SPECIAL COMPETITIVE
STUDIES PROJECT

arm

October 2025

Smarter at the Edge

How Edge Computing Can Advance U.S. AI Leadership and Energy Security

Table of Contents

Executive Summary

Frontier AI models may be trained in the cloud, but the future of AI will increasingly operate across both cloud and the edge — in billions of devices, local networks, and embedded systems that process intelligence close to where it's used. Current AI deployment patterns – characterized by massive, centralized data centers – present growing challenges in electricity consumption, water usage, and operational costs that could undermine the technology's transformative potential. In the United States, data centers already consume 4% of domestic electricity capacity and could reach 20-25% by 2030 based on the current trajectory.[1]

In the geopolitical context, concerns are already growing that the energy demands of compute could be a strategic bottleneck as the United States and China race to scale and deploy AI applications across their respective economies to spur economic growth and jumpstart the next generation of scientific and technological innovation. Governments and industry are not standing still. In the United States, the White House in July released its AI Action Plan, which set an ambitious agenda for the government to both accelerate AI innovation and streamline the AI infrastructure buildout process. Not to be outdone, in August, Beijing released details for its "AI+" initiative, an industrial plan to integrate AI across strategic sectors to boost productivity and growth. At the same time, companies on both sides of the Pacific are redesigning chips and moving away from general-purpose processors toward highly specialized AI accelerators tuned for inference – the stage where models meet their end-users. The aim is clear: to reduce energy and cost while expanding AI's reach.

Smarter, Not Bigger

The next wave of AI will not be defined by the largest models, but by the smartest and most efficient ones.

Edge computing – processing AI workloads closer to where data is generated – offers a pathway to resilient, energy-efficient AI. Advances in GPU efficiency and workload scheduling have significantly reduced the energy costs of inference in cloud environments, even as edge computing offers complementary gains for distributed applications. By combining specialized energy-efficient hardware with optimized software architectures, edge AI can reduce energy consumption by up to 60% for equivalent workloads while providing superior performance for latency and

[1] Matthew Connatser, Arm CEO warns AI's power appetite could devour 25% of US electricity by 2030, The Register (2024).

security-sensitive applications.[2] This results from both more efficient processing, often on battery-operated devices, as well as energy saved from moving data from devices over the network to data centers, and back.

Data centers will remain a foundational piece of AI infrastructure, but edge computing will likely be a complementary and necessary offset. Data centers will continue to train and run the most advanced frontier models, while edge deployments will increasingly handle everyday inferences. This hybrid approach will be the key to scaling AI without proportionally scaling its resource demands.

The U.S. government has a vital role to play in realizing this future. Market forces alone will not deliver the infrastructure, research, and incentives needed to ensure U.S. leadership in efficient AI. Industry innovation must be matched with public investment and smart policy. That means expanding funding for energy-efficient semiconductor research, establishing federal procurement preferences that reward efficient AI infrastructure, and creating government-sponsored testbeds to prove edge AI in high-value public missions. It also means continued investment in research and development (R&D) programs at the Department of Energy, National Science Foundation, and other agencies to ensure hardware, software, and model innovation move in lockstep. The next wave of AI will not be defined by the largest models, but by the smartest and most efficient ones, and the U.S. should lead in making that future a reality.

1. AI's Resource Challenge: An Unsustainable Trajectory

The Scale of Energy Demand

Data centers, the backbone of modern AI infrastructure, already account for roughly 4% of U.S. electricity consumption today.[3] Some individual data center facilities now consume more electricity than entire cities like Pittsburgh, Cleveland, or New Orleans.

Globally, AI-driven computing could nearly double the electricity consumption of data centers from 536 terawatt-hours (TWh) in 2025 to over 1,065 TWh by 2030.[4] These

[2] AWS Graviton4: Revolutionizing EC2 Performance with Best-in-Class Price-Performance and Energy Efficiency, Amazon Web Services (2024).
[3] Arman Shehabi, et al., 2024 United States Data Center Energy Usage Report, Berkeley Lab (2024). .
[4] Karthik Ramachandran, et al., As generative AI asks for more power, data centers seek more reliable, cleaner energy solutions, Deloitte (2024).

figures rival the total power usage of mid-sized nations, placing unprecedented stress on power grids and increasing operational costs for all consumers.

The Inference Revolution

The nature of AI's energy demands has undergone a fundamental transformation. While training large models like OpenAI's ChatGPT-4 represented large but one-time training events, the focus has now shifted to inference – the process of generating outputs for end user queries. A single ChatGPT query consumes approximately 0.3 watt-hours, roughly ten times the power of a standard search engine query.[5]

Inference is the New Frontier

More than 75% of computational demand in the U.S. will soon be for inference – a shift that creates long-term energy challenges for utilities and regulators.

Analysts estimate that more than 75% of computational demand in the United States will be for inference in the coming years, a figure that is already 100 times greater than it was at the dawn of large language models (LLMs).[6] This transition from intermittent training demand to continuous inference creates systemic, long-term challenges for utilities and regulators.

The Efficiency Paradox

Even as the industry achieves efficiency gains in hardware and software, these improvements are often reinvested to fuel more compute power and stimulate greater demand. Bloomberg Intelligence forecasts a 4x increase in AI energy demand by 2032, even accounting for efficiency improvements.[7] This creates a self-reinforcing cycle where technological progress drives net increases in resource consumption despite individual task optimization.

Meeting this challenge will require more than incremental private sector innovation. The scale of AI's resource footprint demands national-level infrastructure and strategy, on par with past U.S. efforts in nuclear security and supercomputing.

---

[5] Jasmine Capen, What's the Shocking Truth Behind AI Energy Consumption?, Integrity Energy (2025).
[6] Emma Sheppard, 'The amount of inference compute needed is already 100x more': How Europe's AI companies can meet demands, Sifted (2025),
[7] Mark MacCarthy & David M. Klaus, Why AI demand for energy will continue to increase, The Brookings Institution (2025).

2. Edge Computing: A Strategic Solution

Redefining AI Architecture

From a silicon perspective, two activities consume the bulk of the power – computation and data movement. Therefore, optimizing communication between compute components is critical to efficient AI in silicon.

Efficiency at the Edge

Edge computing offers a pathway to more resilient, energy-efficient AI – reducing energy use by up to 60% for equivalent workloads.

In this context, edge computing represents a fundamental architectural shift: Embedding machine learning (ML) capabilities directly into devices, such as smart cameras, industrial sensors, automotive systems, and mobile devices. This enables real-time processing without constant internet connectivity, reducing reliance on centralized cloud infrastructure. This distributed approach creates synergistic benefits that go beyond energy savings, encompassing operational cost reduction, performance gains, and enhanced security.

Energy savings directly translate to reduced operating expenses. For example, manufacturing facilities implementing edge-based process optimization have reported 30-50% reductions in energy use compared to traditional methods.[8]

But beyond that, edge computing is transforming almost every sector of innovation:

- **Automotive:** By bringing intelligence on-board, edge AI can deliver split-second perception and hazard detection – a capability central to the next generation of autonomous mobility.
- **Industrial Automation:** Edge-specific AI chips allow sub-millisecond response times for factory robotics, where cloud latency would mean system failure.
- **Healthcare Devices:** Smartwatches and medical wearables now run AI locally to detect arrhythmias, sleep apnea, or early signs of hypertension.
- **Mobile & PCs:** Smartphones can now translate languages instantly without needing the internet, and new laptops come with AI features that make everyday

[8] Emily Khym & Nayella Vasquez, The Power of AI in Clean Energy: Transforming Sustainability for the Future, Yale Clean Energy Forum (2025).

tasks like work, creativity, and online safety faster, more personalized, and more secure.

- **IoT & Smart Homes:** Edge AI in cameras, sensors, and smart speakers reduces response times, optimizes energy use, and avoids transmitting sensitive data.

The Edge AI Evolution Matrix (2025–2029)

Key trends indicate a shift away from basic data processing at the edge – such as filtering and cleaning – to **advanced, real-time AI-driven analysis and decision-making,** fueling new use cases and business models:

Year / Phase	Key Capabilities	Example Applications
2025	Predictive maintenance, localized rule automation, real-time monitoring	Manufacturing anomaly detection, smart grid management, remote patient monitoring
2026–2027	On-device generative AI, adaptive ML models that learn locally, privacy-preserving analytics, seamless device–cloud handoffs	Smart assistants that improve over time, edge healthcare diagnostics, secure enterprise data analysis
2028–2029	Multi-agent edge environments, hyperlocal automation, agentic AI collaboration, energy optimization, autonomous systems	Fleets of vehicles, smart city infrastructure, distributed energy markets

Source: Arm research

Alongside the expansion of application possibilities, edge computing enhances privacy and security. That's because local inference keeps sensitive data – such as health metrics or proprietary enterprise data – on the device, reducing exposure to cyber risk. Additionally, users can now create text, images, or music directly on the device, avoiding both cloud dependency and potential data leakage.

Overcoming Challenges to Scale

For all its promise, edge computing faces a significant scale challenge: Building out edge infrastructure at the scale of today's hyperscale data centers is complex. But progress is rapid. Hardware innovation for CPUs, GPUs, and AI accelerators ensures the right processor is used for the right workload. Advances in memory systems and dataflow designs help reduce the cost of moving data within chips.

Frontier to Edge in 18 Months

In as little as 12-18 months, models once considered frontier are moving to the edge.

Smaller, more efficient AI models – known as small language models (SLMs) – are also emerging. Slimmed-down models optimized for devices with limited power are increasingly performing cutting-edge tasks; in as little as 12-18 months, models that were once considered frontier are moving to the edge.[9]

Innovations on the software side are squeezing out more performance per watt by optimizing workloads across different types of hardware, compressing models, and intelligently orchestrating tasks from cloud to edge. This ensures that efficiency gains translate into real-world energy savings.

The Complementary Ecosystem

Edge computing will not replace data centers; it will complement them. Data centers will remain vital for training and hosting frontier AI, but much of the inference will migrate to the edge, just as past supercomputer capabilities now fit into smartphones. The cloud will remain the backbone for massive model training and resource-intensive development. However, edge devices will increasingly handle the bulk of latency-sensitive, real-world inference tasks. McKinsey forecasts that up to 70% of AI inference will occur at the edge by 2030.[10] This hybrid, multi-tiered architecture enables greater AI capability without proportional growth in resource consumption, supporting both business performance and energy efficiency goals.

[9] Ylli Bajraktari, et al., AGI Will Arrive in Three Ways, Special Competitive Studies Project (2024).
[10] Rishabh Mishra, Verizon Launches AI Connect as McKinsey Predicts That 60-70% of AI Workload Will Shift From Training to Inferencing By 2030, Benzinga (2025).

Edge AI Market Expansion by Sector

The global edge AI market is expected to grow rapidly, with estimates ranging from $20–26 billion in 2024 to over $66 billion by 2030, and a compound annual growth rate (CAGR) between 20 to 37% depending on the segment.[11] Application domains will expand from today's pilots to mission-critical deployments, including:

Sector	Edge AI Applications
Utilities	Grid balancing, outage prediction, real-time energy trading
Healthcare	On-device diagnostics, privacy-first patient monitoring
Industrial IoT	Process automation, quality control, safety monitoring
Consumer Devices	AI-powered smartphones, smart home automation

Source: Arm research

3. From Energy Consumer to Optimizer

AI is not just a consumer of energy, but it can also be a tool for managing it. Smarter, more efficient AI accelerates adoption of AI in energy optimization tasks, from balancing power grids and data center capacity to improving industrial processes.

AI-powered "smart grids" use predictive analytics and machine learning to optimize power generation, distribution, and consumption across electrical networks.[12] These systems predict energy demand with unprecedented accuracy, optimize power routing to minimize transmission losses, and better integrate variable renewable energy sources.

The concept of "data center flexibility" transforms AI infrastructure from passive energy consumers into active grid management tools.[13] AI workloads can be automatically moved to regions experiencing low grid demand or high renewable energy generation. Non-urgent AI training can be scheduled during off-peak hours when excess generation capacity is available.

[11] Edge AI Software Market, Markets and Markets (2025) and Edge AI Market 2025-2030, Grand View Research (2024).
[12] M Balamurugan, et al., Role of artificial intelligence in smart grid - a mini review, PubMed Central (2025).
[13] EPRI, Epoch AI Joint Report Finds Surging Power Demand from AI Model Training, PR Newswire (2025).

Beyond optimizing its own energy consumption, AI enables significant efficiency improvements across industrial processes. AI-enabled process optimization in manufacturing can reduce energy consumption and waste by 30-50% compared to traditional methods.[14] AI-driven efficiencies in daily freight operations, better capacity use, and shifts to lower carbon transport modes, could cut freight logistics emissions by up to 10-15% at full scale.[15]

4. Enabling Technologies and Innovations

Heterogeneous Computing

Computing power used to advance by cramming ever more transistors into smaller chips. But that 'brute force' approach is no longer cost effective.

No single chip can efficiently handle the diversity of AI tasks. AI depends on a range of computing technologies working together. Instead, systems rely on *heterogeneous computing* – the coordinated use of different types of chips, often in different types of devices, to optimize performance, cost, and energy use.

CPUs act as the general-purpose managers, directing overall system activity and running everyday software. GPUs excel at crunching massive amounts of data in parallel, such as training large language models. NPUs and other AI accelerators are specialized for highly compute-intensive AI tasks and workloads, like facial recognition or voice assistants, delivering faster results with less power. Finally, Application-specific Integrated Circuits (ASICs) are custom chips built for one purpose – such as processing payments or securing communications – where efficiency and speed are critical. Using the right chip for the right job makes AI systems faster, cheaper, and more resilient.

New approaches like modular chip designs ('chiplets') are emerging to cut power consumption by keeping computation and memory tightly integrated. Instead of building ever-larger monolithic processors, these small, specialized components can be combined like building blocks, minimizing the energy-intensive process of moving data across distant parts of a system.

---

[14] Emily Khym & Mayella Vasquez, The Power of AI in Clean Energy: Transforming Sustainability for the Future, Yale Clean Energy Forum (2025).
[15] Intelligent Transport, Greener Future: AI as a Catalyst to Decarbonize Global Logistics, World Economic Forum in collaboration with McKinsey & Company (2025).

Software Optimization for Edge Deployment

AI is moving from large, resource-intensive models to smaller, more efficient ones designed for specific tasks – the SLMs mentioned earlier. SLMs are streamlined AI systems that can run on limited hardware, making them well-suited for "edge" devices like smartphones, cars, or medical sensors. By optimizing software, SLMs deliver useful AI capabilities without requiring the massive computing power of traditional large models.

Several techniques make this possible. Quantization reduces the memory needed to run a model – in some cases by up to 80 percent – so they can fit on a smartphone.[16] Pruning removes unnecessary parts of the model, making it smaller and faster, like trimming excess code. Knowledge distillation allows a large AI system to "teach" a smaller one to perform the same task with fewer resources – useful for things like chatbots or customer service tools. Federated learning trains models directly on edge devices, such as smartwatches or home assistants, reducing reliance on cloud networks and improving data privacy.

Recently developed SLMs are now capable of running directly on edge devices – like laptops, smartphones, IoT systems – without requiring cloud infrastructure. Examples include OpenAI's open-source gpt-oss-20b, Microsoft's ultra-low-latency Phi-4-mini-flash-reasoning, and compact models like Phi-3-mini, TinyLlama, and MobiLlama. These models demonstrate that AI innovation is increasingly decentralized, efficient, and accessible – shifting computing capability closer to where data is generated.

These advances enable AI to operate closer to where data is generated – on the edge – bringing benefits in speed, security, and accessibility for real-world applications.

5. Government Actions to Date: Successes and Gaps

Industry innovation has led the way in making AI more energy-efficient, but the U.S. government took important steps of its own. These efforts demonstrate recognition that AI's energy challenge is not just a commercial issue, but a matter of national infrastructure, competitiveness, and security. At the same time, promising federal initiatives can be underfunded or fragmented across agencies.

[16] Emmett Fear, AI Model Quantization: Reducing Memory Usage Without Sacrificing Performance, Runpod (2025).

Successes:

White House AI Action Plan: The ambitious White House plan sets an important policy demand signal and represents the Administration's first coordinated effort to frame artificial intelligence as a strategic priority. The plan directs agencies to prioritize AI R&D, accelerate the buildout of domestic AI infrastructure, and forge public-private partnerships to export an American AI stack to overseas markets, while strengthening measures to counter China's AI development.

Department of Energy (DOE) Research: DOE has supported energy-efficient computing through its Advanced Scientific Computing Research program and new initiatives like the Frontiers in Artificial Intelligence for Security, Science, and Technology (FASST) program. These efforts link AI efficiency directly to national energy priorities. Replicating the success of national labs with high-performance computing for AI with efficiency and resilience in mind is critical.

NSF's National AI Research Resource (NAIRR): The NAIRR pilot is designed to democratize access to compute, data, and tools for researchers, with efficiency built in as a guiding principle. Though still modest in scale, it represents a model for shared, public-facing AI infrastructure.

CHIPS and Science Act: This landmark legislation authorized funding to expand domestic semiconductor R&D and manufacturing, including advanced packaging, heterogeneous compute, and microelectronics research centers. It incentivized critical infrastructure to be built in the U.S., creating a foundation for more efficient architectures.

Early Testbeds: Agencies such as Department of Homeland Security (DHS), Department of Defense (DoD), and DOE have launched pilot programs using AI to optimize grids, monitor critical infrastructure, and enhance emergency response. These experiments point toward the potential of government-led edge AI deployments in mission-critical domains.

Gaps:

Implementation Without Investment: Many CHIPS Act provisions for microelectronics research centers and advanced efficiency programs remain

either unfunded or funded at a fraction of what Congress authorized. This gap risks stalling momentum at a critical moment. Relatedly, while the White House AI Action Plan lays out an essential policy framework, it lacks sustained funding and the institutional mechanisms needed to translate ambition into execution.

AI R&D Spending Gap: The United States currently invests roughly $3 billion annually in AI R&D, far short of the $32 billion target set by the National Security Commission on AI and reinforced by the Special Competitive Studies Project (SCSP).[17] Without a significant increase, federal R&D will remain too limited to match the scale of private sector advances.

Procurement Practices: Federal procurement has not yet been leveraged as a strategic tool to create demand for efficient AI infrastructure. Current purchasing patterns risks locking agencies into less efficient and interoperable architectures for years to come.

The successes show that government action can meaningfully shape AI's energy trajectory, but the gaps make clear that current efforts are insufficient when the stakes are so high. Industry is innovating at pace, but without scaled and coordinated federal action, the U.S. could face growing energy challenges and risks to its technological leadership. To secure leadership, government must build on early wins, close funding gaps, and set clear policy signals that efficiency is a national priority.

6. China's Strategic Play for Energy Efficient AI: Policy, Infrastructure, and Capital

The United States is not competing in a vacuum. Beijing is already preparing for the next phase of AI competition — not just who builds the largest models, but who can deploy and scale them most efficiently across society. Edge computing sits at the heart of this strategy. China is executing coordinated policy, infrastructure, and financial initiatives that together reveal a deliberate plan

The Efficiency Race

Beijing is already preparing for the next phase of AI competition – not who builds the largest models, but who can deploy them most efficiently.

[17] Final Report, National Security Commission on Artificial Intelligence (2021); Nyah Stewart, Funding for the Future: the Case for Federal R&D Spending, Special Competitive Studies Project (2024).

to field energy-efficient, distributed AI systems at a national scale.

Policy Planning: In August, China launched its "AI+" initiative – a nation-wide push to embed AI across strategic sectors such as energy, transportation, and manufacturing.[18] In practice, the success of "AI+" will likely depend in large part on the deployment of edge computing technologies. Many of the envisioned applications, from factory automation to smart grids, to autonomous vehicles and precision agriculture, will require processing vast amounts of data close to where it's generated to reduce latency, preserve bandwidth, and ensure data security. As Beijing pushes local governments and firms to deploy AI in real-world industrial settings, investments in edge infrastructure, such as specialized chips, local data centers, and industrial Internet-of-Things (IoT) systems, will become a practical prerequisite for turning "AI+" from policy into reality.

Infrastructure Build-out: Beijing plans to deploy AI and edge computing, however, not at the expense of data centers and cloud computing. Since 2022, Beijing's *Eastern Data, Western Computing* program has directed industry to build a geographically distributed compute ecosystem that eases pressure on local grids and resources.[19] Large, energy-hungry data centers are clustered in renewable-rich western provinces such as Guizhou and Inner Mongolia, while edge and regional centers near major cities handle real-time inference and data processing. This tiered architecture embodies edge-computing principles — processing data close to its source and reserving remote facilities for large-scale training — and has pushed China's AI and cloud infrastructure past 750 EFLOPS of capacity, achieved through a hierarchical network spanning frontier clusters to billions of edge nodes.[20]

Capital Mobilization: To finance this shift, Beijing in March 2025 created a $138 billion National Venture Capital Guidance Fund for AI, semiconductors, and clean-energy technologies structured as an industry-government partnership.[21] While details of how the fund will work remain murky, it will likely make investments in advance efficiently, distributed computing in line with higher-level national initiatives and objectives. By aligning public and private investment across the AI-

[18] Kendra Schaefer, "The AI Plus Initiative - China's Plan for AI Diffusion," Trivium China (2025).
[19] Sunny Grimm, "China invested $6.1 billion in state data center project in two years – the 'Eastern Data, Western Computing' project aims to utilize the country's undeveloped land," Tom's Hardware (2024).
[20] China's AI Infrastructure Surge, Strider and Special Competitive Studies Project (2025).
[21] Nectar Gan and Juliana Liu, "China announced high-tech fund to grow AI, emerging industries," CNN (2025).

energy value chain, China aims to lower the energy cost of computation while scaling a nationwide edge-to-cloud ecosystem — transforming efficiency itself into a competitive advantage in the global race for sustainable AI.

These programs underscore a global shift toward coupling AI competitiveness with energy sustainability. The risk is clear: without a deliberate push to lead in efficient AI, the U.S. could win the innovation sprint but lose the efficiency race, leaving American infrastructure strained and competitors setting the rules of the road.

7. Policy Recommendations for Efficient AI Leadership

1. Expand Investment in Efficient Infrastructure and R&D

Building on the CHIPS and Science Act: The CHIPS and Science Act established a strong foundation for U.S. semiconductor innovation. Even as federal priorities shift, policymakers should sustain and expand targeted funding for the next frontier of innovation. This includes prioritizing architectures and design approaches that deliver greater performance per watt, expanding heterogeneous computing research to ensure the right processor is matched to the right workload, and accelerating advanced packaging technologies to reduce the energy cost of moving data.

DOE Energy Efficiency and AI Infrastructure Initiatives: DOE plays a central role in advancing efficient AI computing. Its Energy Efficiency Scaling for Two Decades (EES2) initiative aims to increase semiconductor energy efficiency a thousand-fold over twenty years, an essential step toward managing AI's growing power demands affordably and reliably.[22] This effort aligns with the AI Science Cloud, a national-scale, interoperable AI infrastructure being developed under the One Big Beautiful Bill Act. While the AI Science Cloud will primarily serve frontier-scale computing, its focus on interoperability, modular design, and avoiding vendor lock-in supports DOE's broader goal of sustainable, efficient computing. Together, these two programs can help build momentum for U.S. leadership in efficient AI infrastructure, driving innovation at scale while ensuring energy efficiency remains a core design principle.

[22] DOE EES2 Pledge, U.S. Department of Energy (2024).

2. Incentivize Efficient AI Through Procurement and Competitions

Federal Procurement Preferences: Government procurement policies should establish preferences for infrastructure-grade AI computing that minimizes lifecycle energy costs. This approach leverages the government's substantial purchasing power to create market incentives for efficiency, while reducing federal operational costs.

Efficiency Competitions: Government-sponsored competitions should challenge developers to create the most efficient, application-specific AI models for federal use cases, accelerating innovation while identifying best practices for broader adoption.

3. Develop Public-Sector Edge AI Testbeds

Targeted Application Testbeds: Government-sponsored testbeds should focus on high-value applications where edge AI can provide clear public benefits.[23] Priority areas include wildfire monitoring, critical infrastructure protection, emergency response systems, and resource monitoring.

Interagency Collaboration: Testbed development should involve collaboration between the Department of Energy, National Science Foundation, Department of Defense, Department of Homeland Security, and other relevant agencies to ensure comprehensive evaluation of edge AI capabilities across government missions.

Public-Private Partnerships: Testbed operations should include partnerships with private sector technology providers to accelerate development and ensure practical applicability of research results.

[23] AI R&D Testbed Inventory, The Networking and Information Technology Research and Development (last accessed 2025).

8. Industry Action Items

1. Hardware-Software Co-Design

Companies should continue investing in integrated development approaches that optimize entire systems rather than individual components, generating the greatest efficiency improvements and competitive advantages.

2. Best Practices Sharing

Industry associations should facilitate sharing of efficiency and interoperability best practices and benchmarking data to accelerate adoption of optimal approaches across the sector, while maintaining security standards.

3. Policy Alignment

Product roadmaps should consider policy incentives and regulatory trends to ensure that innovation investments align with long-term market conditions and government priorities.

9. Conclusion: Enabling Resilient AI Leadership

The challenge of AI's resource demands represents a critical test of America's ability to lead in transformative technologies in a manner that is economically sustainable. Edge computing, powered by energy-efficient hardware and optimized software, offers a pathway to resilient AI deployment that enhances rather than constrains the technology's transformative potential.

Where Innovation Meets Policy

Energy-efficient AI will emerge where hardware and software innovation intersect with clear policy signals.

Edge and cloud are complementary, not substitutes. The future of AI deployment requires a strategic balance between centralized cloud computing and distributed edge processing. This multi-tiered approach enables greater AI capability without proportionally greater resource consumption by intelligently distributing workloads across the most appropriate computing environments.

Ultimately, energy-efficient AI will emerge where hardware and software innovation intersect with clear policy signals. The window for proactive action is limited but still open. To realize this future, government and industry must work together, investing in the technologies and policies that make efficient AI not only possible, but inevitable. The U.S. has begun to lay the groundwork for this future, but without sustained investment, interoperable infrastructure, and strong governance, they risk remaining pilot projects. The policies and investments made today will determine whether America leads in resilient, innovative AI technologies or struggles to manage the consequences of less efficient approaches.

By pursuing these recommendations in concert, policymakers and industry leaders can ensure that AI's next chapter is defined not by runaway resource demands, but by intelligent, cost-and-resource-efficient innovation that strengthens both economic competitiveness and technological leadership.
