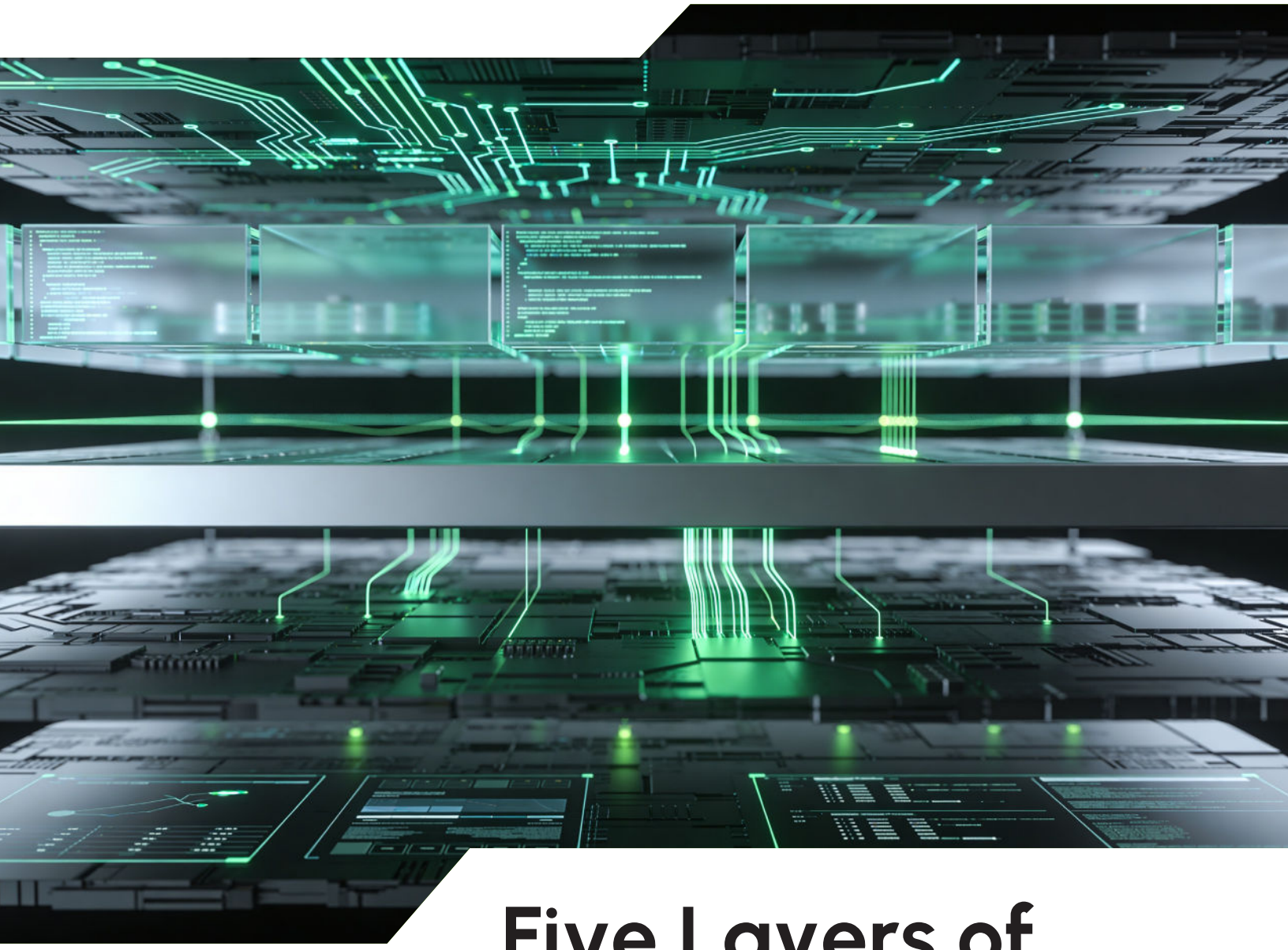
Futurum®

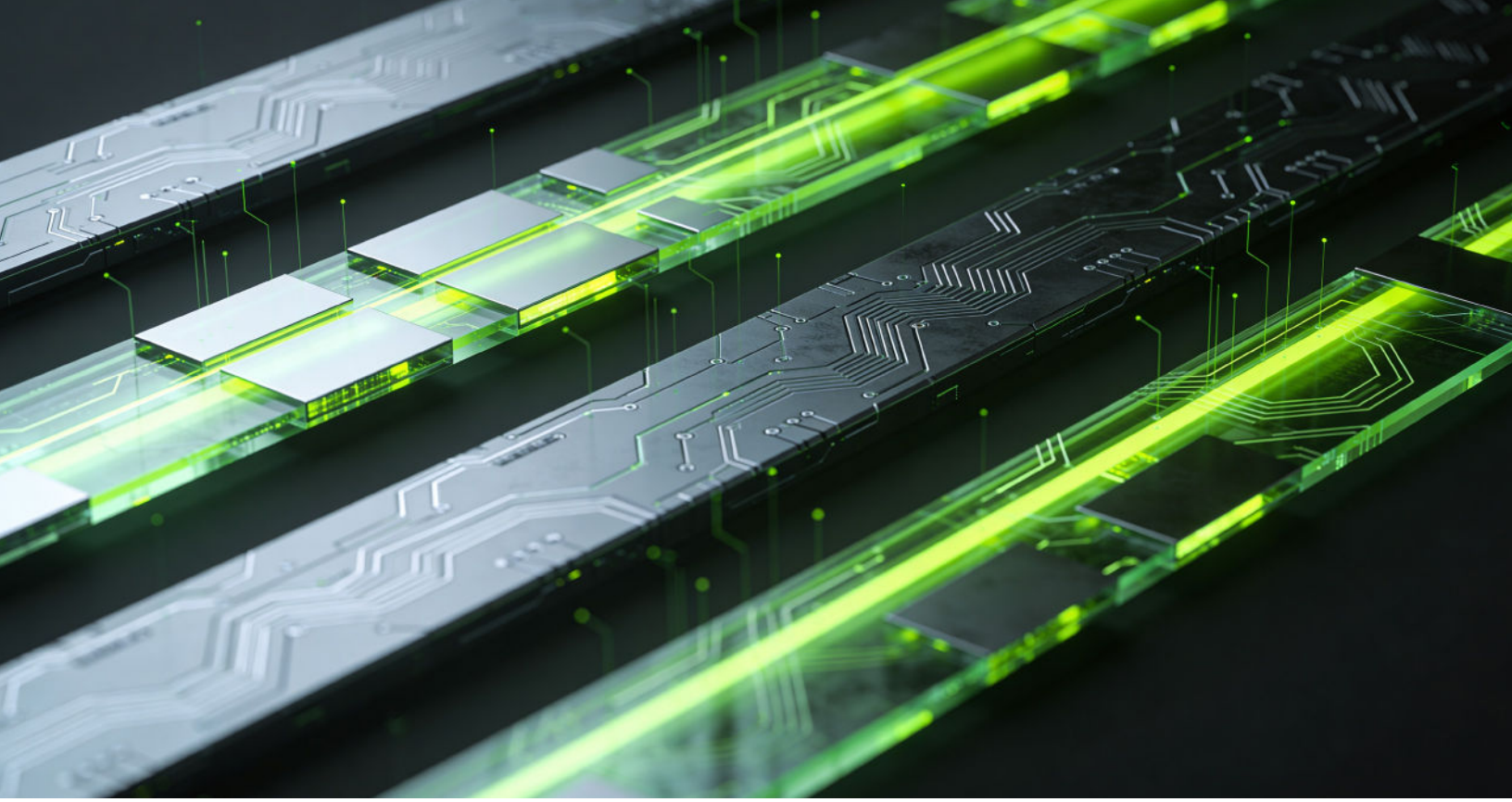
NVIDIA



Five Layers of the AI Cake

A Framework for Policymakers, Investors, and Enterprise Leaders Navigating the AI Build-Out

MAY 2026



AI Is Infrastructure, Not Just Software

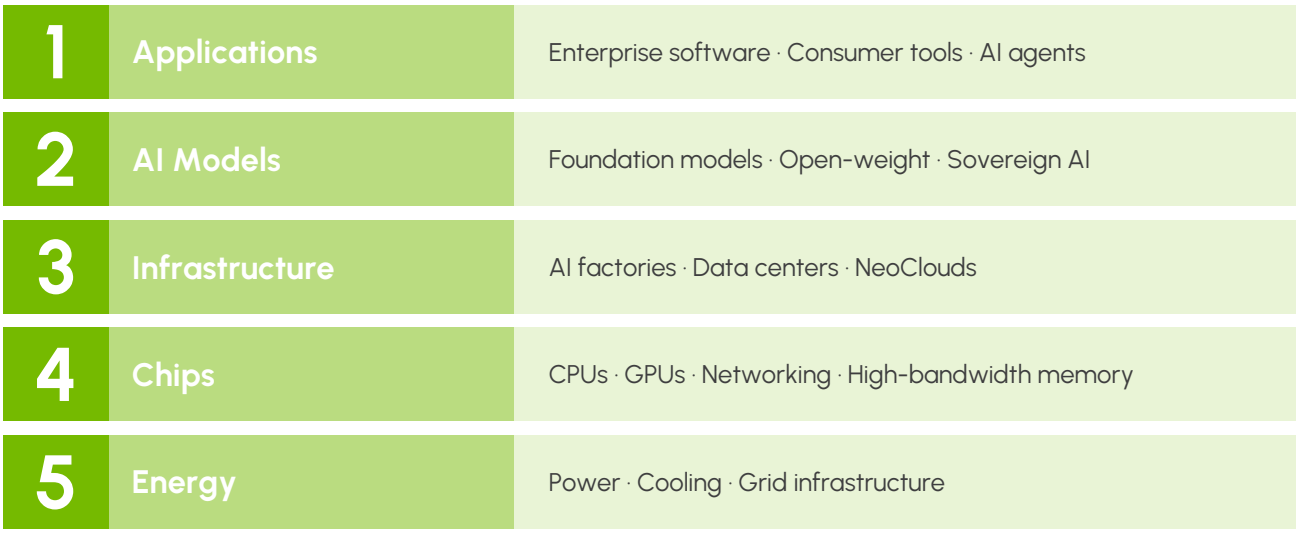
The sheer scale of the AI infrastructure build-out is best captured by the numbers: by 2026, the five largest U.S. hyperscalers are set to pour up to $690 billion into infrastructure, according to Futurum Research. This is a near-doubling of 2025 investment levels in just twelve months. Such massive, concentrated capital expenditure suggests we are witnessing more than a simple product cycle; we are seeing the birth of a foundational utility on par with the electric grid or the global telecommunications network.

Public discussion of artificial intelligence has coalesced around a narrow set of topics: large language models, chatbots, and the question of whether they will automate white-collar work. That framing is incomplete and, from a policy standpoint, misleading. Speaking with BlackRock CEO Larry Fink at the World Economic Forum in Davos 2026, NVIDIA Founder and CEO Jensen Huang described AI as a five-layer cake: energy, chips, computing infrastructure, models, and applications. Each layer must be built, financed, staffed, and operated. Each has distinct economics, distinct bottlenecks, and distinct national-security implications.

Our view is that the cake metaphor is more than rhetorical. It is the most useful analytical lens currently available for understanding where value is created in the AI economy, where the real bottlenecks lie, and how countries and enterprises should allocate attention. This paper walks through each layer, examines the workforce and economic implications, and sets out what countries and enterprises need to consider if they want to participate in the AI economy rather than simply consume its outputs.

The Five Layers

Figure 1. The Five-Layer Framework for AI infrastructure



Source: Adapted from NVIDIA (Huang, Davos 2026) and Futurum Research

Layer 1: Energy

The base of the cake is energy, and it is increasingly the gating factor. AI data centers are among the most energy-intensive facilities ever built. Futurum's 2026 outlook concluded that energy and cooling constraints have now surpassed silicon availability as the primary bottleneck for AI expansion, which is a meaningful shift from the picture 12 months ago.

This is driving renewed investment in energy generation including nuclear, natural gas and renewables, alongside grid modernization and behind-the-meter power. One underappreciated dynamic is that the U.S. grid already holds an estimated 100 to 150 gigawatts of stranded capacity sitting between baseload and peak demand: power that could serve AI workloads if those workloads can flex around grid conditions. NVIDIA DSX Flex is designed to enable exactly this: AI factories that can ramp inference workloads, power-cap GPUs, or pause training jobs when utilities hit peak demand, then consume available capacity the rest of the time.

Unlike traditional data centers, which run ERP, email, and other workloads requiring near-continuous availability, AI factories can be architected with this flexibility from the ground up. That distinction matters because obtaining a new grid connection in many markets now takes 8 to 12 years. At the same time, the AI infrastructure build-out is catalyzing much-needed investment in the U.S. energy grid itself: large, long-duration demand signals from AI facilities are supporting grid modernization, incentivizing new generation capacity and helping to spread fixed infrastructure costs more broadly.

There is also a counter intuitive pricing dynamic that policymakers should understand. Most grid infrastructure costs - poles, wires, substations - are fixed and shared across all users. When large-load customers such as AI data centers add consistent demand, those fixed costs are spread across a broader base, which can moderate per-customer bills rather than raise them. A Lawrence Berkeley National Laboratory study found that states with growing electricity consumption saw real prices fall, while states with shrinking consumption saw prices rise[1]. Additionally, PG&E forecasts have shown that each new 1 GW of new data center load would reduce electricity prices by 1-2%. This does not mean AI data centers are cost-free for local grids, but

the premise that they automatically translate into higher household energy bills deserves scrutiny. Where unique grid upgrades are required, regulators can require the operator - not local taxpayers - to bear those costs.

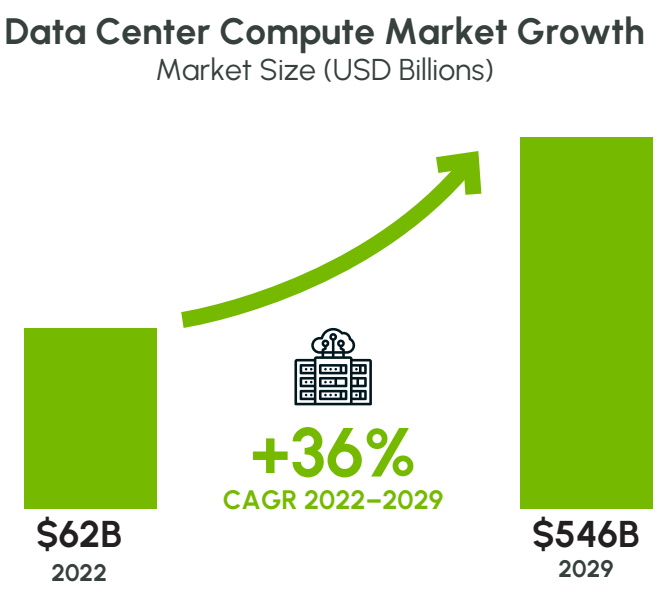
Water and cooling are no longer secondary considerations. 45°C warm-water liquid-cooling, led by NVIDIA's Blackwell platform, are designed to maximize water efficiency and enable operation without running chillers in most climates and most of the year. This represents a step change versus traditional cooling approaches, with NVIDIA systems already engineered for high-temperature liquid cooling while much of the industry is still transitioning. As AI data center build-out accelerates, however, watershed-level planning and permitting have not kept pace, making water, cooling, and power increasingly interconnected considerations for operators and policymakers.

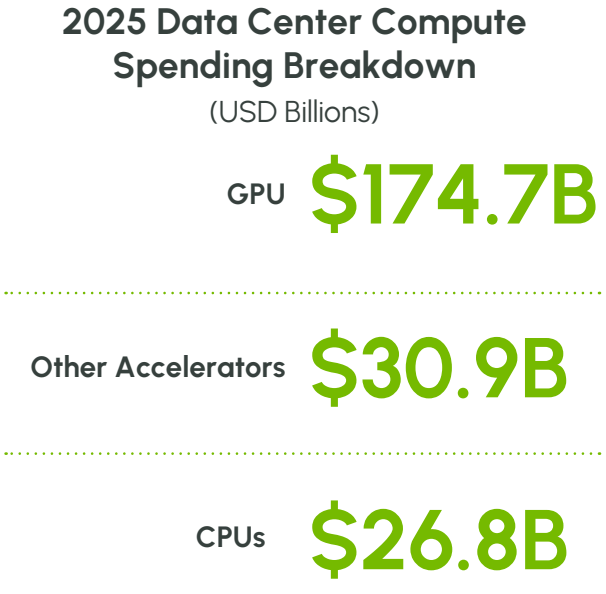
A second dimension - one that flows from the application layer back down to the energy layer - is AI for energy. Energy providers are applying AI to grid simulation, distributed-resource management and predictive maintenance, and are building AI-native tools to optimize both generation and consumption. This creates a flywheel: AI consumes energy, but it also helps produce and distribute more of it more efficiently.

NVIDIA ecosystem partners at this layer include Schneider Electric, GE Vernova, and Jacobs Engineering. A concrete example of that collaboration is the 800-volt DC power delivery standard for next-generation AI factories, backed by more than 30 industry participants including Vertiv, ABB, and Texas Instruments. By replacing conventional AC distribution with a single high-voltage DC conversion at the substation perimeter, the architecture enables 85% more power through the same cabling and reduces total cost of ownership at gigawatt scale by around 30%, which is a meaningful efficiency gain as rack densities push toward one megawatt.

Layer 2: Chips and Semiconductor Manufacturing

The silicon layer underpins everything above it: GPUs, networking components (InfiniBand, NVLink), CPUs and high-bandwidth memory. Futurum Research tracks the data center compute market growing from $62 billion in 2022 to a projected $546 billion by 2029, a compound annual growth rate of roughly 36%. In 2025, GPUs accounted for $174.7 billion of that spending, compared with $30.9 billion for other accelerators (XPUs) and $26.8 billion for CPUs.





1 Wiser, R. et al. "Factors Influencing Recent Trends in Retail Electricity Prices in the United States." Lawrence Berkeley National Laboratory and Brattle Group, October 2025.

What is less well understood outside the industry is that chips are being redesigned from the ground up because efficiency - measured in performance per watt - now determines how fast intelligence can scale. This requires extreme co-design of the entire system at data-center scale, alongside continuous software optimization. NVIDIA is the leader as the designer of the accelerated-computing platform, but the competitive picture is not static. What is worth noting is the pace at which NVIDIA continues to innovate: the Rubin and Vera architectures, NVLink 6 and the breadth of the CUDA software ecosystem create a compounding advantage, forcing competitors to meet an ever-advancing standard, rather than a stationary one. AMD's MI-series accelerators have gained traction in specific inference workloads and custom silicon from hyperscalers, such as Google's TPU, Amazon's Trainium and Microsoft's Maia is gradually being used to serve portions of first-party demand. Our view is that NVIDIA's platform advantage will hold for the balance of this decade, even as silicon layer becomes multi-vendor..

New fab investments run into tens of billions of dollars per facility, and many of these plants are being built in the United States under the CHIPS Act. This is part of a broader reindustrialization effort that is also rebuilding a domestic manufacturing labor base - a point frequently lost in discussions that frame AI as purely a knowledge-economy story.

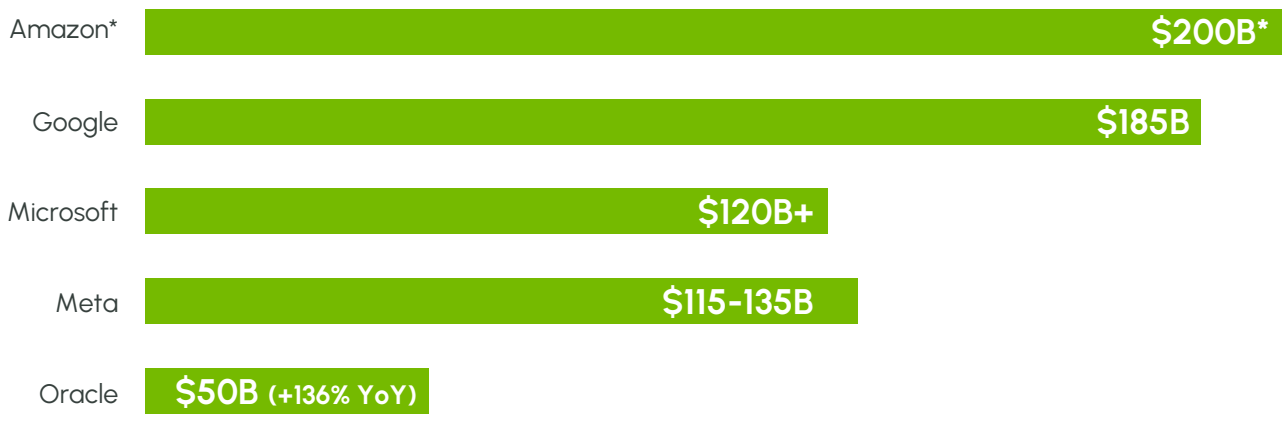
Layer 3: Computing Infrastructure and AI Factories

This is where AI factories come into focus. Unlike conventional data centers, which store and retrieve pre-recorded data, AI factories generate intelligence at scale. Three kinds of new facilities are going up simultaneously: chip fabs, systems-manufacturing plants, and AI factories themselves.

The systems layer involves rack-scale integration of GPUs, networking, liquid cooling, and software; NVIDIA's GB300 NVL72 platform illustrates the complexity. Infrastructure builders such as DPR, Caterpillar, and Vertiv construct and equip these facilities. Server OEMs including Dell, HPE, Lenovo, and Supermicro assemble and deliver the compute itself. Capacity is being built out by hyperscalers (AWS, Google Cloud, Microsoft Azure, CoreWeave) alongside a rapidly emerging set of regional and sovereign providers, including neoclouds such as Nebius and Nscale that are enabling sovereign AI deployments in dozens of countries.

The capital intensity is extraordinary (see Figure 2). Microsoft is tracking toward $120 billion or more in AI-related capex in its 2026 fiscal year. Meta's range of $115 to $135 billion includes a 1-gigawatt data center in Ohio and a Louisiana facility that could eventually scale to 5 gigawatts. Oracle's projected $50 billion represents a 136% year-over-year increase, underpinned by $553 billion in remaining performance obligations. These are not speculative figures; they are contracted demand.

Figure 2. Selected Hyperscaler AI Capex, 2026 (Projected)



*Amazon figure covers total company capex including logistics and fulfilment, not AI/cloud only.
Futurum Intelligence projects combined capex for the five largest U.S. hyperscalers at USD 660–690B in 2026.

Source: Futurum Research

Layer 4: AI Models

This is the layer the public sees most directly: large language models, multimodal models, and reasoning models have reached nearly a billion users in roughly two years. Frontier model training runs now might involve hundreds of thousands of GPUs operating in concert, with reinforcement learning post-training becoming a distinct and compute-intensive workload to train domain-specific agents. Inference for reasoning models has become an at-scale challenge in its own right, requiring performant compute, memory, fabrics, and networking to deliver responsive user experiences cost-effectively.

Open-weight and proprietary ecosystems coexist across geographies. In the West, Meta's Llama, Mistral, Cohere and NVIDIA's Nemotron anchor the open-weight camp, while OpenAI, Anthropic, and Google hold the dominant positions in proprietary frontier models by usage. Chinese open source models DeepSeek and Alibaba's Qwen have become the most-downloaded open-source models globally, with Chinese labs collectively shipping a new top-performing model roughly every four to six weeks. That cadence has direct implications for sovereign AI strategy: any country evaluating its model options in 2026 is choosing from a genuinely global menu, not just a U.S.-centric one.

Models are becoming more efficient and the cost per million tokens has fallen dramatically, approximately 35x cheaper on Blackwell than on Hopper, but aggregate compute demand continues to rise driven by two compounding forces: the number of users and applications grows constantly, while each interaction consumes more tokens as reasoning models and agentic workflows become standard. Scaling laws, mixture-of-experts architectures, and agentic AI are absorbing the efficiency gains and then some.

At the national level, our view is that countries should develop models trained on their own languages, cultural contexts, and domain data rather than relying solely on U.S.-built systems. This is not about rejecting American models; it is about ensuring that national intelligence is part of the national ecosystem. The relationship between model proliferation and the hardware platforms those models are optimized for will also shape who leads the global AI race. Open-weight model leadership matters here: countries and enterprises choosing infrastructure will increasingly favor platforms where their preferred models run best, which gives the United States a strategic interest in maintaining leadership in open-weight AI alongside its proprietary frontier models.

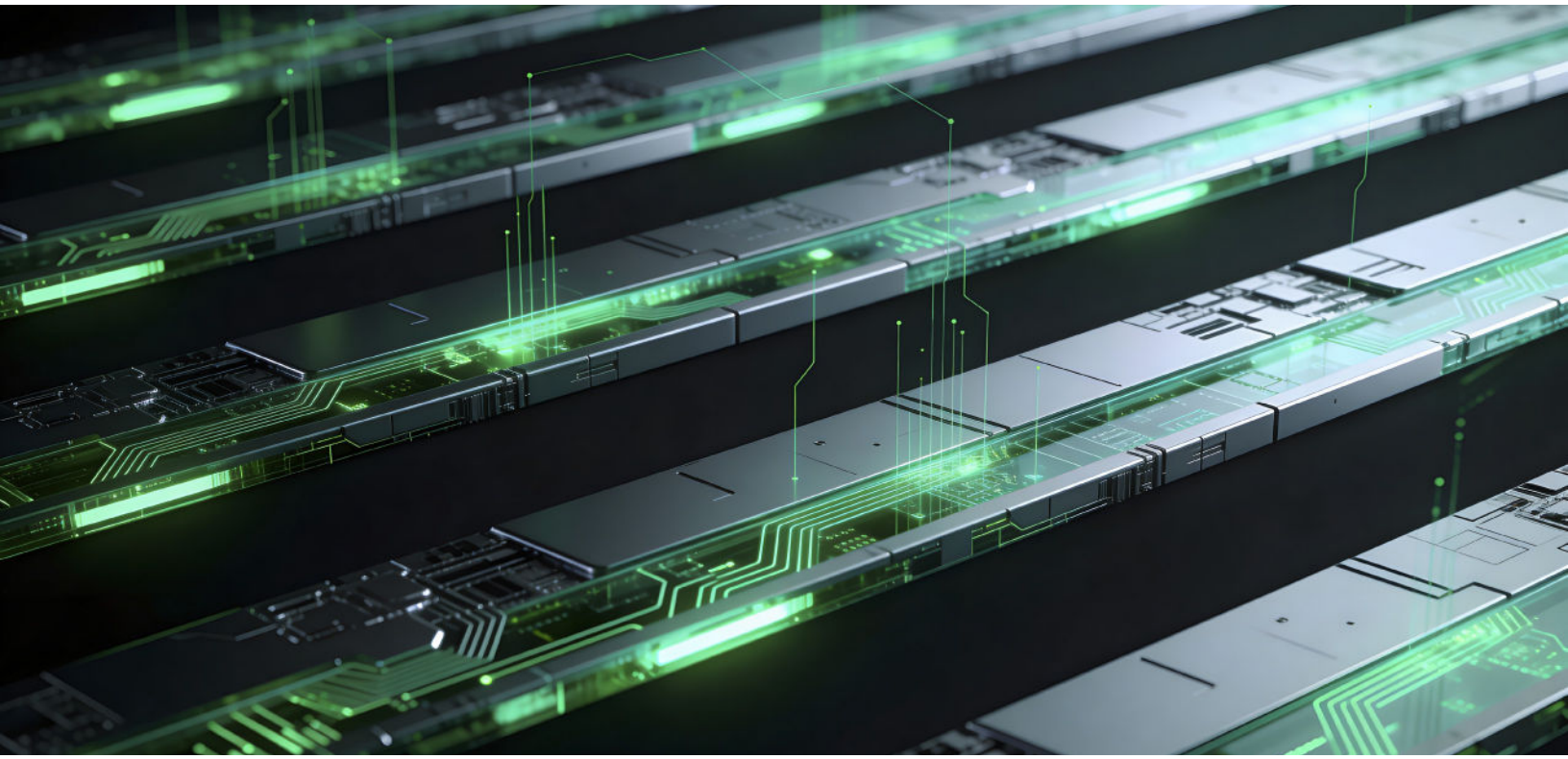
Layer 5: Applications

The application layer is where value reaches end users: enterprise software, consumer tools, and industry-specific applications and it is ultimately the reason the entire stack exists. Productivity gains, new revenue streams and operational efficiencies are realized here; that demand for economic returns is what pulls investment through the four layers beneath, from energy to chips to infrastructure to models. Capital committed to those lower layers is underwritten by the returns realized at this one It attracted the highest deal volume of any AI segment in 2025, with more than 4,600 venture-backed transactions recorded by PitchBook[2] - a signal of where entrepreneurs and investors see the closest path to commercialization.

That investment wave is generating justified enthusiasm, but also a recurring concern among policymakers: whether AI applications will displace workers at scale rather than augment them. The evidence so far is more nuanced. Radiology is the most frequently cited example: AI-powered scan analysis has so far been associated with continued or increased demand for radiologists, who can now process more cases and devote more time to high-complexity diagnosis. Causation is debated, but the pattern cuts against the simple displacement narrative.

The application layer is also being reshaped by AI agents: software that can use tools, automate multi-step workflows, and act autonomously on behalf of users. Our view is that this is the most consequential development at the application layer - not because agents will replace end users, but because they become end users themselves, consuming APIs and applications on behalf of humans. That fundamentally changes how software is built, priced, and distributed. Enterprises planning for the next five years need to treat agents as a new customer category, not merely a new feature.

The critical insight of the five-layer framework is this: the application layer only works if the four layers below it are functioning. A country with world-class AI startups but inadequate energy, no fabrication base, limited data-center capacity, and no domestic model development is not participating in the AI economy - it is dependent on it.



2 https://pitchbook.com/news/reports/q4-2025-ai-vc-trends

Workforce and Economic Impact

The dominant public narrative about AI and jobs focuses on displacement. The five-layer framework points to a different story, and one that deserves more attention from policymakers than it currently receives.

AI infrastructure build-out and ongoing maintenance is generating sustained demand for tradespeople: electricians, plumbers, steelworkers, construction workers, HVAC technicians, and network installers. These roles are not peripheral; they are central to whether AI factories can be built and operated at all. Compensation for these positions is rising materially, with skilled roles in leading-edge fab construction moving into six-figure territory.

This is not confined to the United States - sovereign AI initiatives worldwide are producing comparable domestic demand, from Germany to India to the Gulf states. The workforce impact spans the full educational spectrum. PhDs train frontier models. Software engineers build applications. Tradespeople build the physical infrastructure. Facilities managers, HVAC specialists, and grid engineers operate it. Each layer of the cake employs different skills, and every country building AI capacity will need to develop all of them. Our view is that the breadth of this job creation is underappreciated, and that workforce-development policy for the trades is as relevant to AI competitiveness as graduate-level research funding. Plus, the build-out value is ultimately a means to an end: the larger economic prize is the value derived from applying manufactured intelligence to the entire global economy, i.e. the productivity gains, new industries and compounding efficiencies that Jensen Huang describes when he says the application layer is where most of the economic value ultimately accrues.

Implications for Countries and Enterprises

AI is becoming critical infrastructure in the same sense that electricity and the internet are: every company will use it, and every country will want to build as much of it as possible. No country wants to import all its intelligence.

The implication for national strategy is that policy cannot focus solely on the model and application layers. Treating AI as national infrastructure means making energy capacity, semiconductor access, and data-center build-outs part of industrial strategy, alongside domestic model development and application ecosystems. Countries with strong industrial bases are particularly well-positioned for robotics and physical AI, where manufacturing expertise converges with AI to create opportunities that pure software economies cannot easily match.

For enterprises, the question is different but related: which layers do you participate in, which do you consume as a service, and where do your dependencies lie? Answering that question is now a core part of strategic planning rather than a technical detail.

For enterprises, the question is different but related: which layers do you participate in, which do you consume as a service, and where do your dependencies lie? Answering that question is now a core part of strategic planning rather than a technical detail

The role of institutional capital - pension funds, sovereign wealth, long-duration investors - in financing this build-out is one of the most important unresolved questions in global capital markets. The Davos exchange between Fink and Huang framed it directly: the relevant question is not whether this is a bubble, but whether current investment levels are sufficient to meet demand. We lean toward the latter framing, with a caveat: concentration risk among a small number of hyperscalers and sovereign anchors means any material policy shock, such as export controls, regulatory reversals, or capital-market contraction, could disrupt the sequencing of the build-out even if the underlying demand picture remains intact.

A final point, directed at the country that currently holds the strongest hand at every layer of the stack. The United States has built a genuine lead in AI across chip design, cloud infrastructure, foundation models, and applications. That lead is not guaranteed. Domestic regulatory uncertainty and skepticism about AI can slow adoption and chill investment. Inconsistent export policy risks pushing other countries toward non-U.S. technology stacks. The meaningful risk is not that AI develops too fast; it is that the United States creates the conditions for its own leadership to erode while others build.

Conclusion

The five-layer framework is a practical tool, not a theoretical construct. Understanding what AI actually is - five interdependent layers, with the largest economic payoff at the application layer, not just at chips or models - is the crucial first step to making sound decisions. Investors allocating capital, policymakers designing industrial strategy and enterprise leaders deciding where to participate and where to consume: all of them reason better about AI when they think about the whole stack. AI is not just a software phenomenon; it is an infrastructure build-out with physical, economic and workforce dimensions at every layer. Ultimately, it is confidence in the enormous value that manufactured intelligence will deliver to the global economy that makes this infrastructure investable and scalable.

For policymakers, investors, and enterprise leaders, the key takeaway is simple: think about the whole stack. The variables to watch over the next 24 months are the pace at which energy and water constraints get resolved, the progress of sovereign AI initiatives globally, and whether workforce-development programs for the trades keep up with demand. Critically, U.S. policy on AI development and technology exports must remain coherent and consistent, as regulatory unpredictability would be self-defeating for a country that currently leads at every layer. Ultimately, competitive position will be determined by execution - specifically, the speed to power, speed to silicon, and speed to deployment - making reference architectures a material consideration for enterprises and sovereigns evaluating where to start.

About this Brief and Futurum Research Cited

This market brief was produced by The Futurum Group and commissioned by NVIDIA. Market-sizing and capex figures are drawn from Futurum Intelligence research, including AI Capex 2026: The $690B Infrastructure Sprint and the Futurum Data Center Compute, AI Platforms, and Sovereign AI research programs. Views expressed are those of the author as an independent analyst.

Important Information About This Report

AUTHORS

Nick Patience
Vice President & Practice Lead,
AI Platforms | The Futurum Group

Daniel Newman
CEO | The Futurum Group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

NVIDIA

ABOUT NVIDIA

NVIDIA (NASDAQ: NVDA) is the world leader in AI and accelerated computing.

Futurum®

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.