Allianz

Allianz Trade

Thinking fast, building slow: The energy cost of the US AI boom

12 May 2026

Allianz Research

Content

Executive Summary

Patrick Hoffmann
Economist, ESG & AI
patrick.hoffmann@allianz.com

Katharina Utermoehl
Head of Thematic and Policy Research
katharina.utermoehl@allianz.com

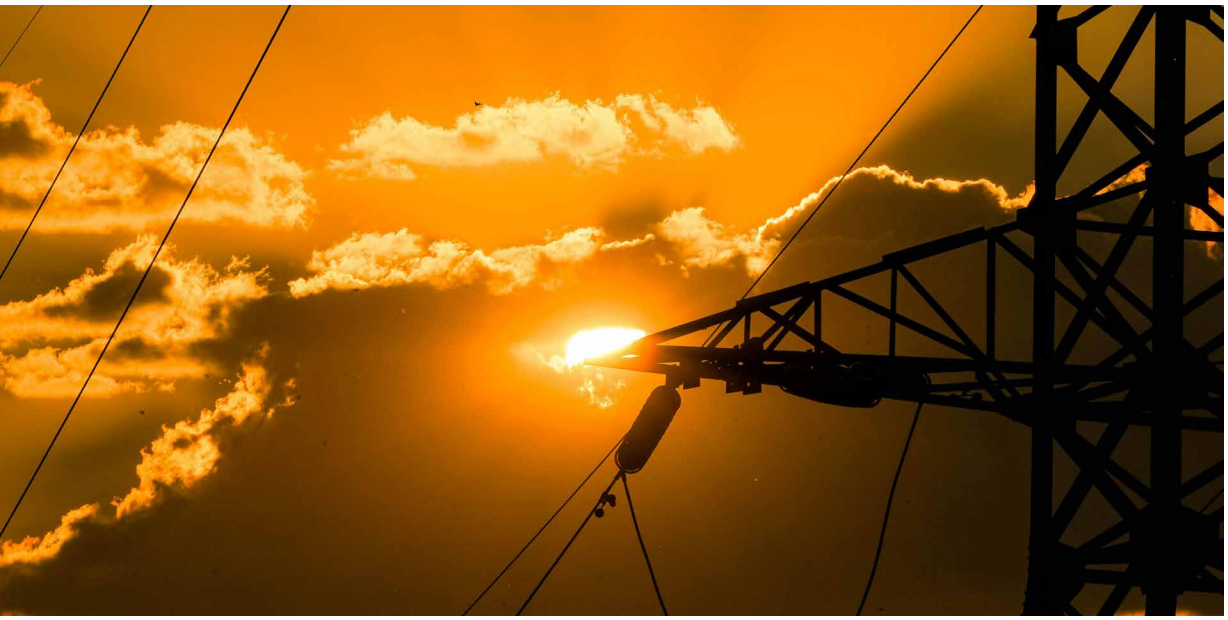
- **Artificial intelligence is about to impose the largest sustained demand shock on US electricity infrastructure in decades.** By 2030, data-center power consumption is expected to nearly double, lifting the sector's share of total US electricity demand from roughly 5% to around 9%. Although planned generation additions look sufficient on paper, data centers may absorb nearly half of projected new capacity, leaving thin margins if electric-vehicle adoption or industrial electrification accelerate faster than expected. Yet demand itself is evolving faster than forecasts can capture: generative AI reached 53% population-level adoption in just three years, agentic systems consume far more energy per interaction than conventional workloads and a 280-fold decline in inference costs since 2022 is driving a rebound effect that efficiency gains alone cannot offset.

- **The critical bottleneck, however, is not generation capacity but the grid itself.** Data centers can be built in under two years, but grid connections can take up to seven years in congested markets such as Northern Virginia. Nationwide interconnection requests now total 1.84 terawatts, exceeding total installed US generating capacity. Supply-chain shortages have pushed lead times for critical grid equipment to several years, while projected demand would require building roughly 8,000 km of high-voltage transmission lines annually – around ten times the current pace. The strain is already showing: In early 2026, the Department of Energy invoked emergency powers to shift data centers onto backup generation during peak demand periods. Rising public opposition and legislative scrutiny add a further layer of uncertainty that conventional supply forecasts have yet to fully capture. The strain is already visible in project pipelines, with half of the 12 GW of US data-center capacity planned for 2026 being delayed or cancelled.

- **Aggregate electricity prices have not yet fully reflected these pressures, but a growing „data-center premium" is emerging.** States hosting the highest concentration of data-center activity have so far seen price inflation below the national average, reflecting favorable grid conditions, economies of scale and the lagged structure of utility rate-setting. Beneath the surface, however, the impact has become increasingly visible since 2023. US residential customers are already paying USD1.4bn more per year on their electricity bills as a direct result of data-center demand, with just five utilities serving 4.4mn households in Northern Virginia, the Pacific Northwest and Arizona accounting for more than 40% of that total. The sales-weighted average price effect across all utilities sits at just 0.6%, but for the most exposed utilities roughly 7.8pps of a 24.5% cumulative price increase between 2020 and 2024 are directly attributable to data-center demand, adding 0.19pp to headline inflation over four years through the direct electricity channel alone. These markets historically benefited from below-average electricity costs, a

gap that has already narrowed from 5% to 3.7% since 2020 and is set to close further. Data-center investment grew 32% in 2025 and is set to rise a further 75% in 2026 alone, pointing to an additional electricity price effect of close to 14pps for the most exposed utilities over 2025–2026, nearly doubling the cumulative four-year effect in just two years. Meanwhile investor-owned utilities filed USD18bn in rate-increase requests in 2025, the highest since the mid-1980s, with costs largely falling on existing rate-payers rather than the facilities driving them.

- **Preparing energy infrastructure for AI and addressing community concerns is as urgent as building the AI infrastructure itself.** The immediate priority is interconnection reform: binding timelines, penalties for speculative queue filings and priority treatment for shovel-ready projects with firm power commitments would help relieve the most acute bottlenecks. Cost allocation is equally important. Unless data centers bear a proportionate share of the infrastructure costs they create, public opposition and permitting delays will intensify further. Mandatory energy-use disclosure, incentives to redirect investment away from saturated regions, stronger efficiency standards, demand-flexibility mechanisms and stricter additionality requirements for power-purchase agreements would fill the most critical gaps in a policy framework that remains largely inadequate to the scale of the challenge – and lay the foundation for AI ambitions that the grid can actually support.

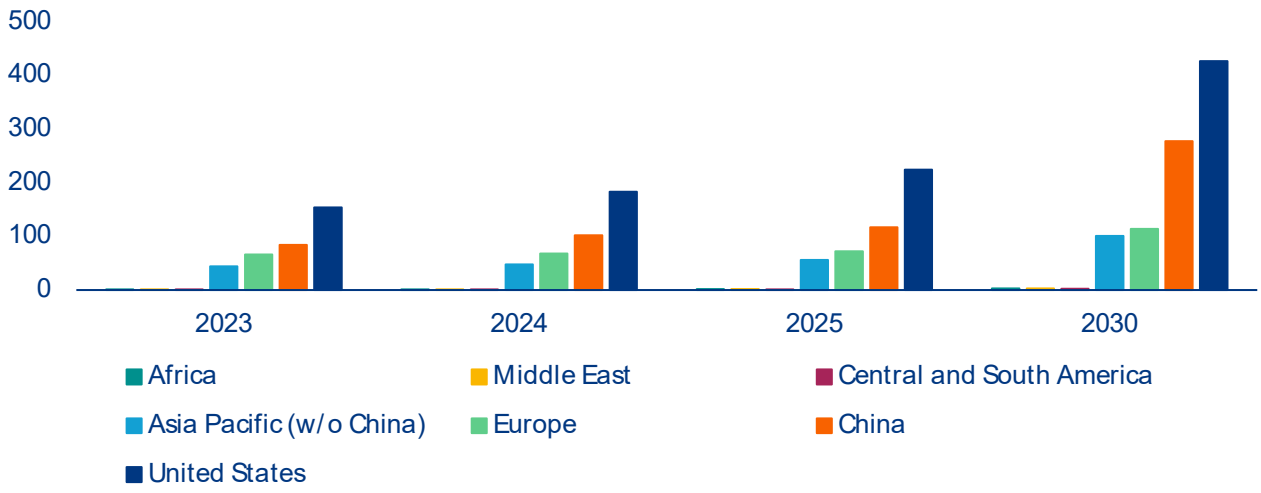


Can power markets handle the AI surge?

The AI revolution is reshaping global economies, starting with power markets. In the IEA's baseline scenario, AI-fueled growth in data-center power demand is expected to increase between 58% and 137% across major world regions between 2025 and 2030, supported by annual investment that now rivals the scale of the entire global oil and gas industry (Figure 1).[1] The shift is particularly pronounced in the US, where data-center power demand is projected to reach 426 TWh by 2030, nearly double the 2025 level. Based on long-term EIA projections, our base case estimates suggest this figure could double again by 2050, reaching roughly 810TWh. While the precise share attributable to AI remains difficult to isolate, bottom-up modeling suggests AI inference already accounts for 12–14% of total consumption, with total AI data-center capacity reaching 29.6 GW by end-2025.[2] The implications for power systems are already visible in the US, where the data center share of total electricity consumption could jump from around 5% today to between 7% and 12% by 2028, with our estimate, based on current supply pipelines and baseline demand assumptions, pointing to a share of around 9%. [3]

[1] Key Questions on Energy and AI (IEA)
[2] The 2026 AI Index Report | Stanford HAI
[3] 2024 United States Data Center Energy Usage Report | Energy Technologies Area

Figure 1: Data-center power demand growth by region (in TWh)



Source: IEA baseline projection

This significant demand growth is fueling a massive expansion in generation capacity. According to the Energy Information Administration's (EIA) Short-Term Energy Outlook, power generation in the US is expected to grow by 441 TWh between 2023 and 2027, an increase of 10.5%.[4] This would mean the five years since the breakthrough of generative AI see a larger increase than the preceding two decades, over which power generation grew by only 7.7%. 76% of the increase in power generation is expected to be driven by wind and solar (21.9% wind and 54.1% solar) with the remaining 23% supplied by gas (15.3%), nuclear (5.2%) and hydropower (3.0%). Coal remains on its post-2007 declining trend, though the pace has slowed markedly from -4% annual reductions before the AI era to just -0.5% today.

On paper, the capacity additions currently included in the EIA's pipeline of planned and under-construction projects appear broadly sufficient to accommodate rising data-center demand, with data centers expected to absorb around 47.1% of projected new supply.[5] That apparent adequacy, however, rests on the fragile assumption that competing sources of electricity demand remain relatively contained. In reality, the grid is coming under pressure from multiple directions simultaneously. Beyond data centers, electrified transport alone could materially tighten the balance. Under the IEA's baseline scenario, US electricity consumption from electric vehicles is projected to rise from 25.2TWh in 2024 to 91.4TWh by 2030. Near-term signals are mixed, but the risks remain tilted to the upside: Secondary-market EV sales continue to rise despite the expiry of the federal tax credit, while gasoline prices elevated by the escalation in the Middle East are reinforcing the economic case for adoption. Historically, sustained fuel-price pressure has been one of the strongest catalysts for accelerating EV uptake.[6] Industrial electrification adds a further layer of uncertainty: Manufacturing reshoring and the electrification of industrial processes point towards structurally higher power demand that remains difficult to forecast with precision. Taken together, these competing demand vectors suggest that the supply cushion that seems comfortable could narrow far more quickly than current projections imply.

The bigger issue lies on the supply side and is rooted in US power grids, which struggle to keep pace with data-center interconnection demands. While physical construction typically takes under 24 months, grid-connection queues can more than double total

[4] IEA STEO, April 2026 update
[5] Based on EIA-860M (February 2026 update)
[6] Global EV Outlook 2025 - IEA

project timelines, with hotspots like Northern Virginia reporting waiting times of up to seven years.[7] This bottleneck reflects a deeper structural problem as both new generation capacity and new demand face the same overloaded interconnection queues (Figure 2). In Texas (ERCOT region), generation and storage projects queuing for grid connection nearly quadrupled between 2020 and 2024, while nationally active interconnection requests now stand at 1.84 TW, exceeding total US installed generating capacity. That figure overstates the genuine pipeline as only around 20% of requests ultimately reach commercial operation, with the remainder abandoned due to speculative filings, rising costs and mounting delays. Resolving these constraints requires grid expansion on a scale the US has not attempted in decades: In their 2024 National Transmission Planning study, the DOE estimates that in a lowest-cost scenario the total transmission system would increase by 2.1-2.6 times until 2050 and up to 3.3 times in a high demand case. This would imply building around 8000km of high voltage transmission lines per year, a pace that appears far beyond current construction rates, which have averaged well below 1,000 km annually in recent years.[8]

Getting there is complicated by constraints that span the entire supply chain. The grid equipment needed (transformers, switchgear and cables) faces lead times of up to four years alongside a 30% supply shortfall.[9] Building new transmission lines compounds the problem further, taking four to eight years in advanced economies, a timeline fundamentally incompatible with the pace of data-center deployment. Beyond power infrastructure, the AI build-out is also running into semiconductor supply constraints: According to the IEA, HBM chip shortages have emerged as a binding bottleneck over the past six months, with insufficient supply of this specialized memory threatening to slow AI data-center deployment independently of grid availability.[10] A construction workforce shortage of around 439,000 skilled workers and increasingly complex permitting processes compound the problem further. As a result, almost half of the approximately 12 GW of US data center capacity planned for 2026 is expected to be delayed or cancelled, with only around 5 GW currently under active construction.[11]

Figure 2: US power capacity in queue for grid connection by electricity market region (in GW)

Source: LBNL, interconnection.fyi

[7] Epoch AI and Energy and AI
[8] DOE and Grid Strategies
[9] Wood Mackenzie
[10] Key questions on energy and AI - IEA
[11] Slightline Climate Data and Bloomberg

Meanwhile grid stress from data centers has moved from projection to reality, helping explain the prolonged integration assessments that slow their expansion. Unlike traditional commercial or residential load, data centers represent large, concentrated blocks of power demand that can disconnect or transfer to backup power in a coordinated manner during grid disturbances. Compounding this, AI-optimized data centers can exhibit sharper and more frequent load swings than conventional facilities: As workloads shift between high-intensity processing and lower-utilization periods, power draw can fluctuate rapidly, placing greater demands on frequency regulation and reserve markets. When a disturbance triggers a simultaneous disconnect, as occurred in Virginia in 2024, where 60 facilities representing 1.5 GW dropped off the grid at once, the sudden loss of load forces operators to curtail generation equally fast to avoid a frequency imbalance that can cascade into widespread blackouts.[12] The reverse risk is equally real: When data centers draw at full capacity during peak periods, they can overwhelm local transmission capacity, forcing power to reroute through congested lines and driving up congestion costs across the broader grid. This strain is increasingly forcing active intervention, as seen in January 2026 when the DOE invoked emergency powers to authorize PJM and ERCOT to shift data centers onto backup generators to prevent residential blackouts during a severe winter storm. Both failure modes reflect the same underlying mismatch: a grid designed for distributed, flexible demand is being asked to absorb loads that are geographically concentrated and operationally rigid.

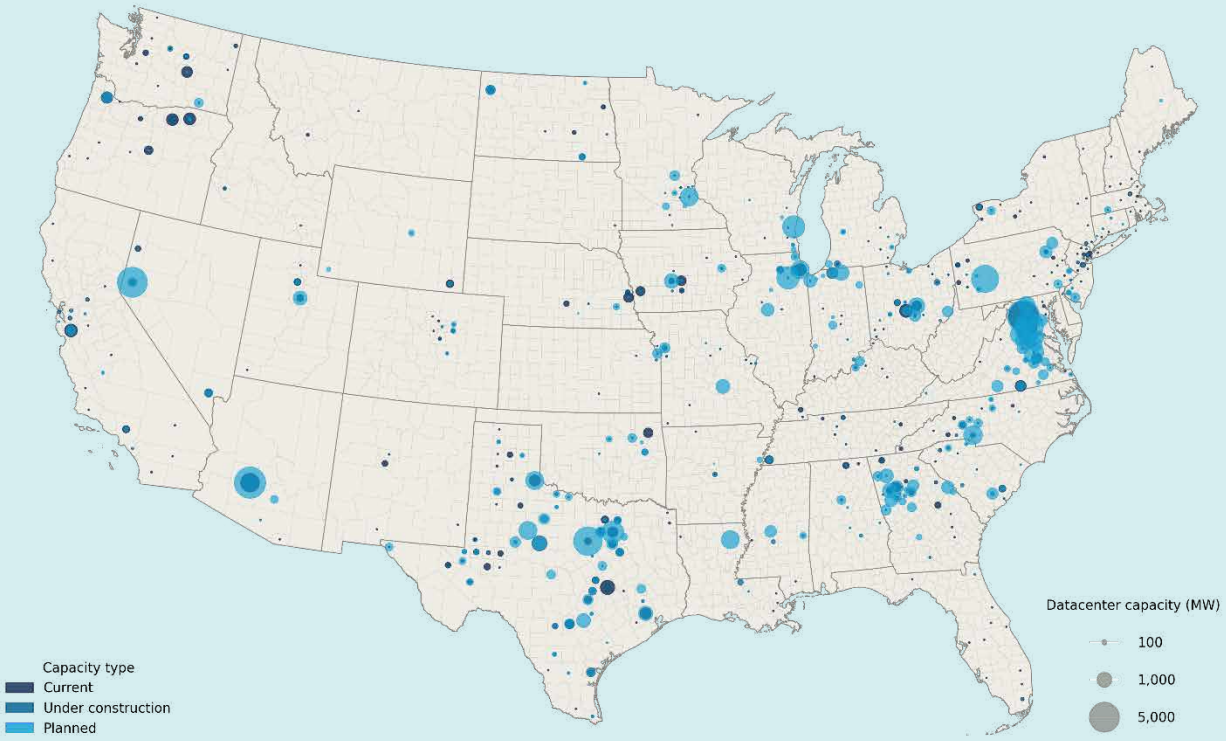
These physical bottlenecks are now increasingly intersecting with political and social constraints. Organized opposition to data-center development has emerged as an additional factor shaping the outlook. In 2025, an estimated USD156bn worth of data-center projects were blocked or delayed by local opposition, moratoriums and litigation in the US, with project cancellations more than quadrupling from six in 2024 to 25 in 2025.[13] The opposition is largely rooted in concerns about electricity costs as the infrastructure required to accommodate data-center load growth is typically funded through broader rate-payer tariffs rather than charged directly to the facilities driving demand. In PJM alone, the grid's independent market monitor estimates that data-center load growth added USD16.6bn in capacity costs across the 2025/26 and 2026/27 delivery years, spread across 67mn consumers.[14] Legislative responses are following, with moratorium bills introduced in at least 11 states in 2026, with some like Maine considering statewide construction pauses.[15] While unlikely to halt the AI infrastructure buildout, rising opposition introduces permitting delays and regulatory uncertainty that compound existing supply-side constraints and underscores a broader question about who ultimately bears the cost of AI infrastructure.

[12] DCD
[13] Data Center Watch and Construction Dive
[14] Comment of the Independent Market Monitor for PJM
[15] Axios

Where are data centers built and what determines their location?

Data center location reflects a specific set of infrastructure, cost and regulatory requirements that have historically concentrated capacity in a small number of markets. Today, Texas and Virginia alone account for 42% of total US data center capacity, though the nature of that concentration differs markedly between the two. Virginia, and Northern Virginia's Loudoun County in particular, dominates current installed capacity and represents the largest data-center cluster in the world. Texas, by contrast, carries the largest pipeline of planned and under-construction capacity, reflecting its abundance of land, a deregulated electricity market, and aggressive state-level incentives (Figure 3).

Figure 3: US data center capacity by county (in MW)



Source: National Laboratory of the Rockies.

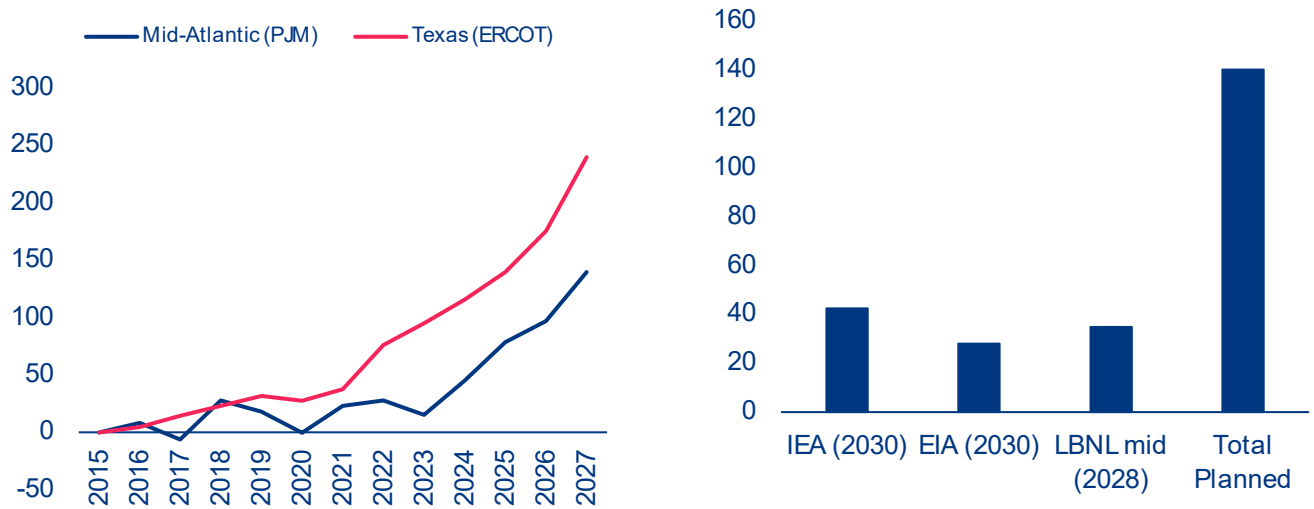
Four factors dominate siting decisions. Power availability and cost are the most fundamental as cheap and reliable electricity underpins the economics of large-scale compute. Fiber connectivity to major internet exchange points determines latency and operational viability. Land availability and permitting speed determine how quickly large campuses can be developed. Tax incentives, particularly data-center exemptions offered by Virginia, Texas and several other states, have actively directed investment toward specific markets.

The consequence of this concentration is both operational and systemic risk. When grid capacity tightens in saturated markets, the AI infrastructure buildout slows not just regionally but globally, given the role these clusters play in serving international demand. A regional grid failure in these clusters would ripple globally, translating into higher prices for AI and cloud services, slower model deployment and capacity-driven rate limiting for end users. Overclustered markets also face compounding vulnerabilities across power supply, water availability, cooling infrastructure and skilled labor, making geographic diversification of future capacity an increasingly urgent priority.

While aggregate power-supply growth appears sufficient on paper, the regional picture is more nuanced. In the two most exposed markets, ERCOT (Electric Reliability Council of Texas) and PJM (PJM Interconnection), power generation growth has accelerated sharply since 2022, jumping from 10 TWh p.a. to 22 TWh p.a. in Texas and from 4 TWh p.a. to 31 TWh p.a. in PJM (Figure 4 a). This momentum is expected to continue, with the EIA's Short-Term Energy Outlook projecting an additional 99 TWh of supply in Texas and around 60 TWh in PJM by end-2027. On the demand side, Texas and Virginia each carry approximately 30 GW of data center capacity in their pipelines, which, if fully implemented, would imply an increase in annual electricity demand of around 140 TWh per region. In practice, however, given typical build-out timelines and project attrition, a more realistic estimate puts incremental data center demand at 30–60 TWh per region by 2030 (Figure 4 b). At those levels, both markets can realistically absorb expected AI-driven load growth, provided their supply pipelines materialize.

Delivering on that supply potential is, however, not without obstacles. In ERCOT, the supply pipeline is theoretically comfortable, but interconnection queues have grown sharply in recent years, raising the risk that capacity additions lag behind the pace of data center buildout. In PJM, the challenge is more structural. Generation growth has been slower and surrounding regions offer limited import relief, given mostly flat supply growth expectations. Data centers are also not the only source of incremental demand, with electrification of transport and industry adding further pressure on available grid capacity. With 105 GW currently queued in PJM's interconnection queue, processing delays alone could push a significant share of planned capacity beyond the 2030 horizon. Should capacity additions in either market fail to keep pace with demand, persistently tight grid conditions would translate into additional upward pressure on regional electricity prices.

Figure 4: Estimated power generation growth by grid region vs 2015 (lhs, TWh) and estimated data center power demand in Texas and Virginia (rhs, TWh)



Source: Allianz research based on EIA (STEO), IEA, LBNL and National Laboratory of the Rockies; Note: Planned power demand was estimated based on state level NLR capacity data for facilities planned or under construction. IEA, EIA and LBNL country level estimates were downscaled to regions based on planned capacity shares.

Beyond the baseline: Agentic AI, adoption velocity and the rebound effect

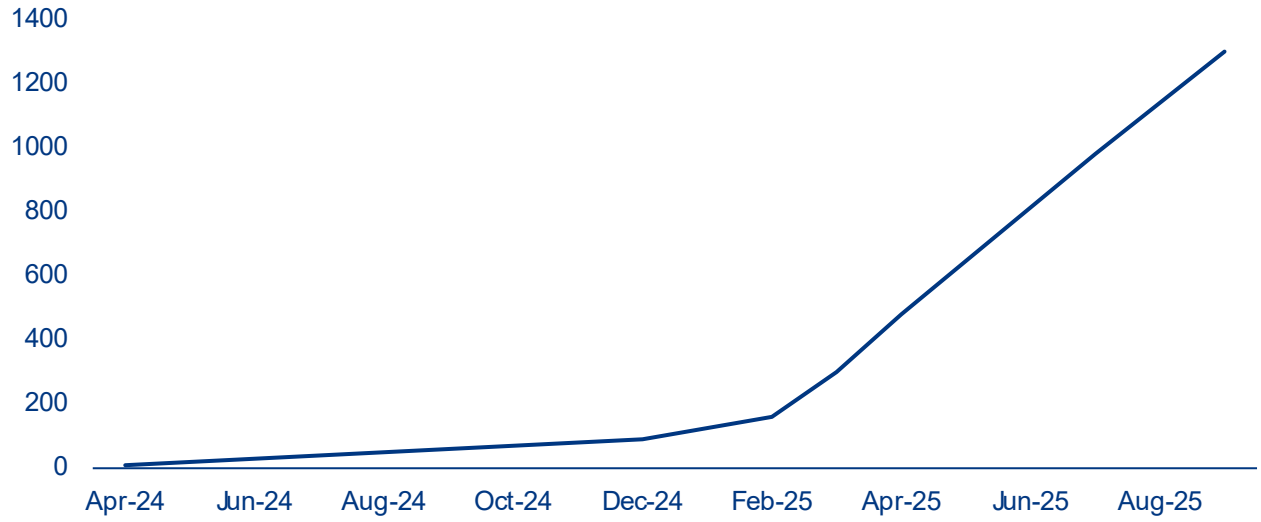
The future trajectory of AI's impact on power markets is shaped by two opposing forces. Rapid efficiency improvements are materially reducing the energy cost per token, while surging adoption, growing usage complexity and the emergence of agentic workflows are pushing demand in the opposite direction. Which force dominates will determine whether current supply projections prove adequate or significantly understated.

The energy efficiency of AI systems has improved dramatically across all layers of the technology stack... Compute performance per watt has increased by around 34% per year since 2008, with successive GPU generations delivering compounding gains.[16] At the model level, architectural innovations such as Mixture-of-Experts designs, quantization, speculative decoding and model distillation have each contributed further reductions in energy per task, in some cases cutting compute requirements by several multiples without meaningful loss in output quality. On the infrastructure side, hyperscale and AI-specialized data centers have become significantly more efficient at converting electricity into useful computation, with overhead losses from cooling and power distribution falling from around 40–50% of total facility energy in older facilities to just 10–20% in modern ones.[17] According to the IEA, widespread adoption of existing AI applications across industry, buildings and transport could deliver aggregate energy savings of over 13 EJ by 2035, equivalent to around 3% of global final energy consumption.[18] Realizing this potential, however, is far from guaranteed, as widespread adoption remains contingent on overcoming persistent barriers including fragmented data, limited digital skills and inadequate regulatory incentives.

[16] Epoch AI
[17] LBNL
[18] Key Questions on Energy and AI - IEA

...yet demand is expanding at least as fast, driven by rapid adoption, growing model capability and a structural shift in how AI is used. Today, generative AI has reached 53% population-level adoption within three years of its mass-market introduction, faster than the personal computer or the internet, with 88% of companies reporting AI use in 2025 according to Stanford's AI Index.[19] The models powering this growth have scaled at an equally striking pace, with the compute required to train frontier AI models growing at roughly 4–5x per year since 2020 and GPT-4 estimated at around 1.8trn parameters, roughly 10 times larger than GPT-3 released just three years prior.[20] Usage intensity has followed the same trajectory. According to OpenRouter's platform data, average prompt length has grown roughly fourfold since early 2024 as users shifted from simple queries to context-heavy, multi-step workflows, with reasoning models growing from a negligible share to over 50% of token usage by late 2025.[21] This growth is reflected in token consumption figures across major AI companies. OpenAI's API alone grew from 300mn tokens per minute in 2023 to over 6bn by late 2025, a 20-fold increase in under two years, while Google's Gemini token consumption increased approximately sevenfold between February and September 2025 (Figure 5).[21]

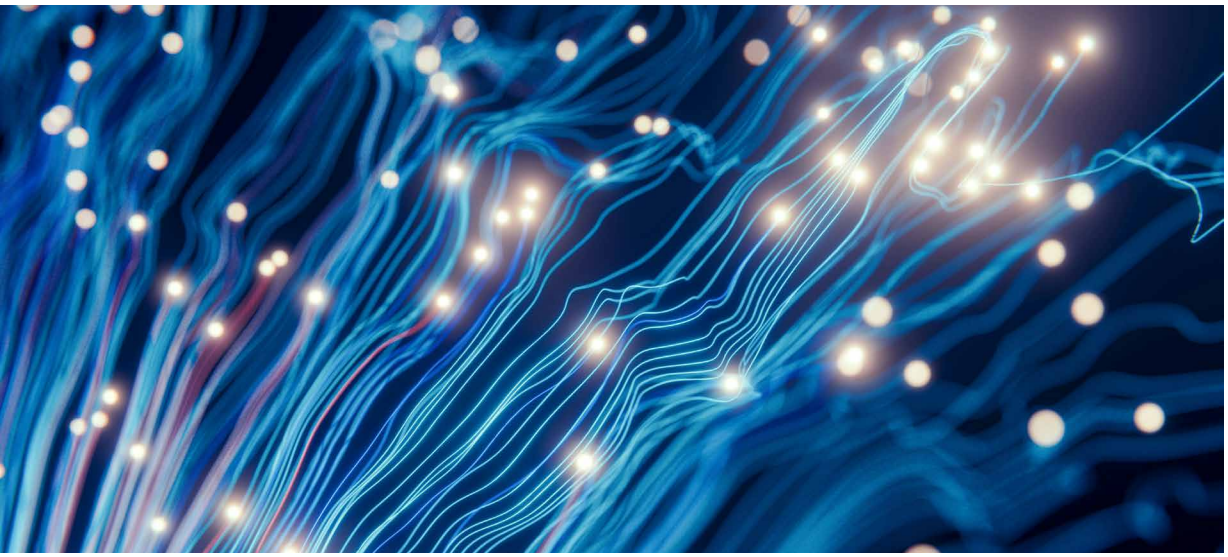
Figure 5: Google Gemini monthly token consumption (trillion tokens)



Sources: Google DeepMind; Alphabet Q2 2025 earnings.

The primary driver behind this acceleration is a fundamental shift in how AI is being used. Unlike conventional chat interactions, agentic systems plan and execute multi-step tasks autonomously, dramatically increasing token consumption per interaction. The energy implications are substantial, with an agentic task with reasoning consuming around 50Wh compared with roughly 0.3Wh for a standard text query, roughly 150 times more energy per interaction. Another reinforcing factor is the sharp decline in inference costs, which fell more than 280-fold between late 2022 and late 2024.[23] As lower prices stimulate broader and more intensive use, efficiency gains at the model and hardware level risk being offset by a classic rebound effect, whereby falling costs per unit of consumption drive total consumption higher. Taken together, these trends suggest that AI-driven electricity demand is likely to continue growing faster than efficiency improvements can offset in the near term, reinforcing the supply-side pressures on regional power markets.

[19] The 2026 AI Index Report - Stanford HAI
[20] Epoch AI
[21] OpenRouter
[22] Gemini's token growth partly reflects its integration into Google Search, which significantly expanded query volumes beyond standalone AI product usage.
[23] The 2025 AI Index Report - Stanford HAI



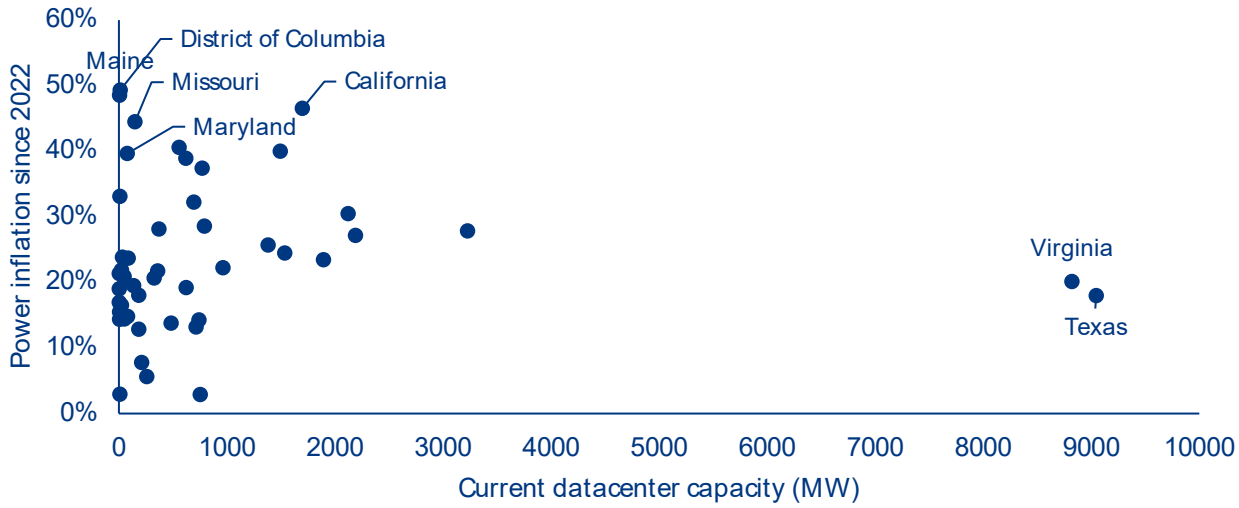
The hidden bill?: AI's impact on power prices

Data-center demand affects electricity prices through several compounding channels. Large concentrated loads drive up capacity market auction prices as grid operators must procure additional firm generation to meet anticipated peak demand, with costs socialized across all ratepayers. Grid-connection infrastructure represents a second channel as transformers, switchgear and dedicated transmission lines are typically recovered through broader utility tariffs rather than charged to the facilities driving demand. These pressures land on grids already strained by decades of underinvestment, amplifying the cost impact of each new interconnection. A third channel operates through generation mix: Data centers require firm, around-the-clock power that renewables alone cannot provide, pushing utilities toward more expensive gas and nuclear capacity. Finally, the renewable energy that data centers rely on for sustainability commitments is often located far from load centers, creating congestion costs that ripple across the broader grid. Network charges have so far remained relatively stable, but are structurally bound to rise in the regions where AI infrastructure is most concentrated. Because each channel operates on a different timescale and affects different consumer groups, the overall price impact is neither immediate nor evenly distributed across consumers and regions.

So far, aggregate state-level data tell a surprisingly benign story. Despite hosting more than 40% of total US data-center capacity, Texas and Virginia have seen power price inflation of only 18% and 20% respectively since 2022, below the 24% US average (Figure 6). This is consistent with research from Lawrence Berkeley National Laboratory (2026), which finds that state-level load growth from 2019 to 2025 was generally associated with lower all-sector average retail electricity prices as rising fixed infrastructure costs are spread across a larger base of consumers and kilowatt-hours sold.24 A further moderating factor is that data centers have historically clustered in regions with favorable grid conditions and low base energy costs, insulating those markets from the price pressures that concentrated demand might otherwise create. Regulatory lag compounds this dynamic: Utility rate cases can take well over a year from filing to approval, and capital investments are recovered over asset lifetimes rather than immediately, meaning that retail prices today largely reflect costs incurred years earlier. That lag is, however, beginning to unwind: Investor-owned utilities filed USD18bn in rate increase requests in 2025, the highest level since the mid-1980s, with regulators approving 64% of requests.

[24] Energy Markets & Planning

Figure 6: Data-center capacity and retail electricity price inflation since 2022, by state



Source: Allianz Research based on EIA and NLR.

State-level data mask considerable heterogeneity across individual utilities. A utility directly serving a major data-center cluster faces fundamentally different demand dynamics than a rural cooperative in the same state, yet both are folded into the same state-level price average. To move beyond the aggregate picture, we estimate the effect of data-center demand growth on residential electricity prices at the utility level, exploiting the fact that utilities entered the AI investment boom with very different pre-existing exposure to data center capacity.[25] The intuition is straightforward. Utilities that happened to serve counties with significant data center infrastructure were more directly in the path of the subsequent investment wave than those that did not. By interacting each utility's data center footprint with the national surge in construction investment – a design known as a Bartik shift-share instrument – we can estimate the price effect of data center demand while accounting for time-invariant utility characteristics and aggregate macroeconomic trends. The analysis covers approximately 1,200 US utilities over 2020–2024, drawing on EIA utility data matched to county-level data center capacity data from the National Laboratory of the Rockies (NLR) and national construction investment figures from the Census Bureau.

The results suggest that price pressures are more visible at the utility level than aggregate data imply, with three findings standing out. First, the price effect is concentrated entirely in the post-2023 period, precisely when AI-driven investment accelerated sharply, while the pre-2023 period shows no significant relationship between data center exposure and price changes (Table 1, column 2). This timing is difficult to explain by anything other than the AI demand shock itself, and provides confidence that we are capturing a genuine effect rather than a coincidental correlation with pre-existing trends. Second, the demand-side results validate the approach: utilities with higher data center exposure show markedly stronger commercial electricity consumption growth post-2023, consistent with hyperscale facilities connecting to utility grids and driving up load (Table 1, column 3). The effect is broad-based across utilities rather than driven by any single outlier. Third, in terms of magnitude, a utility with 4% of national data center capacity in its service territory, broadly representative of the most exposed markets outside the largest hubs, is associated with a cumulative residential price increase of approximately 3% over 2020–2024, based on the USD22bn growth in national data-center construction investment observed over that period. Modest in isolation, this effect is best understood

[25] Data center capacity shares reflect 2025 installed capacity, the large majority of which was planned and permitted before the generative AI boom took hold in 2023, making them a reasonable proxy for pre-existing exposure rather than a response to the shock we are trying to measure.

as an early signal rather than a final estimate – given the regulatory lag discussed above, much of the cost pressure already absorbed by utilities is still working its way through to consumers.

Table 1: Estimated effect of data-center demand on electricity prices

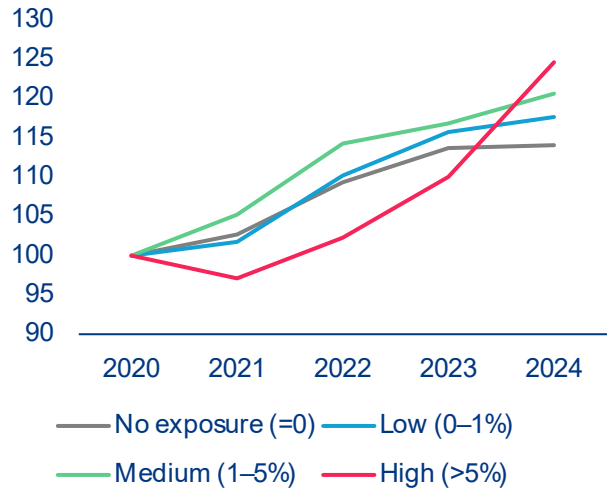
	Dep. var: log(Residential Price)		Dep. var: log(Commercial Sales/Customer)
	(1) Baseline	(2) Pre/Post Split	(3) Post-2023
Bartik IV	**0.0342*** *(0.018)*	—	—
Bartik IV × Post-2023	—	**0.0321**** *(0.0156)*	**0.2416***** *(0.0556)*
Bartik IV × Pre-2023	—	0.0145 *(0.021)*	—
Observations	6,586	6,586	6,388
Within R^2	0.0098	0.0115	0.033

Source: Allianz Research; Note: All specifications include utility fixed effects and year fixed effects. Standard errors are clustered at the utility level and reported in parentheses. The Bartik instrument is B_ut = z_u × G_t, where z_u is utility u's predetermined share of national data-center capacity and G_t is annual US data center construction investment (US Census Bureau). Specifications (1) and (2) use log residential prices as the dependent variable; (1) uses the full-period IV while (2) splits it at 2023 to test for pre-trends. Specification (3) uses log commercial sales per customer as the dependent variable with the post-2023 IV, providing demand-side validation of the instrument. Significance: *** p<0.01, ** p<0.05, * p<0.10.

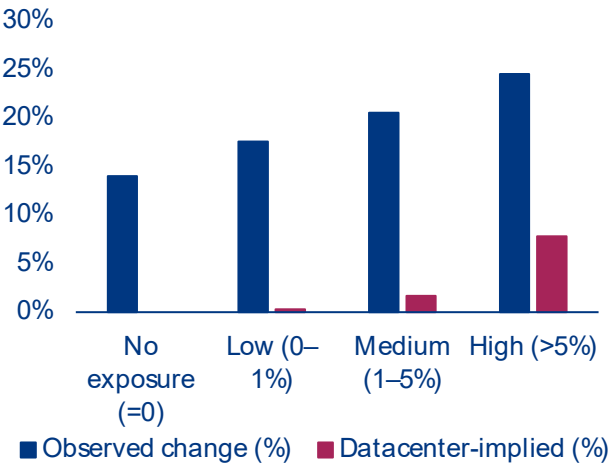
Data-center demand is leaving a measurable imprint on electricity prices, with effects concentrated in the most exposed markets. Figure 7a) shows that while all utility groups experienced meaningful price increases over 2020–2024, the pattern is far from uniform. The no-exposure group saw the smallest increase at 14%, while the high-exposure group ended the period 24.5% above its 2020 level, a gap that widened noticeably after 2023 as AI-driven investment accelerated. Of that 24.5% rise, approximately 7.8pps are directly attributable to data-center demand based on our regression estimates (Figure 7b)), with the remainder reflecting the broader energy cost pressures that affected all utilities over the period, including fuel-price volatility, rising infrastructure investment costs and general inflation. The difference between the two groups of around 10pps therefore reflects both the data-center premium and other structural differences between high and low-exposure utility markets.

Figure 7: a) Price index by exposure group (2020=100) and b) Observed vs data center-implied price change 2020-2024 (in %)

a



b



Source: Allianz Research based on EIA and NLR; Note: Exposure groups defined by utility share of national data center capacity: no exposure (n=718), low 0–1% (n=442), medium 1–5% (n=73), high >5% (n=5). Data center-implied effect = beta × avg. group capacity share × DC investment growth (USD21.91bn)

Across the US, residential customers are already paying around USD1.4bn more per year on their electricity bills as a direct result of data-center demand. Just five utilities, serving 4.4mn households – roughly 4% of residential customers – in Northern Virginia, the Pacific Northwest and Arizona, account for more than 40% of the total, with their customers facing an average extra annual bill of USD139, attributable to data-center growth. A wider group of 73 medium-exposure utilities pays around USD28 per customer per year, while the vast majority of US utilities are not affected. Strikingly, the most affected markets are not pricier than the national average to begin with. Their historical 5% cost advantage has already narrowed to 3.7% by end-2024, and the gap is set to close further as investment accelerates, eroding a buffer that has so far masked the true scale of the data-center premium.

Table 2: Estimated financial impact of data center demand on residential electricity prices by utility exposure group in 2024

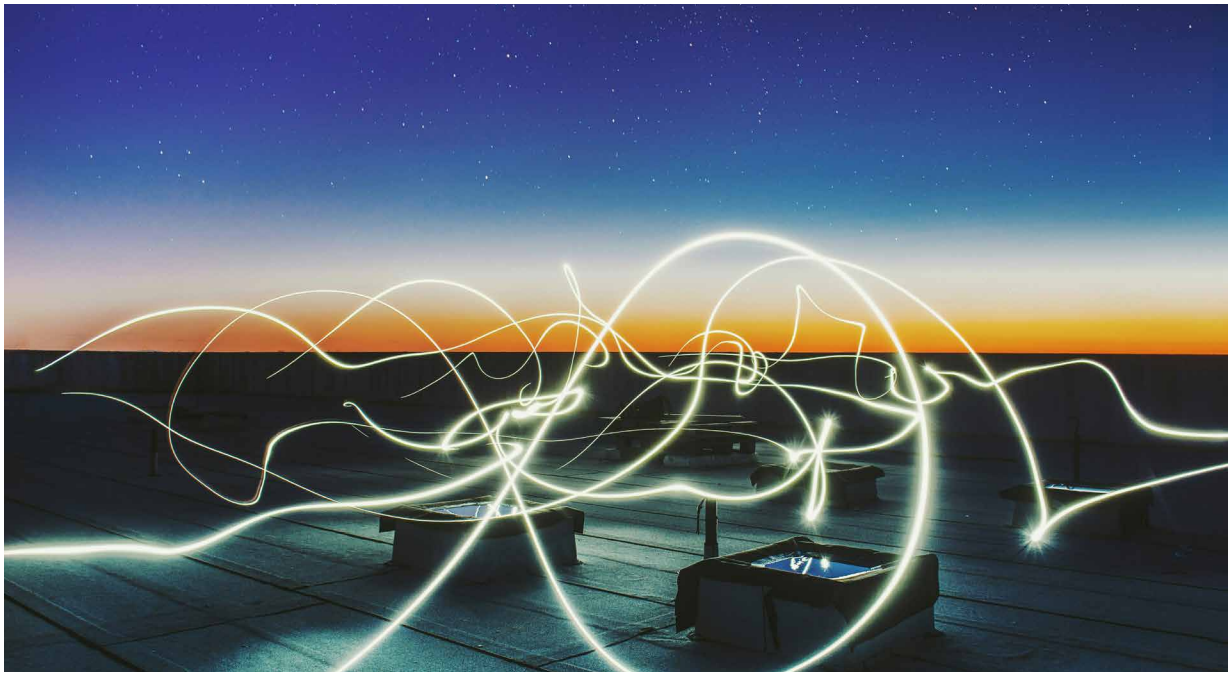
Metric	No exposure	Low (0–1%)	Medium (1–5%)	High (>5%)
Utility characteristics				
Number of utilities	718	442	73	5
Avg. exposure share (%)	0	0.17	2.3	10.47
Market size				
Avg. residential price (USD/kWh)	0.143	0.161	0.159	0.147
Total sales (TWh)	203.2	724.5	247.4	52.5
Total customers (mn)	17.3	71.42	24.08	4.37
Estimated price impact				
Estimated price effect (%)	0	0.13	1.72	7.84
Extra bill per customer (USD/yr)	0	2.1	28.05	138.65
Extra bill total (USD mn)	0	150	675	605
Inflation impact				
Direct CPI contribution (pp)	0	0.0032	0.0427	0.1945

Source: Allianz Research based on EIA and NLR

Translated into aggregate consumer prices, our findings confirm an overall modest but highly skewed data-center price premium. Across the full sample, the sales-weighted average residential price increase attributable to data centers amounts to around 0.6% over 2020–2024, translating into a direct headline CPI contribution of just 0.015pp. For the most exposed utilities, however, the same channel adds roughly 0.19pp to headline CPI, around thirteen times the full-sample figure.[26] Because these estimates capture only residential electricity and exclude commercial pass-through, they should be read as a lower bound on the true inflation impact, with the full effect on aggregate consumer prices likely materially larger once higher input costs for electricity-intensive sectors are reflected in final goods prices.

This premium is likely to grow as the investment wave continues. Data-center investment already grew by 32% in 2025 and is set to increase by a further 75% in 2026 alone.[27] As pending rate cases translate accumulated cost pressures into retail tariffs, our estimates point to an additional price effect of close to 14pps for the most exposed utilities over 2025–2026, equivalent to a further CPI contribution of around 0.34pp, nearly doubling the cumulative four-year effect in just two years. With infrastructure costs still largely socialized across all rate-payers rather than borne by the facilities driving demand, the question of who ultimately pays for the AI buildout is set to become increasingly central to the policy debate.

[26] Direct CPI contribution estimates assume an electricity share of 2.48% in the consumer price basket, based on BLS CPI relative importance weights (December 2024)
[27] Key Questions on Energy and AI - IEA



Preparing AI for energy and energy for AI: policy recommendations

The scale and pace of AI-driven power demand growth points to a set of policy priorities that are urgent, tractable and largely absent from current frameworks, which would help lay the foundation for AI ambitions that the grid can actually support.

1) Expand and accelerate

- **Reform interconnection queues.** Interconnection queues are the single most binding near-term constraint on both generation capacity and data-center deployment. Streamlining queue processes, introducing binding timelines for interconnection studies, strengthening financial penalties for speculative filings that inflate backlogs and delay genuine projects and prioritizing shovel-ready projects with firm power commitments would significantly accelerate the pace at which new capacity reaches the grid.

2) Govern & allocate

- **Improve transparency and demand tracking.** Mandatory disclosure of data-center energy consumption, water use and interconnection requests would materially improve capacity planning. Incorporating large-load interconnection requests into publicly accessible data systems would make the pace and geography of data-center expansion visible to regulators, utilities and the public in real time.

- **Address overclustering risks in saturated markets.** The concentration of approximately 42% of US data-center capacity in Texas and Virginia creates systemic risks that extend beyond local grid stress. Overclustered markets face compounding vulnerabilities across power supply, water availability, cooling infrastructure, and skilled labor, while their position as critical nodes in

global AI value chains means that a regional grid failure or prolonged supply disruption could have disproportionate consequences for AI deployment globally. Policymakers and grid operators should assess concentration thresholds beyond which new approvals in saturated markets require enhanced scrutiny, consider incentive structures that actively redirect investment toward underutilized regions with available grid capacity, and prioritize locations with proximity to renewable energy resources to reduce long-run dependence on firm fossil fuel generation.

- **Require data centers to internalize their infrastructure costs.** The socialization of capacity costs onto existing ratepayers is both economically inefficient and politically unsustainable. Policymakers should establish frameworks requiring large new loads to bear a proportionate share of the transmission and capacity infrastructure their demand necessitates. Time-of-use and peak-pricing tariffs that reflect system stress would more accurately align data-center costs with the grid impacts they create, reducing the cross-subsidization of large loads by residential ratepayers. Where market-based pricing mechanisms are insufficient, direct compensation requirements, obliging data centers to offset the costs their demand imposes on host communities, provide an alternative route to addressing the equity dimension and reducing the political friction that is increasingly delaying project approvals.

3) Optimize & decarbonize

- Implement mandatory energy-efficiency standards. No binding energy performance standards currently apply to data centers in most jurisdictions despite their growing share of national electricity consumption. Minimum PUE thresholds, water-usage effectiveness requirements and mandatory energy intensity reporting would establish a baseline accountability framework, with stricter requirements for new hyperscale builds where best-in-class efficiency is already technically and economically achievable.[28]

- **Develop demand-flexibility frameworks.** Full curtailment of data center operations is economically impractical given facility capital intensity, but partial demand flexibility is more tractable and largely untapped. Batch workloads such as model training and non-time-sensitive inference can in principle be shifted to off-peak periods, reducing peak demand pressure and congestion costs without compromising operational continuity. Regulators and grid operators should develop demand-response frameworks that reward this flexibility through capacity market mechanisms or dynamic pricing signals.

- **Ensure AI powers the energy transition rather than slowing it.** The risk that AI-driven data-center growth locks in years of additional fossil-fuel generation is real and underappreciated. Behind-the-meter generation, power purchase agreements tied to additionality requirements, and prioritization of nuclear and long-duration storage for firm power needs would reduce this risk. AI infrastructure and decarbonization are not inherently in tension. The former should be designed to accelerate the latter.

[28] Data Centers and Their Energy Consumption - Congressional Research Service


Our
team

Chief Investment Officer
& Chief Economist
Allianz Investment Management SE

Ludovic Subran
ludovic.subran@allianz.com

Head of Economic Research
Allianz Trade

Ana Boata
ana.boata@allianz-trade.com

Head of Macroeconomic & Capital Markets Research
Allianz Investment Management SE

Bjoern Griesbach
bjoern.griesbach@allianz.com

Head of Outreach
Allianz Investment Management SE

Arne Holzhausen
arne.holzhausen@allianz.com

Head of Corporate Research
Allianz Trade

Ano Kuhanathan
ano.kuhanathan@allianz-trade.com

Head of Thematic & Policy Research
Allianz Investment Management SE

Katharina Utermoehl
katharina.utermoehl@allianz.com

Macroeconomic Research

Lluis Dalmau Taules
Economist for Africa & Middle East
lluis.dalmau@allianz-trade.com

Maxime Darmet Cucchiarini
Senior Economist for UK, US & France
maxime.darmet@allianz-trade.com

Jasmin Gröschl
Senior Economist for Europe
jasmin.groeschl@allianz.com

Françoise Huang
Senior Economist for Asia Pacific
francoise.huang@allianz-trade.com

Maddalena Martini
Senior Economist for Southern Europe & Benelux
maddalena.martini@allianz.com

Luca Moneta
Senior Economist for Emerging Markets
luca.moneta@allianz-trade.com

Giovanni Scarpato
Economist for Central & Eastern Europe
giovanni.scarpato@allianz.com

Outreach

Heike Baehr
Content Manager
heike.baehr@allianz.com

Maria Thomas
Content Manager and Editor
maria.thomas@allianz-trade.com

Corporate Research

Guillaume Dejean
Senior Sector Advisor
guillaume.dejean@allianz-trade.com

Maria Latorre
Sector Advisor, B2B
maria.latorre@allianz-trade.com

Maxime Lemerle
Lead Advisor, Insolvency Research
maxime.lemerle@allianz-trade.com

Sivagaminathan Sivasubramanian
ESG and Data Analyst
sivagaminathan.sivasubramanian@allianz-trade.com

Pierre Lebard
Public Affairs Officer
pierre.lebard@allianz-trade.com

Thematic and Policy Research

Michaela Grimm
Senior Economist,
Demography & Social Protection
michaela.grimm@allianz.com

Patrick Hoffmann
Economist, ESG & AI
patrick.hoffmann@allianz.com

Simon Krause
Economist, ESG & Insurance
simon.krause@allianz.com

Hazem Krichene
Senior Economist, Climate
hazem.krichene@allianz.com

Kathrin Stoffel
Economist, Insurance & Wealth
kathrin.stoffel@allianz.com

Markus Zimmer
Senior Economist, ESG
markus.zimmer@allianz.com

Recent Publications

12/05/2026 | Behind the gate: The promises and perils of private markets democratization
06/05/2026 | US large banks: The peak of the cycle is not the time to be complacent
05/05/2026 | Automotive: Will the Middle East crisis supercharge EV momentum?
29/04/2026 | Staycation summer? Jet-fuel crunch reshapes the peak holiday season
23/04/2026 | Energy shock and policy response: Once bitten, twice shy?
22/04/2026 | Global Insolvency Outlook: Brace for Middle East spillovers
31/03/2026 | Economic outlook 2026-27: The Fog of War
25/03/2026 | AI capex cycle: war-proof for now
23/03/2026 | Signal without response: Why the EU ETS needs resolve, not redesign
17/03/2026 | Not all Emerging markets are equal: Hormuz, triple deficits, and the new energy risk premium
16/03/2026 | Warsh's double dilemma: when the Middle East rewrites the Fed's playbook
11/03/2026 | Allianz Social Resilience Index 2025: The Middle-Resilience Trap
10/03/2026 | The second energy shock: Why Europe still isn't energy secure**t**
05/03/2026 | Closing the gender income gap: from paycheck to pension
03/03/2026 | Conflict in the Middle East: Implications for markets and macro
25/02/2026 | Mining for the future: Addressing liabilities and unlocking sustainable transition opportunities
24/02/2026 | Variable geometry for European trade: Building resilience and diversification
23/02/2026 | Schroedinger's tariffs
20/02/2026 | Private equity in transition: from distribution drought to selective recovery
19/02/2026 | Eurobonds – A window of opportunity for a strategic necessity
16/02/2026 | Country Risk Atlas 2026: Under the surface
13/02/2026 | Fragmentation tests US market primacy and reshapes the global investment landscape
11/02/2026 | High prices, thin buffers: America's affordability crisis persists
10/02/2026 | Europe's households after the rate shock: A windfall for some, a squeeze for others
04/02/2026 | Team Italy: An economic performance worthy of a gold medal?
04/02/2026 | Five things that could derail the ECB
02/02/2026 | A new decade high for major insolvencies driven by services, retail and construction
29/01/2026 | From Japan with love: New policy stance creates both market opportunities and liquidity risks
27/01/2026 | Trade receivables in a fragmented world: Navigating collection complexity
26/01/2026 | EU-India trade deal: EUR30bn of combined yearly exports gains in a fragmented world
23/01/2026 | Eyes back on the Fed (and on interventionist financial policies)
21/01/2026 | Tackling the insurance protection gap
20/01/2026 | The heat is on: Unlocking Germany's heat-pump potential
16/01/2026 | Geopolitics heats up from Venezuela, to Greenland to Iran, but investors shrug. For how long?
14/01/2026 | Allianz Risk Barometer - Identifying the major business risks for 2026
17/12/2025 | Economic Outlook 2026-27: Stretching the limits
11/12/2025 | What to watch
10/12/2025 | Convertible bonds: The Financial Roadster for Dynamic Markets
05/12/2025 | What to watch
02/12/2025 | High hopes, heavy footprint: Aviation's quest for climate-neutral skies

Discover all our publications on our websites: Allianz Research and Allianz Trade Economic Research

Director of Publications
Ludovic Subran, Chief Investment Officer & Chief Economist
Allianz Research
Phone +49 89 3800 7859

Allianz Group Economic Research
https://www.allianz.com/en/economic_research
http://www.allianz-trade.com/economic-research
Königinstraße 28 | 80802 Munich | Germany
allianz.research@allianz.com

@allianz

allianz

Allianz Trade Economic Research
http://www.allianz-trade.com/economic-research
1 Place des Saisons | 92048 Paris-La-Défense Cedex | France

@allianz-trade

allianz-trade

About Allianz Research
Allianz Research encompasses Allianz Group Economic Research and the Economic Research department of Allianz Trade.

Forward looking statements

The statements contained herein may include prospects, statements of future expectations and other forward-looking statements that are based on management's current views and assumptions and involve known and unknown risks and uncertainties. Actual results, performance or events may differ materially from those expressed or implied in such forward-looking statements.
Such deviations may arise due to, without limitation, (i) changes of the general economic conditions and competitive situation, particularly in the Allianz Group's core business and core markets, (ii) performance of financial markets (particularly market volatility, liquidity and credit events), (iii) frequency and severity of insured loss events, including from natural catastrophes, and the development of loss expenses, (iv) mortality and morbidity levels and trends, (v) persistency levels, (vi) particularly in the banking business, the extent of credit defaults, (vii) interest rate levels, (viii) currency exchange rates including the EUR/USD exchange rate, (ix) changes in laws and regulations, including tax regulations, (x) the impact of acquisitions, including related integration issues, and reorganization measures, and (xi) general competitive factors, in each case on a local, regional, national and/or global basis. Many of these factors may be more likely to occur, or more pronounced, as a result of terrorist activities and their consequences.

No duty to update

The company assumes no obligation to update any information or forward-looking statement contained herein, save for any information required to be disclosed by law.