on GPU demand, the type of AI workload and energy efficiency improvements in cooling infrastructure, a study by Lawrence Berkeley National Laboratory says. Electricity is now a constraint not only for IT systems, but also for utilities, as its consumption is growing beyond what grids are equipped to handle; industrial policy is becoming an increasingly relevant aspect of energy provision for the digital sector.[4]

This energy challenge is amplified by the demand for hardware. Progress in hardware has been very fast,

but demand grows faster. AI hardware performance has grown at a compounded annual rate of 43% since 2008 and at 40% for energy efficiency, while the cost per unit has dropped by 30% (Stanford 2025 AI Index). Despite this, it costs anywhere from $79 million (GPT-4), $192 million (Gemini 1.0 Ultra), to $170 million (Llama 3.1-405B) to train each model.[5] Industry leaders warn that the frontier training runs will transition from hundreds of millions of dollars into the billions and then to even larger numbers, following trends in model size, data demands and infrastructure needs. While efficiency has decreased the cost to deploy, it has also driven increased usage and thus shortages. AI hardware infrastructure, including GPUs, TPUs and high-bandwidth memory, relies on very complex, fragmented and globalized supply chains. Producing semiconductors, especially at advanced nodes, is an incredibly multi-step process that requires highly specific machinery and rare earth metals like neodymium, dysprosium and terbium. China is not dominant in leading-edge chip manufacturing, but it remains central to the broader strategic contest through its scale, industrial policy, critical minerals position, mature-node capacity and accelerated push for semiconductor self-reliance. The hoarding of both chips and minerals is already considered to be in the nation's best interests, but fabricating advanced chips requires extremely high capital investment and a high level of technology and expertise, hence why the CHIPS and Science Act of the US and similar bills in Europe and Asia, are incentivizing domestic production. Vertical integration through acquiring firms or investing in manufacturing capabilities directly is becoming the favored strategy in the industry to achieve guaranteed access. Collaborative partnerships between industries and with governments are formed to optimize research, share costs and expedite the development of new hardware configurations.

Thirdly, AI is evolving from being an economic driver to a geostrategic struggle. Forty AI models were released in the US in 2024, compared to fifteen in China and while there is a close gap in performance, models in China are outstripping the US in terms of patent and publication numbers. China's data centers constitute the world's second-largest market, with an estimated 32 new AI projects started between 2023 and 2024, with substantial government backing. With both energy and chips becoming highly prized resources, costs may now only continue to rise and motivate additional public and private investment, rather than diminish demand. Restrictions on the export of advanced chips and associated technologies have been placed on China in an effort to prevent its development of advanced military capabilities, but also to impede its commercial technology advancement and alter the structure of supply chains. Such sanctions are often reciprocated with additional national investments in indigenous technologies and the creation of new regional collaborations. Digital sovereignty is a core policy priority; nations increasingly desire for data, computing power and the value it generates to remain within their borders, thus new national clouds, mandatory local data storage and development of new standards are proliferating. With the challenges in defining, implementing and policing global privacy, data ownership and intellectual property rules, the effectiveness of multilateral dialogues and standard-setting organizations is at stake. Infrastructure previously thought of as 'neutral' now serves as a strategic tool and a source of bargaining power.

The central premise of this paper is that the current high costs of AI are not merely inconvenient but are driven by energy, hardware and geopolitical forces. Instead of viewing AI as an 'add-on' for productivity, it should be understood as a complex and resource-dependent system that is constrained by human beings'

ability to generate power, place data centers and procure hardware at scale, thereby overcoming material and geopolitical limits that have previously been overcome by widespread deployment. This echoes historical paradigm shifts in electrification and global telecom, in which demands collided with the limits of what humanity could materialistically construct. The scale and consequences of the present shift now place even higher stakes in political and industrial landscapes, far beyond economics alone.

2 Temporary high cost of AI

What is striking about AI now is not that it will eventually become very cheap; the concern with that has already peaked, but that we are currently in the phase where there are rapidly falling marginal costs for inference and extremely high costs for infrastructure at the scale that many users require. The Stanford AI Index tracked the cost of inference per one million tokens for models around GPT-3.5's capability; over the period November 2022 to October 2024, this fell 280-fold from $20 to 0.07 per million tokens.[6] Meanwhile, top output models remain in the range $2.19 to 60 per one million tokens as of early 2025. This presents a paradoxical experience for users where experimenting with AI may seem cheap, while fully reliable, large-scale deployment of an AI system can be extremely expensive due to fixed capital outlays for infrastructure; the costs associated with the generation of the prompt go well beyond API prices.

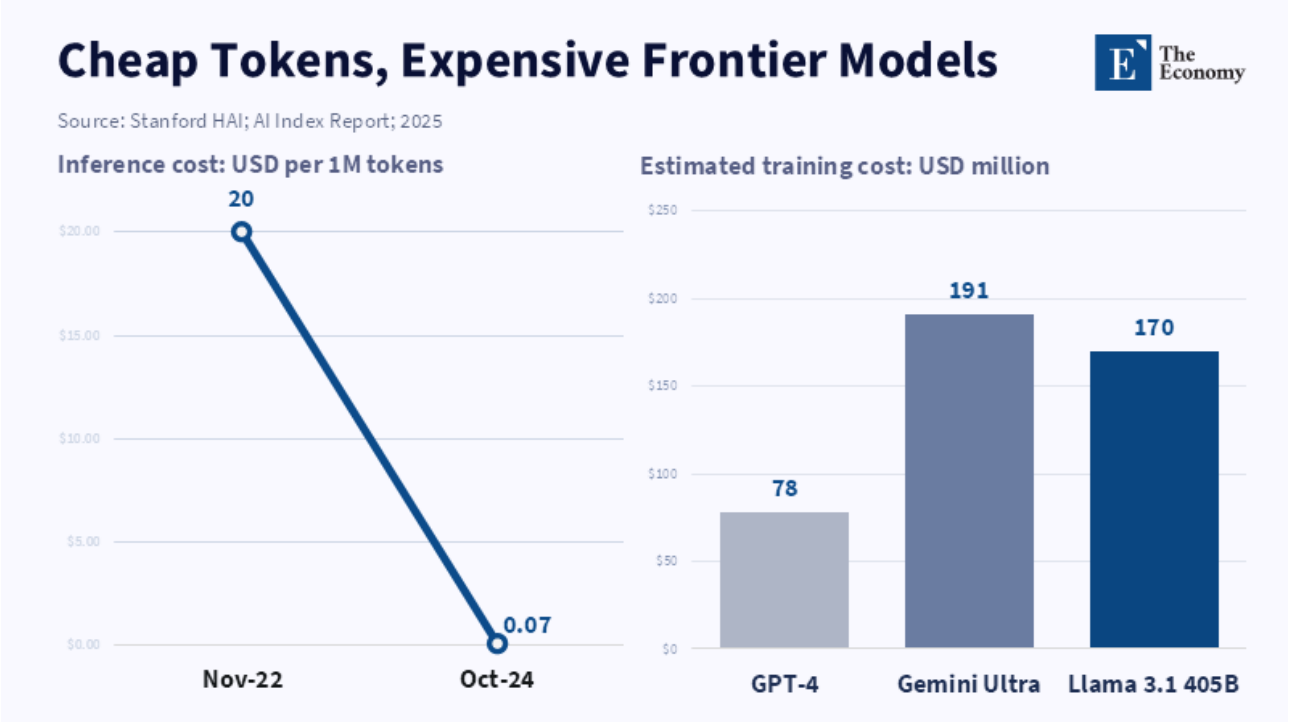

Figure 1: AI adoption is already stratified by firm size, suggesting that infrastructure cost and organizational capacity may shape access as much as software availability.

Capital expenses for AI infrastructure have now begun creating enormous gaps between companies and across regions of the world.

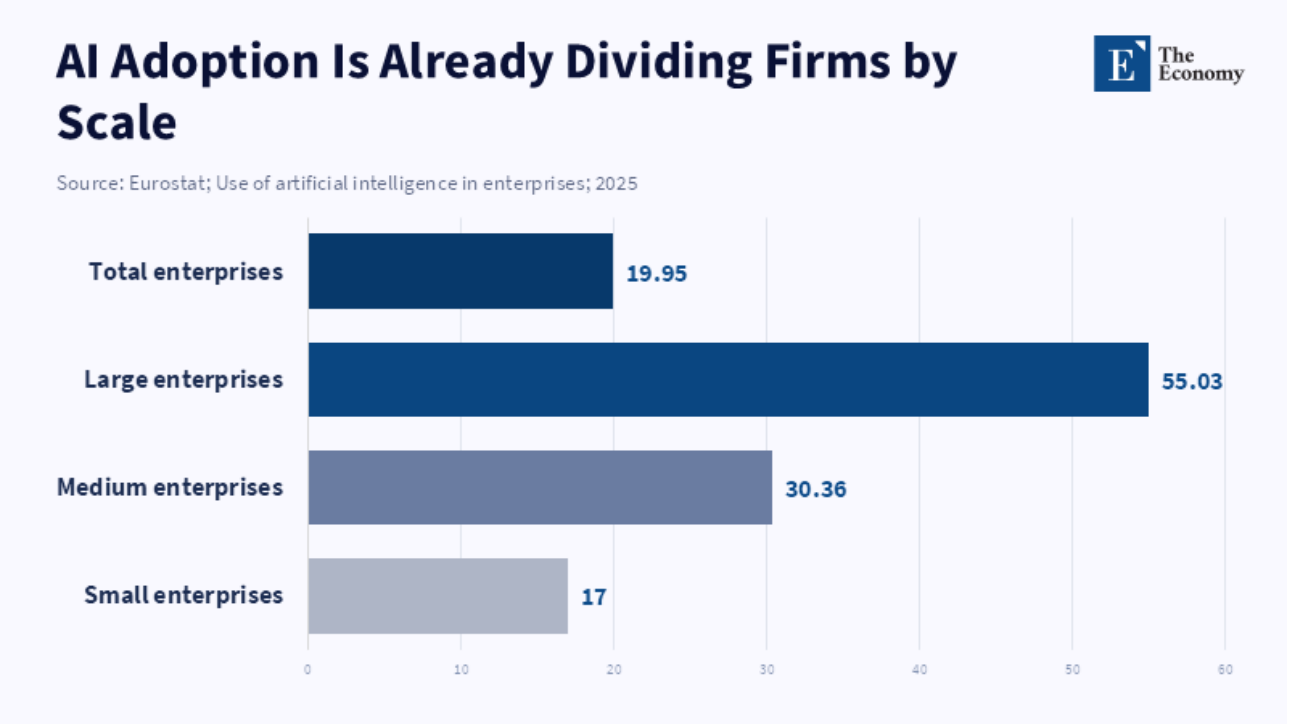

Figure 2: Falling inference prices make AI appear cheaper at the point of use, but frontier model development remains tied to very high fixed infrastructure and training costs.

The companies that currently hold dominant positions are ones with access to a huge quantity of cheap capital and existing cloud infrastructure, where they are able to absorb costs of top-end chips and redundant power, while it remains too costly for small startups to build their own infrastructure. The risks are that incumbents may become entrenched at the expense of faster progress on the margins, while companies look towards massive mergers and takeovers as the market tries to gain economy of scale in hardware and energy supply chains and insurers/financial institutions try to develop models of risk for AI infrastructure due to reliance on single suppliers, potential shifts in regulation and the robustness of power and logistics systems. The large capital expenses of infrastructure will likely be recovered through decreased unit costs over time, especially with the advent of secondary markets for reconditioned chips, but this will be a period of extreme volatility. Not only will adoption rates vary significantly between industries, but also across geographic regions; the finance sector, healthcare and logistics are moving much faster than other industries, which may have long compliance hurdles, older systems, or other resource constraints. It may create or increase the digital divide unless proactive policies and financing structures are deployed to support adoption.

Why this is significant is that the total cost of ownership (including capital expenditure and total cost of changing an existing system), the risks and complexity of changing a system, legal risks and the reliability of output are all significant factors in the decisions a firm makes about whether to invest in AI. Forbes Research found that less than one percent of over 1000 C-Suite executives reported significant (20%+ return on investment) financial gains on the AI they invested in and about a quarter of all firms are expected to delay their planned AI spending for 2026.[7] Reuters, in a report on an EY survey, also finds that nearly all big firms that use AI reported interim losses of $4.4B due to the high cost outlays.[8]

This helps explain the uneven spread. As of 2023, OECD found an average of 8% of all companies and 28%

of all ICT firms were adopting AI.[9] The research also found that firms with over 250 employees were two times as likely to have adopted AI compared to smaller firms and when the research is conducted with SMEs the top barriers include compute cost, privacy concerns and lack of AI skills; on this point,Reuters reports that the German Mittelstand, in a report commissioned by EY, expects their planned 2025 investment into AI at 0.35% of revenues (down from 0.41% in 2024) due to costs, despite worldwide corporate investment rising.[10] While the cost of API services may be falling, these low marginal costs cannot compensate for the very high fixed costs and it may result in great divergence. Some might argue that the above criticism is not valid, that, as a general-purpose technology, AI is like calculators or search engines in that its proliferation should and will be extremely wide. The calculator/search engine analogy is flawed, though. While calculators and search engines reduced information and cognitive load, respectively, neither required companies to make wholesale changes to electricity grids, manufacturing systems, or security infrastructure and they didn't generate the same kind of material demands. The comparison is important because it highlights a critical difference: while calculators and search engines were passive tools that reduced demands upon existing infrastructure, AI actively consumes vast amounts of electricity and it is this material reality of AI, rather than its seeming costlessness via the UI, that will drive its overall deployment among organizations.

This suggests the high cost will persist for the foreseeable future and is structural, not cyclical. It's short-term in that a more efficient chip design or cooling system could reduce the costs and prices are volatile because demand for electricity and cooling now outstrips availability. A crisis anywhere can cause disruption elsewhere, which impacts costs. For smaller economies, this will increase the digital divide. In response, grid development and modernization, conditional permits to ensure efficiency and renewable integration, public-private partnerships to promote shared resource development, subsidizing grid-friendly development rather than simply use and fostering international collaboration on technology standards will likely become increasingly common; regulatory sandboxes are likely to expand, while public procurement might become a key tool in steering development towards beneficial outcomes for all citizens.

3 How to resolve the pressure - Data Center

The obvious immediate bottleneck is the data center itself. The problem itself is hardly trivial; while the lower the latency, the closer it is to end-users, city fiber networks and key exchange points, the closer to the hubs you get the higher the price of land and the more difficult politics. It is not even just the price of the land that can be an issue: a manufacturing plant provides far more ongoing jobs than a data center, even if construction generates substantial benefits. The auditor of the Virginia state legislature reported 74,000 jobs, $5.5 billion labor income and 9.1 billion GSP annually as contributions from the largest cluster of US data centers in the state, but qualified this by noting the majority of value generation occurs during construction, not operation.[11] This is an important distinction, that the tangible costs of land use, new substations, high-power transmission lines, water scarcity, noise and increasingly the question of the cost of power all fall on the locality, while the benefits are disseminated around the broader digital economy.

Siting thus becomes a legitimate problem, not just a real estate one. This is no more obvious than anywhere

in Europe and especially Ireland, where data centers already consume 21 percent of Ireland's total metered electricity consumption (up from 5 percent in 2015), having overtaken household electricity consumption in cities, according to Ireland's Central Statistics Office.[12] The regulator has thus moved from implicit approval to explicit conditionality; under its connection policy for 2025, all new large data centers must provide on-site generation or storage and is to be separately connected and metered, while the demand-side load is to match the grid imports.[13] This is a significant paradigm shift in how the growth of data centers is treated; they will no longer be a standard part of business but rather special cases of high-load infrastructure with part of their costs shifted into the economics of the project.

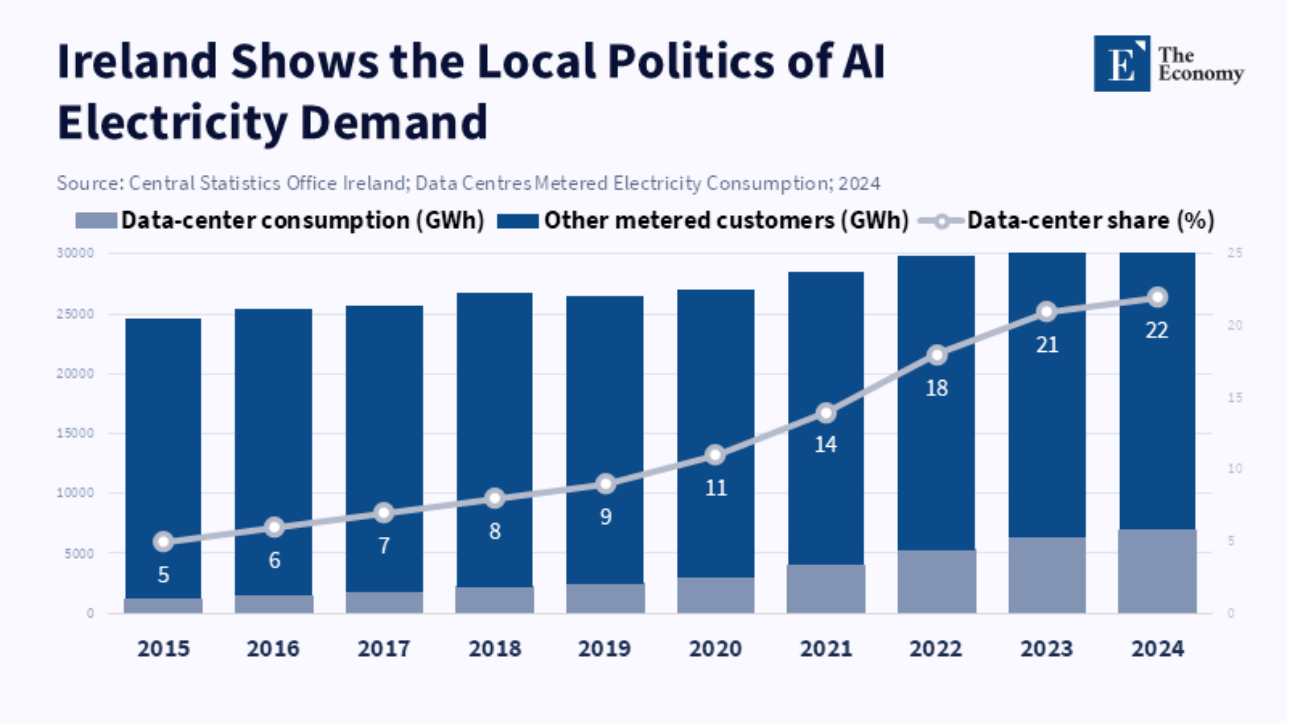

Figure 3: Ireland shows how data centers can move from marginal industrial users to a politically visible electricity category competing with households and other local demand.

This rationale is gaining traction elsewhere, too; it is estimated that by the Brookings Institute, up to 20 percent of planned data centers around the world will be significantly delayed because of issues regarding the grid and that average connection times in mature hubs take seven to ten years, or longer, with up to thirteen in certain cases.[14] Reuters has noted that in April 2026, OpenAI has stalled a main data center project in the UK due to the price of energy and disadvantageous policy environments and is noting community concerns around facilities run by Microsoft, Amazon and Google.[15] These are not merely abstract problems, but practical ones that impact where and at what speed data centers can be powered on and how likely they are to be tolerated politically.

Grid institutions are adapting; the North American Electric Reliability Corporation has estimated that by 2028 data centers could be using 12 percent of total US electricity demand, up from 4.4 percent in 2023, noting the potential for new large loads to cause reliability problems as they may not behave predictably in the same way as standard loads.

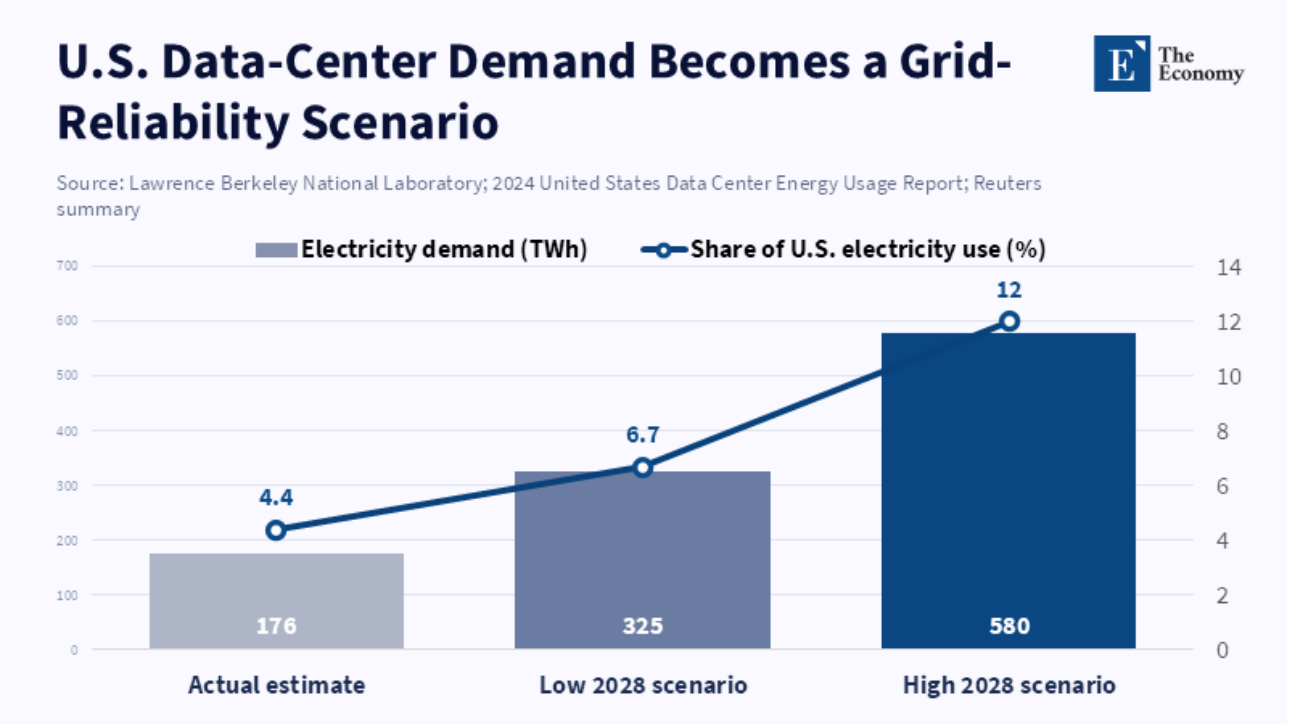

Figure 4: The U.S. data-center outlook is best understood as a reliability range: depending on AI demand and efficiency, compute could become a major grid-planning category by 2028.

NERC has elevated large-load reliability concerns, while FERC has begun examining how AI data centers and co-located large loads should be regulated to protect reliability and consumers[16, 17] It is clear here that problems cannot be solved by private contracts; when the system is overloaded, regulators must participate in the economics of AI.

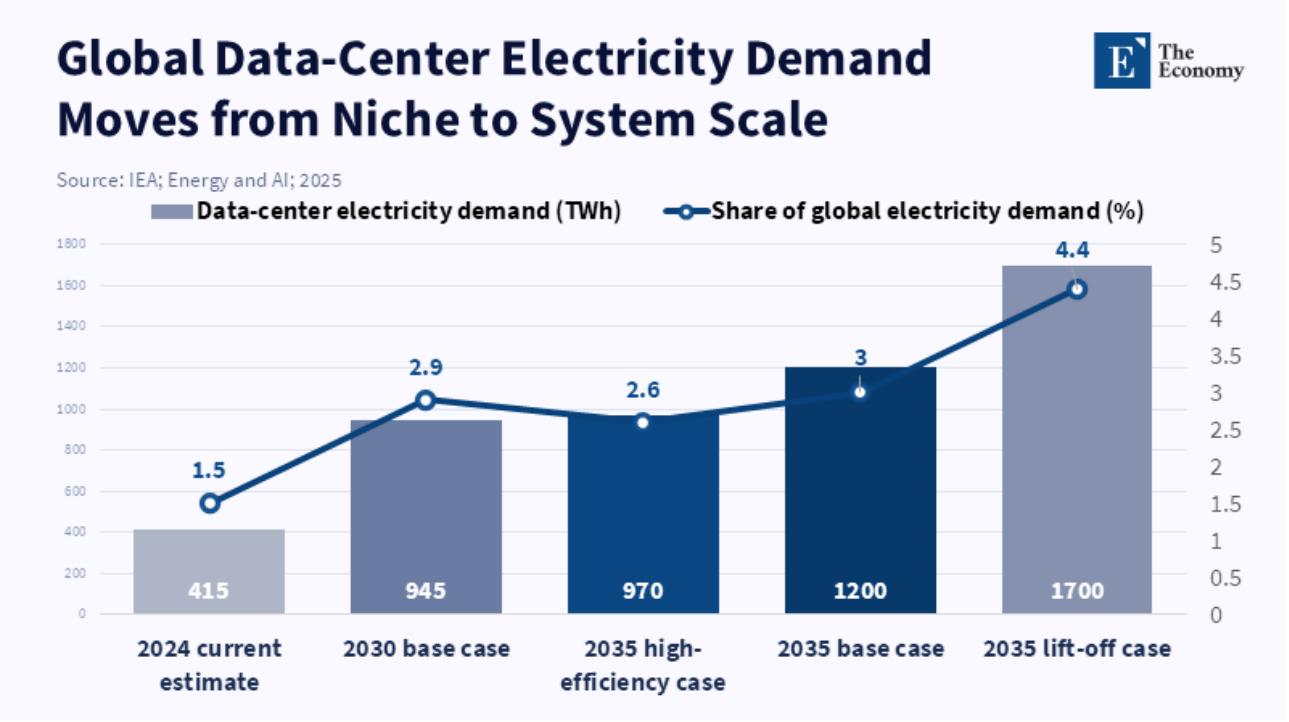

Figure 5: The policy challenge is not simply that data centers use electricity, but that their projected growth is becoming large enough to reshape power-system planning.

There is not likely to be a singular answer to all of these problems, but a series of approaches. The IEA

notes that in the years to 2030 the proportion of electricity generated by renewable sources such as wind and solar globally is to rise from 17% in 2025 to 27% and that by 2035, data centers will account for up to 50GW of flexible generation via the manipulation of workloads and load itself.[18] Operators in places such as Scandinavia and Canada leverage abundant hydropower and cooling to cut down their costs and footprints, while Asia looks at special economic zones and public-private partnerships. The practical answers are ones of operations and placement; less pressing training and batch jobs will occur during periods of low power prices, the cost of power will be internalized via colocation and designs will allow for liquid cooling and load variability rather than assumptions of continuous uniform usage. It is clear that, even if some bottlenecks may be unsolvable in the short term, large improvements may be made via innovation and policy, not necessarily technological change.

This is already manifesting itself in the market; Microsoft has signed a twenty-year deal for the recommissioning of the Three Mile Island Nuclear Generating Station unit, which, upon being brought online will provide 835MW, while Amazon Web Services have extended a nuclear deal with the same provider, ensuring supply of up to 1,920MW until 2042 and are considering small modular reactors, while Google have a deal in place for advanced geothermal power in Nevada which is planned to increase their geothermal capacity from 3.5MW to 115MW.[19] These are not ancillary procurement contracts, but indicate to the market that ensuring a reliable power supply is a crucial part of a top-tier provider's AI strategy, not just a peripheral sustainability concern.

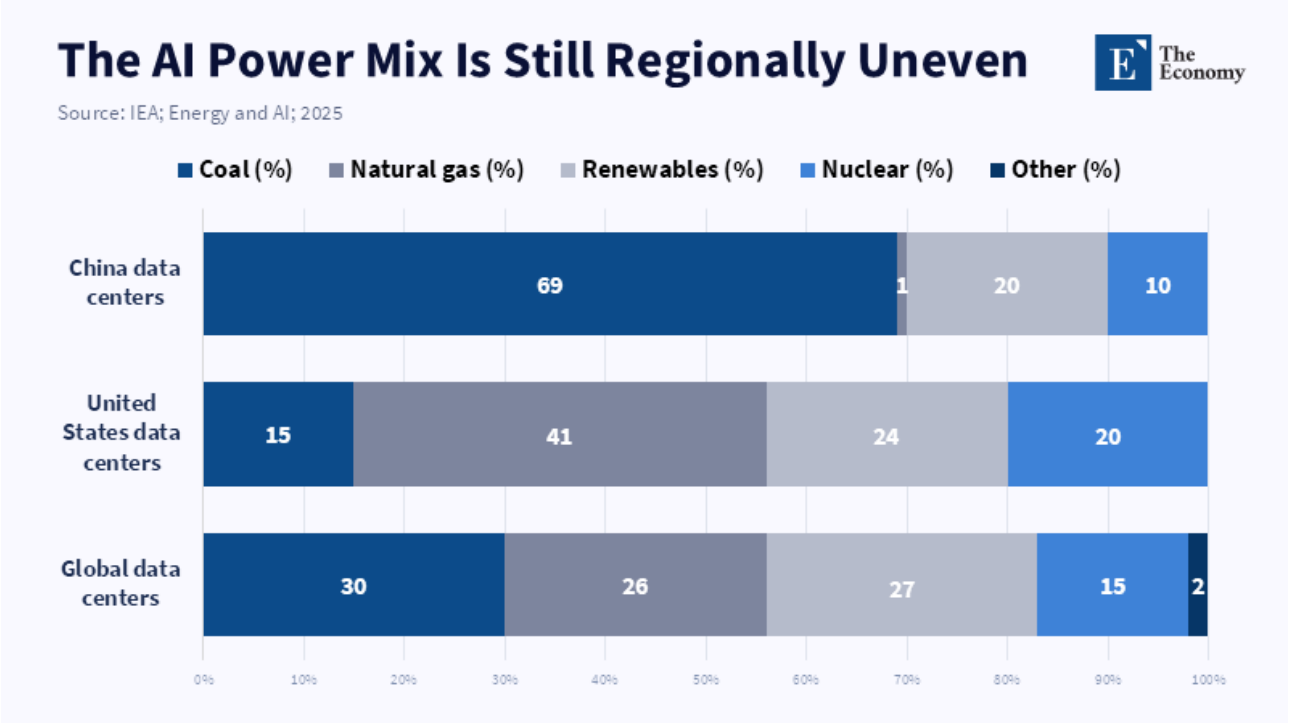

Figure 6: The environmental and political cost of AI compute depends heavily on location, because the power mix behind data centers remains sharply uneven across regions.

There will be other proposed ideas even bolder; Microsoft's Project Natick has proved it is possible to build data centers underwater, where their proposed underwater data center had servers eight times more reliable than land based alternatives[20] (a pilot which Microsoft abandoned), while floating data centers are further down the road commercially with a 6.5MW liquid-cooled site already operational with no towers and an extremely low PUE in Stockton run by Nautilus Data Technologies. Orbital data centers have been proposed for the potential

to leverage a strong renewable component, or perhaps no environmental factors, while being placed far from land constraints or water issues,[21] although it remains far from clear that these can ever become commercially viable, with space companies themselves acknowledging issues related to radiation, maintenance, launch costs and reliability and not for outlandish ideas, but rather because such proposals may never solve the short-term constraints relating to energy and space.

4 How to resolve the pressure - Hardware supply

Even without the power constraints, the AI economy would have bumped into a second wall – supply. While it has been diagnosed primarily as a GPU constraint, this is an incomplete explanation. There were, in fact, a range of supply bottlenecks spanning components – HBM, high-end DRAM, packaging capacity, networking, power, cooling – and a wider range of parts of the software ecosystem as well. By May 2024, SK had sold out of its entire 2024 supply of HBM and had nearly fully allocated its 2025 supply; Micron also sold out of its entire 2024 HBM supply and had largely allocated its 2025 supply by May, later announcing in December 2024 that its entire 2025 supply was sold out and prices already locked.[22] This announcement occurred alongside Reuters reporting, based on TrendForce estimates, that the price of conventional DRAM was up 13%-18% Q/Q in Q2 2024.[23] This is the behavior of a supply constraint that is becoming a real crunch – multiyear forward commitments, not merely high prices, demonstrate that sellers have pricing power.

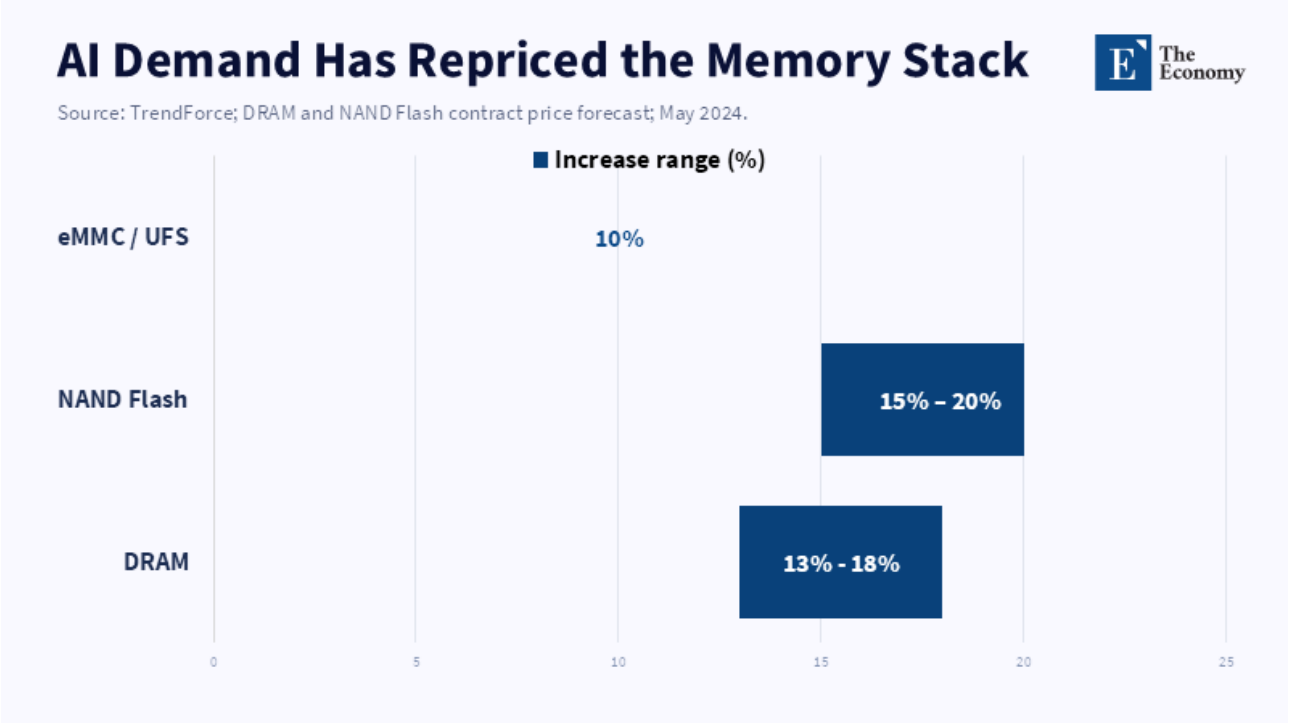

Figure 7: The hardware bottleneck is not limited to GPUs; AI demand has spilled into memory and storage markets, raising the cost base of the wider compute stack.

The difficulty of the situation is compounded by the atypical hardware demands of AI. Although frontier AI accelerators do not require massive wafer throughput, they do demand stacked memory and advanced packaging. As reported by Reuters in March 2024, TSMC intended to build CoWoS advanced packaging capacity in Japan, with current capacity in Taiwan already strained by AI demand.[24] By May 2024, the level of demand pressure

escalated with Reuters reporting major tech companies already directly investing in SK Hynix production lines and funding purchases of EUV lithography machines in cash, guaranteeing memory supply.[25] These steps are described as "unprecedented" in the memory industry, suggesting a shift from a normal cyclical upswing to strategic scarcity, when customers have to subsidize component suppliers.

In light of this, significant research is now being conducted into chiplet architectures, interconnects and novel materials. Companies such as AMD, Intel and NVIDIA are heavily investing in these modular architectures, which provide for scalability and flexibility, though these approaches take some years to materialize. New vendors in India, Korea, Israel and Europe are developing some supply-chain capacity, driven by security and technological autonomy concerns. However, the challenge of high capital cost, skilled labor and the difficulty of integrating into current global value chains will be formidable. The interplay between incumbent players and newcomers as this hardware ecosystem develops will be a significant factor determining future supply-chain stability.

The macroeconomic impacts have already been large. According to the Stanford AI index, while the A100 served as the foundational accelerator for the top AI models, H100 became widely deployed, powering 15 top AI models by the end of 2024.[26] The premium paid for the latest generation is demonstrated by the same report, with the performance-per-dollar and energy efficiency sharply increasing per generation, with simultaneous demand spikes and supply shortages fueled by the shift to high-density cluster configurations. Similarly, a Lawrence Berkeley report highlighted that GPU-accelerated AI server power consumption in U.S. data centers rose from below 2 TWh per year in 2017 to more than 40 TWh per year in 2023.[27] Unlike many prior consumer electronic cycles, better chips now lead to higher cluster sizes, which then lead to increased demand on the memory, packaging and power ecosystem.

National governments have also reacted rapidly to this hardware bottleneck. In the U.S. $6.6 billion in direct federal funding has been made available under the CHIPS Act to TSMC Arizona in support of more than $65 billion in three advanced fab facilities.[28] Japan has also been even more aggressive with significant funding to bolster its domestic semiconductor production capacity at Rapidus and its partners. A further 631.5 billion yen was committed in the 2026 budget, on top of 200 billion yen allocated for FY2025.[29] Moreover, the pursuit of Japanese packaging capacity indicates that Tokyo hopes to establish a fully integrated manufacturing ecosystem. This represents an escalation beyond simple subsidies to national priority investment.

China's policy response also bears notice. By mid-2024, over $6.1 billion has already been invested by China in the Eastern Data, Western Computing initiative.[30] These initiatives operate 8 computing hubs with 1.95 million servers, running at 63% utilization according to authorities. Over 150 computing capacity-related projects initiated in 2023-2024 were completed and operational by the end of 2024 according to Brookings.[31] This signals not only investment in computation capacity but potentially future demand for domestic components due to higher import costs and restrictions. These supply and demand incentives will thus reinforce each other.

Therefore, predictions of price moderation between 2027 and 2028 should be approached with caution. Though increased supply will definitely have an impact, prices will unlikely fall back to where they were before. This is because the current demands are not driven by consumer electronics/enterprise refreshment cycles, but

by the capex of hyperscalers, state-directed computing objectives, import hedged strategies and a race for new model development. Every increase in computational power now leads to a higher financial threshold for larger cluster sizes, fueling even greater demand on the memory, packaging and power stack. McKinsey forecasts that by 2030 global data center capex will total $6.7tn, with $5.2tn spent specifically on AI-ready capacity alone. [32] Supply growth will thus simultaneously relieve bottlenecks while stoking further computational ambition. Expect localized pricing adjustments, not a return to the old normal.

5 How to resolve the pressure - Geopolitical pressure

The third source of stress is geopolitics. Just as energy makes AI materially scarce and hardware makes AI industrially scarce, geopolitics makes AI strategically indispensable. "U.S. Bureau of Industry and Security says it's limiting semiconductor exports to block China from accessing and building the state-of-the-art chips that it relies upon for military applications."[33] Even beyond the delays on shipped chips, this control, introduced in late 2022 and subsequently updated in 2023 and 2024, affects the incentives of all companies and countries along the supply chain and so China is now shifting toward domestic alternatives.[34] As the CSIS predicted in 2026, "The trend toward the engineering out of U.S. Hardware where possible, coupled with continued overall industry growth, will keep overall short-term purchases high, but this does not indicate de-escalation. Rather, it is a contest for long-term independence".

This is crucial because the nature of AI has transformed from ordinary civil productivity into strategic infrastructure with dual use: high-end chips and mass computing underlie the cloud, industrial automation, scientific research and civil and military applications. Therefore, a CSIS AI Index for 2025 can focus less on tracking the models and more on tracking the pace of strategic urgency. Although the U.S. still leads in some important models, China's quality gap over the US is narrowing quickly in a lot of benchmarks, so it's very unlikely there will be a large-scale investment withdrawal from AI; since it will be politically difficult to cut costs in that juncture, as other countries appear to be moving rapidly toward the same capabilities.[35]

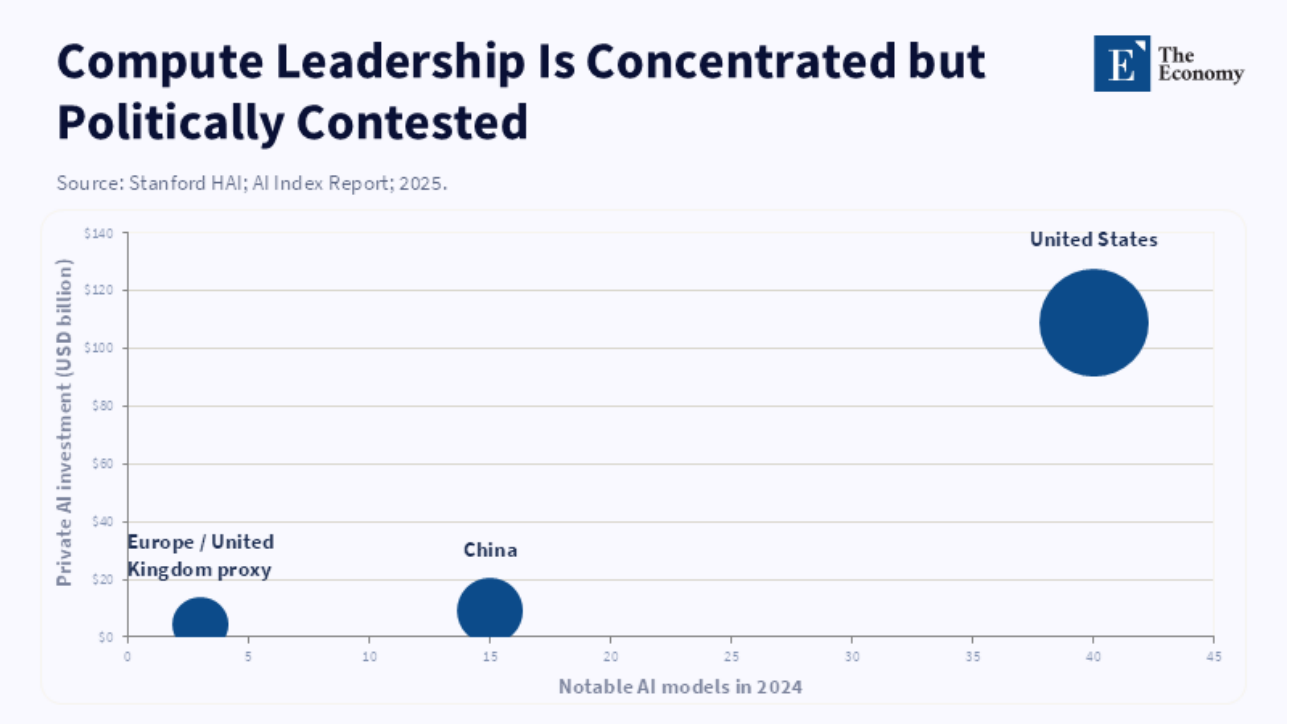

Figure 8: AI leadership remains highly concentrated, but China's scale keeps compute investment politically strategic rather than merely commercial.

The race for high-end AI products among countries is not just a Sino-US contest; European nations are now trying to build ethically sound, trustworthy, sustainable, regulation-harmonized AI, while countries like Japan, Singapore and South Korea are establishing regional alliances to share resources, research and negotiate favorable hardware import deals. Each country (including those not mentioned here) is aiming at a balance of strategic autonomy and cooperation and there is a drive toward digital sovereignty. Nevertheless, different standards, import restrictions and screening regulations have made the global response very dynamic and both competitive and collaborative.

This duality also underlies the growing cold war over infrastructure; By 2026, China is expected to account for a substantial share of global data center energy demand, with some estimates placing it near one-quarter; its vast AI and data infrastructure are growing rapidly, while the U.S. FERC is devising tailored regulatory responses for AI data centers and NERC is now defining high compute loads under a distinct reliability category.[36] In both countries, computation is transforming from privately owned digital infrastructure into strategic national infrastructure. Data centers are increasingly perceived as sensitive and expensive ports, shipyards and oil refineries whose expansion cannot be justified by pure business cases.

This geopolitical stress adds yet another explanation of why prices won't come down as supply increases; more chips will only spur more training runs, more energy will encourage more infrastructure build-outs and more domestic manufacturing will only increase industrial policy competition. The three stresses amplify one another; Energy scarcity further enhances the value of bundled power and infrastructure, increasing countries' appetite for large contracts on both; hardware scarcity heightens demand for subsidized manufacturing and sourcing and geopolitics makes sovereignty indispensable for every choice. As a result, AI infrastructure prices will never go back to median rates as memory prices have with the advancement of technology cycles. It's a

business that has become a strategy and thus the demand curve is shifting outward.

The outcome, therefore, must not be the market-driven AI buildout fiction. The heavy loads of AI should be treated as strategic infrastructure that demands higher levels of transparency about energy, water and emergency power and needs pricing schemes that depend on specific connection costs. Approval of such infrastructure should be contingent on measures including demand flexibility, heat-mitigation plans and ideally, localized/proximal power generation; subsidies should be linked to packaging, memory and grid readiness rather than just chip production and international cooperation should take into account that restricting chip imports, while ignoring energy supply, standards and infrastructure, only shifts the bottleneck and domestic regulations should avoid using consumer prices to subsidize private AI investments through indirect state infrastructure projects.[37]

6 Conclusion - AI Infrastructure Is the Real Bottleneck

The fundamental point from the AI boom is that at its core, the shock to AI is a shock to electricity and infrastructure before it has an impact on broad productivity. This is the initial corrective aspect of this analysis. The surge in electricity demand, scarcity of memory and advanced packaging supply and intensified geopolitical competition in computing were mostly overlooked. Current indicators suggest that none of these limitations can be considered minor; on the contrary, they are the core issues. The demand for electricity by US data centers is projected to reach 6.7 to 12 percent of the national total by 2028, while global demand for data center electricity is predicted to at least double by 2030. The costs of training frontier models are still extremely high, while inference does not impose comparable demands on resources. As countries worldwide are stimulating the construction of chip factories, data centers and purchase agreements through subsidies, instead of ceasing efforts, the conclusion is clear unequivocally: chip supply, coupled with how these three factors interact, will determine AI hardware prices and availability. Sound AI policy must be built upon the understanding that AI constitutes physical infrastructure requiring realistic energy pricing, power-grid interconnectivity reform, conditional permitting and a sufficient supply of dependable and renewable energy sources, while maintaining a rational connection between memories, packages and the power grid. The current misconception of AI as only sophisticated software output rather than as resource-intensive and politically charged infrastructure will only lead to confusion.

References

[1, 2] International Energy Agency (2025) *Energy and AI.* Paris: IEA.

[3] Shehabi, A., Smith, S.J., Hubbard, A. et al. (2024) *2024 United States Data Center Energy Usage Report.* Berkeley: Lawrence Berkeley National Laboratory. doi:10.71468/P1WC7Q.

[4] Shehabi, A., Smith, S.J., Hubbard, A. et al. (2024) *2024 United States Data Center Energy Usage Report.* Berkeley: Lawrence Berkeley National Laboratory.

[5] Maslej, N. et al. (2025) *Artificial Intelligence Index Report 2025.* Stanford: Stanford Institute for Human-

Centered Artificial Intelligence.

[6] Maslej, N. et al. (2025) *Artificial Intelligence Index Report 2025*. Stanford: Stanford Institute for Human-Centered Artificial Intelligence.

[7] Reuters (2025) 'Business leaders agree AI is the future; they just wish it worked right now', *Reuters*, 16 December.

[8] Reuters (2025) 'Most companies suffer some risk-related financial loss deploying AI, EY survey says', *Reuters*, 8 October.

[9] OECD (2024) *Fostering an Inclusive Digital Transformation as AI Spreads among Firms*. Paris: OECD Publishing.

[10] Reuters (2026) 'Germany's Mittelstand cuts AI investments, study shows', *Reuters*, 8 January.

[11] Joint Legislative Audit and Review Commission (2024) *Data Centers in Virginia*. Richmond: Commonwealth of Virginia.

[12] Central Statistics Office Ireland (2024) *Data Centres Metered Electricity Consumption 2023*. Dublin: CSO.

[13] Commission for Regulation of Utilities (2025) *Large Energy User Connection Policy Decision Paper*. Dublin: CRU.

[14] Brookings Institution (2025) *Global Energy Demands within the AI Regulatory Landscape*. Washington, DC: Brookings.

[15] Reuters (2026) 'OpenAI pauses UK data centre project over regulation, costs', *Reuters*, 9 April.

[16] North American Electric Reliability Corporation (2026) *Level 3 Alert: Reliability Guideline Focused on Large Load Challenges*. Atlanta: NERC.

[17] Federal Energy Regulatory Commission (2025) *FERC Directs Nation's Largest Grid Operator to Create New Rules to Embrace Innovation and Protect Consumers*. Washington, DC: FERC.

[18] International Energy Agency (2025) *Energy and AI*. Paris: IEA.

[19] Reuters (2024) 'Google partners with Nevada utility for geothermal power for data centers', *Reuters*, 13 June.

[20] Microsoft (2020) *Microsoft Finds Underwater Datacenters Are Reliable, Practical and Use Energy Sustainably*. Redmond: Microsoft.

[21] Reuters (2026) 'SpaceX says unproven AI space data centers may not be commercially viable', *Reuters*, 21 April.

[22] Micron Technology (2024) *Fiscal First Quarter 2025 Earnings Presentation*. Boise: Micron Technology.

[23] Reuters (2024) 'Samsung flags 15-fold rise in second-quarter profit as chip prices climb', *Reuters*, 4 July.

[24] Reuters (2024) 'TSMC considering advanced chip packaging capacity in Japan, sources say', *Reuters*, 17 March.

[25] Reuters (2026) 'SK Hynix flooded with unprecedented offers from big tech firms to secure chip supplies', *Reuters*, 7 May.

[26] Maslej, N. et al. (2025) *Artificial Intelligence Index Report 2025*. Stanford: Stanford Institute for Human-Centered Artificial Intelligence.

[27] National Institute of Standards and Technology (2025) *TSMC Arizona CHIPS for America Award*. Washington, DC: NIST.

[28] Reuters (2026) 'Japan approves additional funding for chipmaker Rapidus', *Reuters*, 11 April.

[29] Reuters (2024) 'China invests $6.1 billion in computing data center project, official says', *Reuters*, 29 August.

[30] McKinsey & Company (2025) *The Cost of Compute: A $7 Trillion Race to Scale Data Centers*. New York: McKinsey & Company.

[31] Bureau of Industry and Security (2024) *Commerce Strengthens Export Controls to Restrict China's Capability to Produce Advanced Semiconductors for Military Applications*. Washington, DC: U.S. Department of Commerce.

[32] Shivakumar, S., Wessner, C. and Howell, T. (2026) *China's Localization Drive in Semiconductors Gains Impetus from Allied Chip Export Controls*. Washington, DC: Center for Strategic and International Studies.

[33] Maslej, N. et al. (2025) *Artificial Intelligence Index Report 2025*. Stanford: Stanford Institute for Human-Centered Artificial Intelligence.

[34] Brookings Institution (2025) *Global Energy Demands within the AI Regulatory Landscape*. Washington, DC: Brookings.

[35] Federal Energy Regulatory Commission (2025) *FERC Directs Nation's Largest Grid Operator to Create New Rules to Embrace Innovation and Protect Consumers*. Washington, DC: FERC.