Scientific Research Publishing

Energy and Power Engineering, 2026, 18(4), 168-202
https://www.scirp.org/journal/epe
ISSN Online: 1947-3818
ISSN Print: 1949-243X

Artificial Intelligence and Energy Overconsumption: Data Center Electricity Demand, Cooling Burdens, and Regional Sustainability Constraints

David Crovato[1], Jose Rosania[2]

[1]Department of International Business Administration, Erasmus University Rotterdam, Rotterdam, Netherlands
[2]Department of Electrical Engineering, Florida State University, Panama City, Panama
Email: joserosania@icloud.com

How to cite this paper: Crovato, D. and Rosania, J. (2026) Artificial Intelligence and Energy Overconsumption: Data Center Electricity Demand, Cooling Burdens, and Regional Sustainability Constraints. *Energy and Power Engineering*, **18**, 168-202.
https://doi.org/10.4236/epe.2026.184010

Received: February 24, 2026
Accepted: April 10, 2026
Published: April 23, 2026

Abstract

Artificial intelligence is increasingly embedded in everyday life, but its expansion depends on physical infrastructure with significant electricity, cooling, and water requirements. In this paper, "overconsumption" refers not simply to high electricity use, but to levels of AI-related infrastructure demand that intensify grid strain, raise carbon emissions where fossil generation remains dominant, increase cooling and water burdens, or displace other uses of limited energy resources. Rather than treating AI as a purely digital technology, this study examines it as a material system shaped by energy, environmental, and regional constraints. Using secondary sources from peer-reviewed research, international energy institutions, government reports, and industry analyses, the paper distinguishes between total data center electricity demand and the share specifically linked to AI workloads. It compares measured observations with scenario-based forecasts to examine how AI-driven expansion may alter electricity demand across regions. Particular attention is given to cooling intensity, water use, climate conditions, and grid limitations. The analysis finds that AI-related data center growth is becoming an increasingly important driver of electricity demand, especially in regions with high cooling requirements, limited grid capacity, or water stress. It also finds that sustainability outcomes differ sharply across regions, meaning that the environmental burden of AI expansion is unevenly distributed. While improvements in hardware efficiency, cooling systems, and renewable integration may reduce some pressures, continued AI growth without coordinated infrastructure planning risks worsening resource strain and reinforcing regional inequalities.

Keywords

Artificial Intelligence, Energy Consumption, Data Centers, Electricity Demand, Cooling Systems, Energy Infrastructure, Grid Capacity, Renewable Energy Integration, Sustainability, Carbon Emissions, Energy Policy, Resource Constraints, AI Scalability, Power

1. Introduction

Artificial intelligence is often described as virtual technology, yet its operation depends on highly physical systems. Large AI models are trained and deployed in data centers filled with specialized processors and cooling infrastructure that must operate continuously. As AI becomes more integrated into digital platforms, the energy required to support these systems has grown rapidly, raising concerns about sustainability and infrastructure capacity [1].

This paper uses the term "overconsumption" to describe levels of AI-related energy and resource use that create disproportionate strain on electricity systems, cooling capacity, water resources, or surrounding communities. The issue is therefore not simply that AI consumes electricity, but that its scale of growth may exceed the ability of existing systems to support it sustainably.

A key distinction is necessary. Not all data centers for electricity use are attributable to artificial intelligence. Data centers also support cloud services, storage, and general computing. However, AI-focused facilities rely on high-density GPU clusters and generate significantly greater heat, making them more energy-intensive than traditional systems [2]. As a result, AI is becoming a major contributor to the growth of total data center electricity demand, even though the two should not be treated as interchangeable.

This paper examines how AI-driven data center expansion is affecting electricity demand and infrastructure strain. It compares measured electricity use with projected growth and analyzes how cooling requirements, regional climate, water availability, and grid capacity shape sustainability outcomes. The central argument is that artificial intelligence should be understood as energy-intensive infrastructure embedded within real physical constraints.

2. Literature Review

2.1. Global Data Center Energy Demand Trends

Recent research shows that global data center electricity demand is entering a period of rapid growth, largely driven by AI workloads. Projections suggest that total data center demand could nearly double by 2030 (projection) [3]. However, these figures represent total data center consumption and not exclusively AI-related demand.

Regional projections highlight this growth. North America may exceed 386 TWh by 2025 and reach around 755 TWh by 2030 (projection) [3], while Europe is also expected to grow significantly despite infrastructure and regulatory constraints. In the United States, data center demand could reach approximately 1050 TWh annually by 2030 (projection), with around 56% of electricity still generated from fossil fuels (measured) [4]. Researchers warn that meeting this demand is likely to involve expanded natural gas generation, raising concerns about increased carbon emissions and long-term climate impacts. Figure 1, titled "Projected Data Center Electricity Growth," compares the projected growth in data center electricity consumption between the US and Europe, illustrating how both regions are expected to expand significantly.

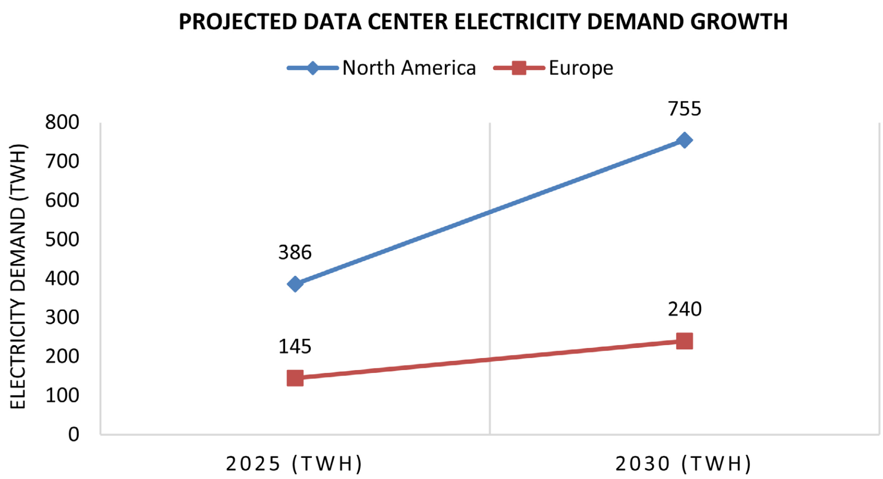

Figure 1. Adapted from S & P Global. (2025). Global data center power demand expected to almost double by 2030.
https://www.spglobal.com/energy/en/news-research/latest-news/electric-power/110525-global-data-center-power-demand-expected-to-almost-double-by-2030

2.2. Difference between "Normal" Data Center Loads vs. AI-Driven Workloads

The literature makes a clear distinction between traditional data centers and AI-focused facilities. Traditional centers mainly support cloud services and business applications, typically relying on CPUs and operating at relatively low rack power densities. In contrast, AI data centers are specifically built to train and run large-scale machine learning models. They depend on specialized hardware such as GPUs, high-speed networking, and advanced storage systems [2]. As a result, their rack power densities are much higher, often exceeding 30 kilowatts compared to roughly 8 kilowatts in conventional facilities.

AI workloads also bring fundamentally different energy and infrastructure demands. Training and real-time inference require continuous, high-intensity computation, leading to more concentrated electricity loads and significantly greater cooling needs. In addition, AI processing can cause rapid and less predictable fluctuations in power demand, creating added pressure on grid stability and power quality. Because of these characteristics, AI-driven facilities contribute disproportionately to overall data center energy growth and introduce new operational challenges for electricity systems.

2.3. Evidence of Cooling

Research increasingly identifies cooling and water use as major sustainability concerns linked to data center expansion, especially in AI-focused facilities [5]. Large data centers can consume up to 5 million gallons of water per day for cooling, an amount comparable to the daily water use of small towns. Across the United States, total data center water consumption reaches hundreds of millions of gallons per day, both through direct cooling and indirectly through electricity generation.

Water use has also risen sharply over time. Direct consumption by U.S. data centers increased from about 21 billion liters in 2014 to roughly 66 billion liters in 2023, reflecting the rapid growth in facility size and number [6]. AI systems intensify this trend, as their higher heat loads often require liquid-based cooling solutions. Although these systems can be more energy-efficient than traditional air cooling, they typically lead to higher freshwater losses, creating a trade-off between energy savings and water sustainability. This challenge has encouraged the development of alternatives such as direct-to-chip and immersion cooling, but these technologies are not yet widely adopted.

2.4. Regional Sustainability Concerns

The literature highlights that the sustainability of AI-driven data center expansion differs significantly across regions. Survey-based studies show clear variations in how experts perceive AI's environmental impact [7]. Respondents in Asia-Pacific and Latin America tend to express more optimism, while those in Europe, North America, and particularly Africa and the Middle East are more cautious. These differences are often linked to variations in infrastructure readiness, climate conditions, investment capacity, and governance frameworks.

Regional conditions also shape practical constraints. In the Middle East, extreme heat and water scarcity intensify cooling and resource pressures. In Europe, limited grid access and regulatory complexity slow expansion. In many developing regions, unreliable electricity infrastructure restricts the deployment of energy-intensive AI facilities. Overall, the literature suggests that AI infrastructure growth is uneven and closely tied to local energy and resource conditions, raising concerns about unequal access to AI's economic benefits.

2.5. Gaps in Research

Despite growing attention to the energy and environmental impacts of artificial intelligence, several important gaps remain in the existing literature.

One major limitation is the lack of transparency in reported data. Many technology companies do not separate AI-specific workloads from general cloud or data center operations in their disclosures, making it difficult to accurately quantify the electricity and water consumption attributable specifically to AI systems

[8]. As a result, much of the current research relies on indirect estimates or aggregated data, which reduces comparability across studies and limits the precision of conclusions.

A second gap is the limited distinction between measured observations and scenario-based projections. While many studies provide forecasts of future data center electricity demand, fewer clearly differentiate between current, empirically measured consumption and estimates based on assumptions about future AI adoption, hardware efficiency, and infrastructure expansion. This lack of consistency makes it difficult to evaluate how rapidly AI is currently contributing to energy demand versus how much impact is expected under different growth scenarios.

In addition, much of the literature focuses primarily on operational electricity use while giving less attention to full lifecycle impacts. The production, transportation, and disposal of hardware such as GPUs, servers, and cooling systems require significant materials and energy, yet these upstream and downstream effects are often underrepresented in analyses of AI sustainability [8]. Similarly, electronic waste generated by frequent hardware upgrades remains insufficiently studied.

Water consumption is another underexplored area. Although some studies estimate cooling-related water use, many rely on indirect calculations rather than direct facility-level measurements, and there is limited research on how data center water use affects local communities, particularly in water-stressed regions [5]. This gap is important because water availability can be a critical constraint on data center expansion.

There is also relatively limited research comparing AI-specific energy demand across regions using consistent criteria. While many studies analyze individual countries or regions, fewer provide systematic cross-regional comparisons that account for differences in grid capacity, climate conditions, water availability, and policy environments. This makes it more difficult to understand how the sustainability of AI infrastructure varies globally.

Finally, policy and regulatory responses to AI-related energy demand remain underdeveloped in the literature. While there is substantial discussion of general energy efficiency and renewable energy policies, fewer studies examine targeted strategies for managing the resource impacts of AI infrastructure, such as location-based restrictions, cooling standards, or reporting requirements.

Together, these gaps highlight the need for more transparent data, clearer methodological distinctions, and more integrated research that connects technical, environmental, and policy perspectives in order to better assess the sustainability of AI-driven data center expansion.

3. Methodology

This paper uses a secondary research methodology based on a comparative literature review and cross-source synthesis. No original experimental or survey data were collected. Instead, the analysis draws on four main categories of sources:

peer-reviewed academic studies, reports from international and governmental institutions, industry and market analyses, and public sustainability disclosures from major technology firms where relevant.

The review process focused on sources addressing one or more of the following topics: data center electricity demand, AI-specific computing loads, cooling energy use, water consumption, regional grid constraints, and policy responses related to energy efficiency or infrastructure planning. Search terms included combinations of phrases such as "AI data center electricity demand," "AI training energy use," "data center cooling electricity," "data center water consumption," "grid constraints and hyperscale data centers," and "AI infrastructure sustainability." Priority was given to recent sources, especially those published between 2023 and 2025, because of the rapid pace of change in AI infrastructure and energy demand.

Sources were selected based on relevance, recency, credibility, and the extent to which they provided either quantitative estimates or policy and regional context. Where multiple estimates conflicted, the paper compared them rather than treating them as interchangeable. In particular, it distinguishes between measured observations, such as current electricity consumption, and scenario-based projections, which depend on assumptions about future AI adoption, efficiency improvements, and infrastructure growth.

The analysis follows a comparative structure in two ways. First, it distinguishes total data center electricity demand from electricity demand attributable specifically to AI workloads whenever the underlying source makes that separation possible. Second, it compares regional conditions across the United States, Europe, the Middle East, and developing economies using consistent criteria: grid capacity, cooling burden, water stress, and policy or regulatory constraints.

This approach has limitations. Many firms report aggregated data that do not isolate AI workloads from general cloud or data center operations. In addition, some market forecasts rely on assumptions that may change quickly as hardware efficiency, model architecture, and deployment patterns evolve. Despite these limitations, secondary comparative analysis remains an effective method for identifying large-scale trends and evaluating how AI expansion interacts with energy and environmental systems across regions.

4. Core Analytical Body

4.1. GPU vs CPU Power Draw

Central Processing Units (CPUs) are designed for general-purpose computing and are used to run operating systems, applications, databases, and other logic-based tasks [9]. Their architecture is optimized for sequential processing, meaning instructions are executed step by step. CPUs usually have a small number of powerful cores, high clock speeds, and layered cache memory that helps reduce latency. They also include energy management features such as dynamic voltage and frequency scaling, which help balance performance with relatively modest power consumption. Because of this, CPUs generally produce less heat and are well-suited for devices with thermal or battery constraints.

However, AI training requires massive parallel computation. Large models perform the same mathematical operations across millions of data points at the same time, which is not well-suited to CPU architecture.

Graphics Processing Units (GPUs), originally developed for rendering graphics, are designed for parallel processing [9]. Instead of a few powerful cores, GPUs contain thousands of smaller cores that can run many operations simultaneously. This structure allows them to handle matrix multiplications and large-scale data processing far more efficiently than CPUs. Modern GPUs also include specialized components such as Tensor Cores that are optimized for AI workloads.

	CPU (Central Processing Unit)	GPU (Graphics Processing Unit)
Primary function	General-purpose processing for sequential tasks	Specialized for parallel processing and data-intensive tasks
Architecture	Fewer, powerful cores optimized for single-threaded performance	Thousands of smaller, simpler cores optimized for parallelism
Processing model	Serial execution: tasks processed one at a time	Parallel execution: multiple tasks processed simultaneously
Core count	Typically 4 - 64 cores in consumer-grade CPUs	Can have thousands of cores in high-performance GPUs
Clock speed	Higher clock speeds (up to ~5 GHz)	Lower clock speeds (~1 - 2 GHz)
Strengths	Precision, sequential tasks, versatility, and logic operations	High throughput for large-scale operations like matrix math
Use cases	Running operating systems, application logic, databases	Graphics rendering, machine learning, scientific computing
Power consumption	Lower due to fewer cores and energy-efficient designs	Higher due to dense cores and memory bandwidth demands
Memory bandwidth	Lower, typically optimized for latency	Higher, optimized for throughput (e.g., GDDR6, HBM memory)
Cost	Relatively affordable and widely available	More expensive, especially for high-performance models
Applications	Laptops, desktops, servers, mobile devices	Gaming systems, workstations, HPC environments, AI workloads
Flexibility	Broad compatibility for diverse tasks	Optimized for specific workloads requiring parallelism

This performance advantage comes with higher power consumption. GPUs rely on dense core structures, continuous parallel activity, and high bandwidth memory systems such as GDDR6 or HBM [10]. A high end GPU can consume several hundred watts under full load, especially during AI model training. Although GPUs often provide better performance per watt for parallel tasks, their overall electricity demand remains high because they operate near maximum capacity for long periods. As a result, large GPU clusters have become a major contributor to the rising energy consumption of modern data centers as AI models continue to grow in size and complexity.

4.2. AI Training vs Inference

Understanding AI energy consumption requires distinguishing between training and inference, as they place very different demands on computing systems [11]. Training teaches a model to recognize patterns in large datasets and connect them to specific outputs, while inference occurs after training when the model applies what it has learned to new data to generate predictions. Training is highly computationally intensive. It involves collecting and preparing large datasets, selecting model architectures such as Convolutional Neural Networks or Transformers, and repeatedly processing the data to calculate errors and update parameters through backpropagation. This process may require thousands of iterations across millions of data points and often runs for days or weeks on high-performance GPUs with large memory and parallel processing.

Inference is less demanding. Once deployed, the model performs only a forward pass to produce predictions without updating its parameters [11]. It can run in cloud data centers or on edge devices such as smartphones, and often operates efficiently on a single GPU or CPU.

From an energy perspective, training drives the highest electricity consumption because of its scale, duration, and reliance on GPU clusters. Inference is continuous and widespread but generally consumes far less energy per operation, meaning most of AI's environmental impact is concentrated in the training phase, especially for large models.

4.3. Data Center Scaling

Data centers have become central to the digital economy, supporting artificial intelligence, cloud computing, and real time data processing. As demand for these services grows, data center infrastructure is expanding rapidly in both scale and complexity. Estimates from McKinsey & Company suggest that global capital spending on data center infrastructure, excluding IT hardware, could exceed $1.7 trillion by 2030, driven largely by AI expansion, edge computing, and high-performance computing.

To meet rising computational needs, data center campuses are also growing in size. Facilities that once operated at tens of megawatts are now designed for hundreds of megawatts, with some approaching one gigawatt of capacity. AI training workloads are a major driver of this growth, as large models require sustained access to large amounts of computing power, increasing electricity and cooling demands.

Looking ahead, data centers are expected to evolve into hybrid facilities that support AI training, inference, and cloud services simultaneously. In the United States, data center power demand could rise from about 25 gigawatts in 2024 to more than 80 gigawatts by 2030, while global demand may reach around 220 gigawatts, highlighting the scale of the infrastructure challenge [12]. **Figure 2**, titled

"Projected Data Center Power Demand Growth (Driven by AI)," shows this projected growth, comparing the power demand in the U.S. for 2024 and 2030 with the global power demand for 2030.

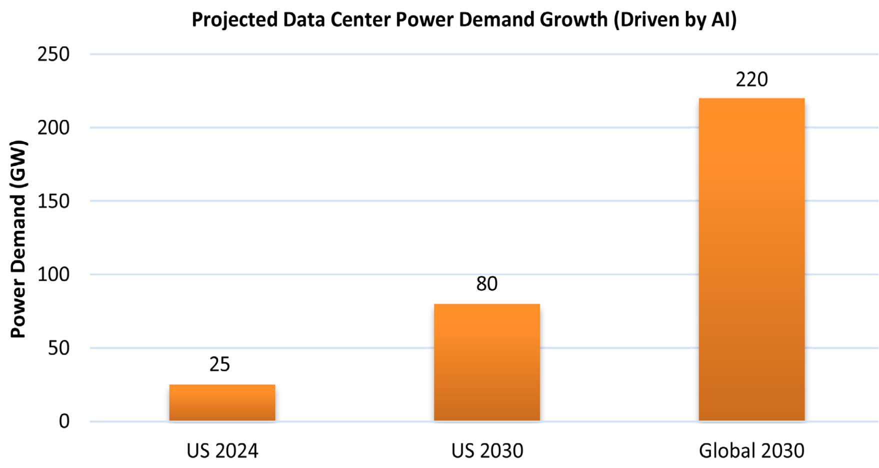

Figure 2. Adapted from McKinsey & Company. (2024-a). Scaling bigger, faster, cheaper data centers with smarter designs. https://www.mckinsey.com/industries/private-capital/our-insights/scaling-bigger-faster-cheaper-data-centers-with-smarter-designs

As AI products proliferate and inference shifts between centralized and edge environments, operators must balance large cloud campuses with distributed computing models. This transition is leading to higher rack-level power densities and the adoption of advanced chip and cooling technologies. While these innovations improve performance, they also increase pressure on electricity systems and reinforce the link between AI growth and rising energy demand.

Spending trends illustrate the urgency of this expansion. Industry estimates suggest that annual global data center infrastructure spending could surpass $1 trillion by 2030 [13]. In 2024 alone, spending reached approximately $290 billion, with major firms such as Alphabet, Microsoft, Amazon, and Meta accounting for a significant share. Capital expenditures are projected to rise further, reflecting the scale at which AI competition is accelerating data center development worldwide.

4.4. Heat Generation Physics

Heat generation is an unavoidable result of electricity use in computing systems. When electrical current flows through components, some energy is converted into heat through resistive losses known as Joule heating. The amount of heat depends on voltage, current, materials, and workload intensity. As systems operate under higher loads, more heat is produced, requiring cooling to maintain stability and prevent damage [14].

This effect is especially evident in data centers, which power artificial intelligence systems and consume large amounts of electricity. Nearly all of this energy eventually becomes heat. As AI workloads expand, global data center electricity demand could reach about 150 gigawatts by 2030, much of which will appear as waste heat [15].

High-performance AI hardware intensifies the problem. GPUs and specialized accelerators operate at high power densities, generating significant heat in compact spaces [16]. Without proper thermal management, excessive temperatures can reduce performance, shorten hardware lifespan, and increase failure risks.

As a result, many facilities are moving beyond traditional air cooling to liquid-based methods such as direct-to-chip and immersion cooling, along with advanced HVAC systems. While these improve heat removal, cooling infrastructure can account for up to 40 percent of a data center's electricity use.

Overall, greater computational intensity produces more heat, which requires additional energy for cooling, creating a feedback loop that makes thermal management a key driver of rising energy consumption in AI data centers.

5. Energy Demand Impact

5.1. How Much Energy AI Data Centers Consume Globally

The rapid expansion of artificial intelligence has created a major new source of global electricity demand. Unlike traditional digital services, AI systems rely on energy-intensive data centers with specialized hardware that operates continuously for model training and real-time inference. As AI becomes integrated across digital platforms, these facilities act as constant high-density power loads that grow with model size, usage, and computational complexity.

In 2024, global data centers consumed about 415 terawatt-hours (TWh) of electricity, roughly 1.5% of global electricity use. [17] Projections suggest this could more than double by 2030 to around 945 TWh, largely driven by AI-optimized data centers that require significantly more power than traditional facilities because of high-performance accelerated servers operating continuously [18].

This growth reflects both an increase in the number of data centers and rising power demand per facility. Modern AI workloads require uninterrupted operation of GPUs and other specialized chips, making current infrastructure far more energy-intensive than earlier generations. Most of the demand growth is expected to come from computing hardware itself, while supporting systems such as cooling will grow more gradually.

Although improvements in chip efficiency and cooling may reduce energy use per computation, these gains are unlikely to offset the rapid scaling of AI models and applications. As a result, AI-driven data centers are likely to become a long-term structural component of global electricity demand rather than a temporary surge.

5.2. Cost Burden on Operators

The expansion of artificial intelligence brings economic and environmental costs that extend beyond the companies operating data centers. While AI drives innovation across industries, the infrastructure required to train and run large models demands large amounts of electricity. Many of these costs are externalized, meaning they are often borne by households, local governments, and vulnerable communities rather than the technology firms developing the systems.

In the United States, the rapid growth of AI-focused data centers has significantly increased electricity demand. Utilities often pass the cost of grid upgrades and new power generation on to consumers through higher electricity rates. In some regions, such as Virginia, electricity bills are expected to rise because of data center expansion. AI data centers also rely heavily on electricity generated from fossil fuels, increasing air pollution linked to respiratory and cardiovascular diseases. Many facilities are located in rural or low-income communities, raising concerns about environmental justice.

At the model level, the energy intensity of AI development further increases these impacts. Research by Strubell *et al.* found that training a single large natural language processing model using neural architecture search produced over 626,000 pounds of carbon dioxide, roughly equal to the lifetime emissions of about five passenger vehicles [19]. Their study also showed that experimentation and hyperparameter tuning greatly increase emissions. While training a single Transformer model without tuning produced 1438 pounds of CO_2, the full experimentation process raised emissions to 78,468 pounds, demonstrating how development practices can significantly increase energy use [19]. **Figure 3**, titled "Carbon Emissions from AI Model Training Scenarios," illustrates these findings, showing the carbon emissions for various AI model training scenarios, from a single transformer model to large NLP model training.

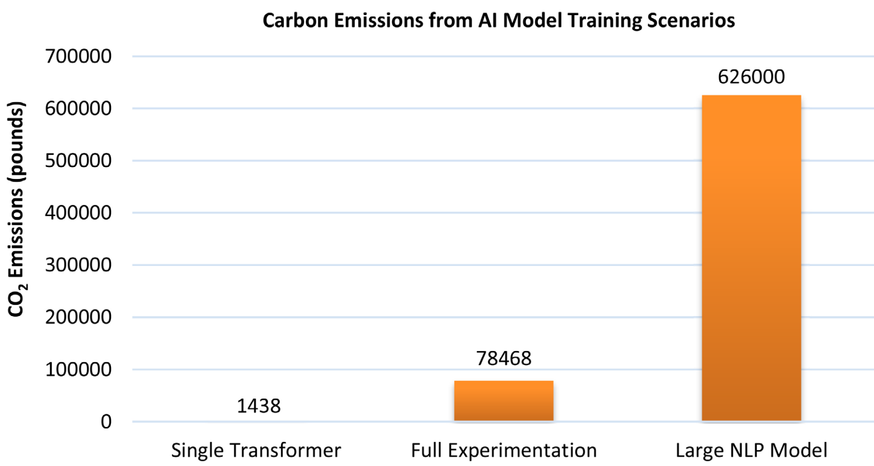

Figure 3. Adapted from arXiv. (2025). Title of technical paper 2507.09611v. https://doi.org/10.48550/arXiv.2503.05804

Beyond emissions, the expansion of AI also contributes to growing electronic waste. Advanced models rely on specialized hardware such as GPUs, TPUs, and large scale data center equipment that require resource intensive manufacturing and frequent upgrades [8]. Global e-waste levels continue to rise, yet only a small share is formally recycled. Much of the remaining waste is improperly disposed of, creating environmental and health risks from hazardous materials.

AI development is also concentrated among a small number of major technology companies, creating structural inequalities. Frontier models require extremely large computational resources that most institutions cannot access, contributing to what researchers call a compute divide. Smaller organizations and public sector institutions often face infrastructure costs that limit participation.

Energy demand also extends to cybersecurity and regulatory compliance. Large scale AI systems require continuous encryption, monitoring, and secure data storage, which increases computational and cooling needs.

Together, these factors show that the costs of AI go far beyond the electricity used by data centers, raising broader concerns about sustainability, fairness, and accountability as AI infrastructure continues to expand.

5.3. Stress on National Grids + Blackout Risks

The rapid expansion of AI data centers is increasing pressure on U.S. electricity grids. Regulators warn that the speed and scale of growth complicate supply planning, demand management, and grid stability. Although reliability remains relatively strong, rising reports of small outages and near-miss events suggest systemic risks are increasing [20].

AI demand could raise U.S. electricity consumption by about 25 percent by 2030, requiring major expansion in power generation [21]. AI workloads are also more energy-intensive than traditional digital services; a single generative AI query can use several times more electricity than a standard online search.

Climate change adds further strain by increasing cooling demand, reducing transmission efficiency, and raising wildfire risks that threaten grid infrastructure. At the same time, much of the U.S. grid is decades old and operating beyond its intended lifespan, making it harder to handle growing demand from electrification, population growth, and AI.

Large AI data centers can also create power quality issues such as electrical distortions called harmonics, particularly in major hubs like Northern Virginia and Chicago, which may damage equipment and affect nearby communities [22].

Utilities and regulators are pursuing grid modernization efforts, including faster permitting and capacity upgrades, but AI infrastructure is expanding faster than grid improvements, raising concerns about long-term reliability.

6. Cooling and Environmental Impact

6.1. Cooling Energy Costs vs Compute Energy

Cooling energy costs are becoming a major part of AI infrastructure as computing demand continues to grow. Energy use does not only come from processors but also from the cooling systems required to keep them running. For sectors that rely on on-premises infrastructure, such as healthcare and government, these combined costs directly affect the total cost of ownership, including power, cooling, and system management. Even in cloud systems, these costs are eventually passed on to customers, which makes energy efficiency an important factor to consider [23].

AI hardware operates at very high power levels. Modern GPUs can consume hundreds of watts, and multi-GPU servers may reach between 5 and 15 kilowatts under full load. When deployed at scale in large clusters, the total electricity demand increases quickly. Most of this energy is converted into heat, which then requires additional energy to remove through cooling systems [23].

Training advanced AI models can require thousands of megawatt-hours of electricity and is often seen as the main source of AI energy consumption. However, recent studies show that this view is incomplete. Large-scale inference, which involves running trained models in real-world applications, can consume equal or even greater amounts of energy over time due to continuous usage across millions of users. For example, research by the International Energy Agency highlights that inference workloads can dominate total energy consumption during the operational phase of AI systems, especially in widely deployed services [24].

As a result, energy demand does not stop after training. Global inference workloads create ongoing and scalable energy use that grows with user demand and application frequency. This also increases cooling requirements, which can account for around 40% - 54% of a data center's total electricity consumption, even in efficient facilities [24].

Category	Typical Power Consumption (kW/rack)	Primary Workloads	Driving Factors
Standard Density	<10 kW (Avg. 7 - 10 kW)	General IT, Enterprise Applications	Traditional servers, standard CPUs
High-Density	10 - 30 kW (Commonly 10, 15, 25 kW)	Virtualization, Cloud, HPC Clusters	Powerful CPUs/GPUs, consolidated VMs
Ultra High-Density/ AI-Capable	30 - 100+ kW (Avg. 60 kW+ for AI)	AI/ML Training, HPC, Big Data Analytics	Advanced GPUs (up to 1000 W), high-core CPUs (up to 400 W), accelerators

Traditional air-cooling systems are starting to struggle with the heat generated by modern AI hardware, especially as rack power densities often exceed 30 kilowatts. Because of this, liquid cooling technologies, particularly direct-to-chip systems, are becoming more common. These systems remove heat more efficiently and can reduce overall energy overhead, even though they may have higher initial costs [25].

Overall, cooling energy is no longer just a secondary concern but a key limitation for scaling AI systems. As computing power increases, cooling demand rises alongside it, directly influencing infrastructure design, operational costs, and environmental impact.

6.2. Water Usage in Cooling Systems

As cooling demands rise, water use has become a major yet often overlooked impact of AI data centers. Many facilities rely on freshwater for cooling, putting pressure on local supplies. A medium-sized data center can use up to 110 million gallons of water per year, while large facilities may consume around five million gallons per day. Across the United States, data centers use billions of gallons annually, yet fewer than one third of operators actively track their water consumption [5].

AI-focused data centers intensify this trend because higher computing density increases cooling requirements. While a single AI query uses little water, consumption grows rapidly when multiplied across billions of interactions. Many advanced AI systems rely on liquid cooling, directly linking AI usage to water demand.

A data center's water footprint also extends beyond on-site cooling. It includes water used for electricity generation and semiconductor manufacturing. Much cooling water is lost through evaporation, especially in evaporative cooling systems, and water use varies depending on climate, facility design, and rack density.

Indirect consumption further increases the total footprint. In 2023, U.S. data centers used about 176 TWh of electricity, creating an estimated indirect water footprint of 211 billion gallons. As electricity demand from data centers could reach 1050 TWh by 2030, water use linked to power generation will likely rise as well. Semiconductor manufacturing also consumes large amounts of water, with chip fabrication facilities using up to 10 million gallons of ultrapure water per day, meaning a single chip may require thousands of gallons of water before being installed in a data center [5]. **Figure 4**, titled "Growth in Data Center Electricity Use and Indirect Water Consumption," illustrates this trend, showing the projected growth in electricity consumption alongside the increasing indirect water consumption linked to data center operations.

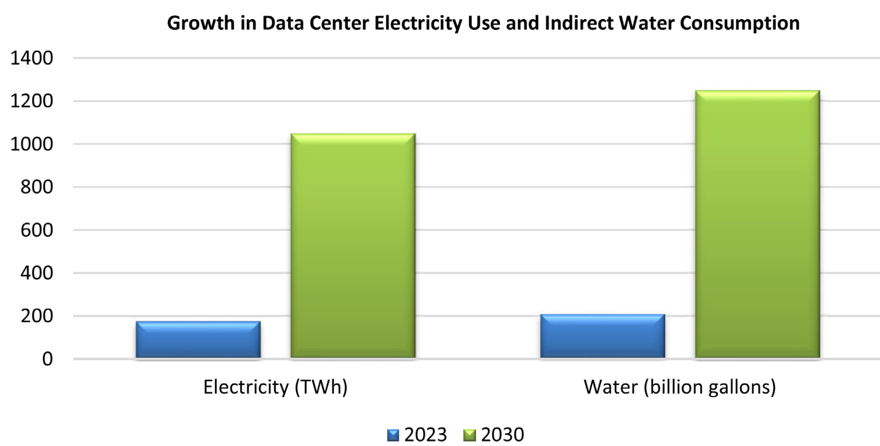

Figure 4. Adapted from environmental and energy study institute (EESI). (n.d.-b). Data centers and water consumption. https://www.eesi.org/articles/view/data-centers-and-water-consumption

Taken together, these factors show that water usage in AI data center cooling is not an isolated operational issue, but a system-wide challenge that spans energy generation, infrastructure design, and supply chains. As AI workloads continue to scale, managing water consumption will become just as critical as managing energy efficiency, particularly in regions facing increasing water scarcity.

6.3. Heat Waste

The rapid expansion of data centers is creating growing challenges for climate and decarbonization efforts. As AI systems expand into image, video, and multimedia applications, computing demand and electricity use continue to rise. Data centers and data transmission networks already account for about 1% of global energy related greenhouse gas emissions [26].

AI optimized facilities, especially hyperscale data centers, drive much of this growth. Some sites require 100 megawatts or more of continuous power. In 2022, global data centers consumed about 240 - 340 terawatt hours (TWh) of electricity, and this figure is expected to increase sharply by 2030 as AI infrastructure expands [27].

In the United States, the impact is particularly large. Data centers are projected to drive nearly half of national electricity demand growth by 2030. A study of 2132 U.S. data centers between September 2023 and August 2024 found they used over 4% of total national electricity, with more than half generated from fossil fuels. This produced over 105 million tons of CO_2 equivalent emissions, about 2.18% of U.S. emissions in 2023 [28].

Metric	Value
Share of US electricity consumption	Over 4%
Share of US greenhouse gas emissions	2.18%
Total CO_2 emissions from data centers	105 million tons (2023)

Environmental pressures extend beyond electricity use. AI data centers require large amounts of water for cooling, contribute to rising electronic waste through frequent hardware upgrades, and require significant land and infrastructure. Together, these factors show that the environmental impact of AI infrastructure goes far beyond energy consumption alone.

7. Regional Sustainability Strain

7.1. U.S. States Struggling with Grid Supply

The rapid expansion of AI data centers is placing growing pressure on electricity grids across the United States, revealing weaknesses in generation capacity, transmission systems, and long term planning. Federal energy authorities warn that current development trends could threaten grid reliability if electricity demand from AI, advanced manufacturing, and electrification continues to grow faster than new power capacity is built. After nearly two decades of relatively stable electricity consumption, demand is now rising sharply. AI driven data centers already represent an increasing share of national electricity use, and projections suggest this demand will continue to grow rapidly. Regions with heavy data center concentration, such as Northern Virginia, have become major pressure points as utilities try to supply enough power for both digital infrastructure and broader electrification needs.

This rise in demand is happening at the same time that many reliable power plants are being retired. The U.S. Energy Information Administration projects that 12.3 gigawatts of generation capacity will retire in 2025 alone, a 65 percent increase compared with 2024. These retirements include 8.1 gigawatts of coal capacity and 2.6 gigawatts of natural gas, both of which provide firm around-the-clock generation that intermittent sources cannot consistently replace. Looking ahead, the U.S. Department of Energy estimates that 104 gigawatts of firm generation could retire by 2030. Although 209 gigawatts of new generation are expected to come online during this period, only 22 gigawatts are projected to be firm baseload sources. Modeling suggests that without adequate replacement, the risk of power outages could increase by a factor of 100 by 2030, with annual outage hours potentially rising from single digits today to more than 800 hours per year, leaving millions of households and businesses vulnerable. Figure 5, titled "US Power Generation Retirements and Replacement Capacity by 2030," displays this trend, comparing the retiring firm generation capacity with the expected new generation and the limited baseload capacity available.

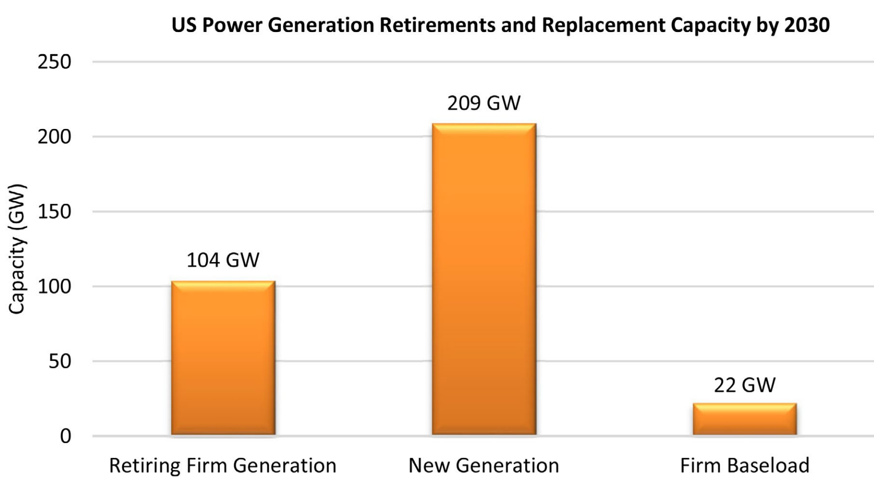

Figure 5. Adapted from U.S. Department of Energy (DOE). (n.d.-a). Department of Energy releases report evaluating U.S. grid reliability and security. https://www.energy.gov/sites/default/files/2025-07/DOE%20Final%20EO%20Report%20%28FINAL%20JULY%207%29.pdf

The imbalance between rising demand and limited new generation is already affecting regional electricity markets. Large grid operators have seen significant increases in capacity prices as utilities attempt to attract investment in new power plants. At the same time, many proposed generation projects remain delayed due to permitting processes, financing constraints, and local opposition. The rapid emergence of AI applications has intensified competition for available grid capacity, as technology firms actively search for locations with sufficient electricity supply for new data centers [29].

Additional risks come from climate and security pressures. Extreme heat, wildfires, and severe storms are increasing stress on transmission systems and pushing electricity demand to record peaks in several states. Meanwhile, the growing complexity of digital grid infrastructure introduces new cybersecurity and physical security challenges. Together, these factors show that the rising cooling and power demands of AI data centers are amplifying existing weaknesses in the U.S. electricity system, raising concerns about long-term reliability, affordability, and infrastructure resilience.

7.2. Europe Energy Crisis

Europe's energy challenge is increasingly linked to electricity grids that cannot keep up with rising demand, renewable expansion, and the rapid growth of energy intensive digital infrastructure. Aging transmission networks and limited grid flexibility are creating bottlenecks that threaten energy security, economic growth, and climate targets across the continent. The Netherlands illustrates this problem clearly. Despite strong renewable deployment, the national grid has reached capacity, with operators such as TenneT and Liander estimating wait times of up to 10 years for new connections or grid expansions. A joint study by Boston Consulting Group and Ecorys estimates that grid congestion is already costing the Dutch economy up to 40 billion euros per year. This situation shows that the Netherlands is not an isolated case, but an early example of the structural grid constraints facing Europe [30].

Indicator	Value
Estimated grid connection wait time	Up to 10 years
Annual economic cost of grid congestion	Up to €40 billion
Status of electricity grid	At full capacity
Impact on new data centers	Effective restrictions and project delays
Policy response	Moratoriums and tighter approval processes

These pressures are appearing across several European countries as electrification accelerates. In Germany, renewable generation has reached record levels, but limited transmission often forces northern wind farms to curtail output because electricity cannot easily reach the south. France faces a different issue, as its nuclear based system must adapt to increasing amounts of decentralized solar and wind. In the United Kingdom, demand from electric vehicles and building electrification is expected to significantly increase peak electricity demand in the coming decade.

Europe's electricity network connects more than 100,000 kilometers of transmission lines and serves about 500 million people. While this interconnection allows countries to share electricity, it also means disruptions can spread across the system [31].

Much of Europe's grid is over 40 years old, and investment has not kept pace with rising demand and renewable expansion. Limited transmission capacity already forces renewable energy to be curtailed because it cannot reach consumers [31].

Geopolitical tensions have added urgency. The Russian invasion of Ukraine exposed Europe's dependence on external energy supplies and led to price spikes after reductions in Russian gas flows. Although the European Union aims to eliminate Russian energy imports by 2027, transmission limits still prevent some regions from accessing cheaper or cleaner power.

At the same time, AI data centers are increasing pressure on electricity systems. Data centers already account for about 3 percent of Europe's electricity demand, and this share is rising quickly in certain regions.

7.3. Middle East Data Center Heat + Water Scarcity Problem

The expansion of AI-driven data centers in the Middle East faces significant sustainability challenges due to extreme climate conditions and severe water scarcity. High ambient temperatures, often exceeding 40˚C during summer months, reduce cooling efficiency and require much more electricity to keep high-performance computing systems operating safely. As AI servers run continuously at high utilization, cooling loads rise sharply, increasing overall power consumption and placing additional pressure on regional electricity systems that already face peak demand stress [32].

Water scarcity further limits large-scale data center development. Many Middle Eastern countries rank among the most water-stressed in the world, with freshwater withdrawals already approaching or exceeding sustainable supply levels. Although liquid cooling can improve energy efficiency compared to air cooling, it requires substantial volumes of water, making it difficult to implement in arid environments. As a result, expanding AI infrastructure risks increasing competition between digital operations and essential water needs such as agriculture and domestic consumption. Together, extreme heat and limited freshwater availability make large-scale AI data center expansion in the Middle East particularly challenging without major advances in cooling technologies, alternative water sources, and stronger energy infrastructure.

7.4. Developing Countries Lacking Infrastructure

In many developing countries, the expansion of AI data centers is limited by gaps in electricity infrastructure. Despite progress in electrification, the World Health Organization's 2025 Tracking SDG7 report estimates that more than 666 million people still lack access to basic electricity, mostly in low income and rural regions where generation capacity, transmission networks, and reliability remain limited [33] [34]. In these areas, electricity is typically prioritized for households and essential services, leaving little capacity for energy intensive facilities such as AI data centers.

Grid reliability is another major challenge. Frequent outages and voltage instability make it difficult to support the uninterrupted power required for data center operations, including computing, data storage, and cooling systems. As a result, many national development strategies focus on expanding basic electrification and improving grid reliability rather than supporting large scale digital infrastructure [35].

These constraints are particularly visible in Sub Saharan Africa, which accounts for most of the global population without electricity access. Even where grid connections exist, power systems often operate near capacity, limiting their ability to support additional large industrial loads. Developing AI data centers under these conditions would require major investment in generation, transmission, and grid management [36] [37].

These infrastructure limitations create an uneven global distribution of AI capabilities. Regions with reliable electricity systems can develop advanced technologies, while many developing countries must prioritize basic energy access, reinforcing global inequalities in AI development [38].

8. Economic and Social Consequences

8.1. Rising Costs for Companies

The rapid expansion of AI-driven data centers is imposing unprecedented financial pressure on companies across the compute power value chain. By 2030, global data center investment is projected to reach nearly 7 trillion dollars, reflecting the scale of infrastructure required to meet growing demand for compute power. Of this total, approximately 5.2 trillion dollars is expected to be dedicated to AI-capable data centers alone, while an additional 1.5 trillion dollars will be required to support traditional IT workloads. These figures highlight the extraordinary capital commitments companies must make in a relatively short time frame, often under conditions of significant uncertainty about future demand, pricing, and utilization rates [27]. **Figure 6**, titled "Projected Global Data Center Investment by 2030," shows the breakdown of projected capital expenditure, with the majority allocated to AI data centers.

Meeting projected AI demand will require a major expansion of global data center capacity, expected to reach about 156 gigawatts by 2030. Investment needs could range from 3.7 to 7.9 trillion dollars, reflecting uncertainty about how quickly AI adoption will grow and whether large infrastructure investments will deliver long term returns.

Several factors are driving these costs. Generative AI is increasing both training and inference workloads, with inference expected to dominate infrastructure

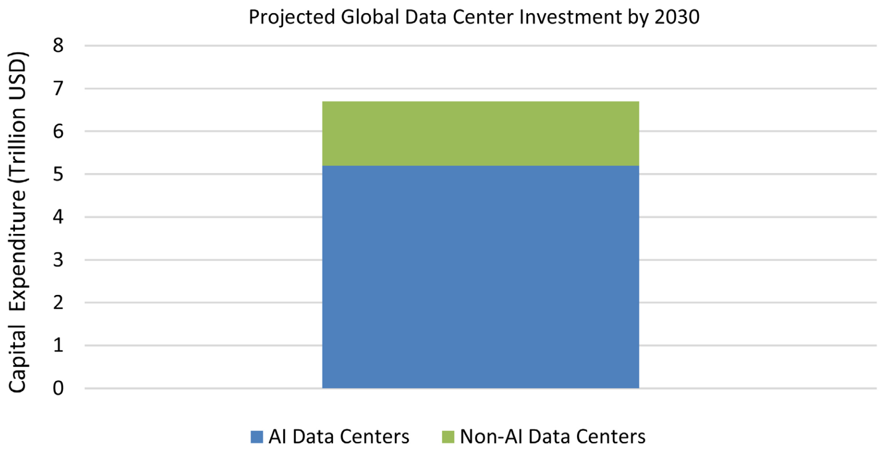

Figure 6. Adapted from McKinsey & Company. (2024-b). The cost of compute: A 7 trillion-dollar race to scale data centers. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers

demand later in the decade. At the same time, industries such as finance, healthcare, automotive, and manufacturing are integrating AI into core operations, increasing reliance on large scale cloud infrastructure. Competition among technology firms and government efforts to build domestic AI capacity are also accelerating new data center construction [39].

As a result, rising costs appear to be a structural feature of the emerging compute economy. Companies must invest significant capital while managing uncertainty related to demand growth, regulation, energy availability, and technological change.

8.2. Barriers to AI Adoption in Poorer Regions

Artificial intelligence can support economic development and improve public services in developing regions, but adoption remains uneven due to infrastructure and financial barriers. Reliable electricity is a major challenge because AI systems require stable, high capacity power for data centers and computing. Limited internet connectivity also restricts adoption, as many AI applications depend on high speed networks, while high data costs and the lack of local data centers increase latency and dependence on foreign infrastructure.

Hardware and investment barriers further limit progress. Advanced AI chips are often expensive and difficult to obtain, and building data centers requires large upfront investment that many countries struggle to finance. At the same time, shortages of skilled engineers and weak digital governance frameworks reduce the ability of local organizations to develop and manage AI systems.

As a result, AI adoption remains concentrated in wealthier regions with stronger infrastructure and financial resources, and without sustained investment in electricity, connectivity, and education, the global gap in AI development is likely to widen.

8.3. Possible Market Concentration

Market concentration in artificial intelligence arises from several layers of the AI ecosystem, including cloud infrastructure, hardware supply chains, research investment, and the geographic distribution of capital and expertise. Together, these factors reinforce one another, concentrating advanced AI capabilities among a small number of companies and countries and widening the global gap between those who can build AI systems and those who depend on external providers [40].

Cloud infrastructure is one of the clearest examples of this concentration. Global cloud spending is dominated by Amazon Web Services, Microsoft Azure, and Google Cloud. According to Synergy Research Group, these providers controlled about 63 percent of the global cloud market in the third quarter of 2025, with AWS alone holding around 29 percent. Because modern AI systems rely heavily on cloud computing for training and deployment, access to scalable compute is often mediated through these platforms, making many startups, universities, and countries dependent on foreign providers [41]-[43]. **Figure 7**, titled "Global Cloud Infrastructure Market Share," illustrates this market distribution, showing the dominance of AWS, Microsoft Azure, and Google Cloud, while highlighting the significant portion held by "All other providers".

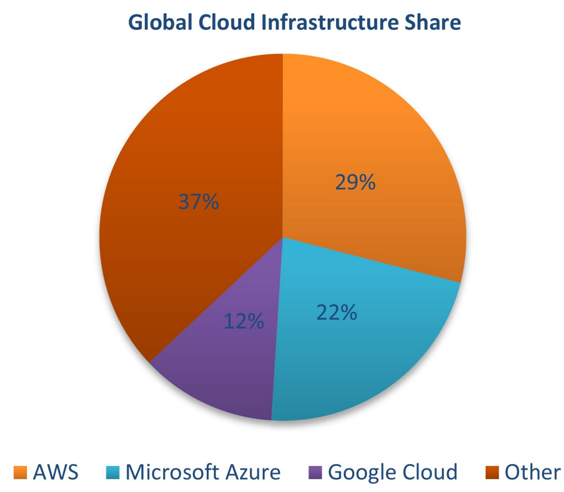

Figure 7. Adapted from Cloud market: AWS, Azure and Google further expand dominance, by Heise Online (n.d.), https://www.heise.de/en/news/Cloud-Market-AWS-Azure-and-Google-Further-Expand-Dominance-11087807.html; The big three grab two-thirds of $107 B cloud market in Q3, by TechTarget (n.d.), https://www.techtarget.com/searchcloudcomputing/news/366634757/The-big-three-grab-two-thirds-of-107B-cloud-market-in-Q3; Cloud market share trends, by SRG Research (n.d.), https://www.srgresearch.com/articles/cloud-market-share-trends-big-three-together-hold-63-while-oracle-and-the-neoclouds-inch-higher.

AI development is also constrained by hardware supply. Training advanced models requires specialized chips produced by a small number of companies, which creates supply limits and allows large technology firms to secure access first while smaller organizations face shortages [44].

Research investment and talent are similarly concentrated. Much of global AI funding and expertise is clustered in a few technology hubs, particularly in the United States and China. Strategic partnerships between cloud providers, hardware manufacturers, and AI research labs further strengthen this concentration by consolidating access to computing power, data, and capital. Governance capacity also varies widely, with developed economies more likely to shape international AI policy frameworks. Together, these advantages reinforce one another, concentrating AI capabilities and economic benefits among a small group of firms and countries [45] [46].

8.4. Environmental Justice Concerns

As artificial intelligence expands globally, its environmental footprint is becoming a growing concern. While AI is often promoted as a tool for solving environmental challenges, the data centers that support it place increasing pressure on energy systems, water resources, and public health [47].

Data centers create impacts throughout their lifecycle. Manufacturing hardware requires significant materials and energy, while operating facilities consume large amounts of electricity and water and contribute to electronic waste. In regions where electricity relies on fossil fuels, data center demand can also increase greenhouse gas emissions and strain power systems.

Water use is another major issue, as cooling systems require large amounts of freshwater, intensifying competition in areas already facing scarcity. Backup power systems and fossil fuel based electricity generation can also contribute to local air pollution and health risks.

The location of data centers often determines who experiences these impacts. Facilities are frequently built where land is cheaper and regulations are weaker, often in rural or disadvantaged communities, reinforcing environmental inequalities. In California, studies using the CalEnviroScreen tool show that many data centers are located in areas with high existing pollution burdens, including regions with elevated diesel particulate pollution.

Indicator	Statistic	What it Shows
Median pollution burden score near data centers	7/10	Data centers are typically located in areas with already high pollution levels
Percentile of environmental impact	Top 20% most polluted areas	Most data centers are sited in communities already facing significant environmental stress
Share of data centers in highest diesel pollution zones	≈33% in top 10% diesel-polluted areas	A disproportionate number of facilities are located where diesel particulate pollution is most severe
Primary pollution sources linked to data centers	Diesel generators, fossil-fuel electricity	Backup power and grid reliance increase local $PM_{2.5}$, SO_2, and NO_x exposure
Communities most affected	Low-income and communities of color	Reinforces patterns of environmental injustice and historical industrial zoning

Growing awareness of these environmental justice concerns has increased public debate about the social costs of AI infrastructure. Critics argue that without stronger safeguards, transparent planning, and community engagement, expanding data centers could deepen inequalities by concentrating environmental and health risks in vulnerable communities. Addressing these issues will require cleaner energy, more efficient technologies, and policies that distribute the benefits and burdens of AI development more fairly.

9. Policy and Regulation

9.1. Energy Efficiency Standards

Energy efficiency is often called the "first fuel" of clean energy transitions because it is one of the fastest and most cost effective ways to reduce CO_2 emissions, lower energy costs, and improve energy security. In the International Energy Agency Net Zero Emissions by 2050 scenario, efficiency is central to limiting future energy demand. The plan aims to reduce global energy intensity by about 4 percent per year this decade, roughly double the 2010-2019 rate [48].

Governments have strengthened efficiency policies in response to rising energy prices and climate risks. Countries representing more than 70 percent of global energy consumption have introduced new regulations. In the European Union, the Energy Performance of Buildings Directive requires new buildings to be zero emission by 2030. The United States, Canada, and United Kingdom have also expanded vehicle efficiency rules and building retrofit programs.

In Asia, China launched the 2024-2025 Energy Conservation and Carbon Reduction Action Plan, targeting efficiency gains across industries including steel, cement, petrochemicals, and data centers. India is promoting electric mobility through the PM E-DRIVE program. Within the EU, the Energy Efficiency Directive (2023/1791) requires large energy users to conduct regular energy audits or adopt certified management systems such as ISO 50001 [48].

Compliance Threshold	Key Requirement
2.5 - 7.5 GWh/year	Mandatory independent energy audit at least every four years, with actionable recommendations
Above 7.5 GWh/year	Implement certified energy management system (e.g., ISO 50001), plus regular energy audits
Any non-compliant organization	Subject to severe financial penalties and mandatory corrective actions

In the United States, the U.S. Department of Energy sets minimum efficiency standards for more than 70 product categories covering consumer, commercial, and industrial equipment. Together, these policies show how efficiency standards have become a central tool for reducing emissions while strengthening long term energy system resilience [49].

9.2. Renewable Energy Mandates

Renewable energy mandates and targets support the transition to cleaner energy by guiding investment and requiring utilities to increase renewable generation. Targets signal political goals, while mandates are legally binding and influence how quickly renewable energy expands.

In Europe, the Renewable Energy Directive (EU/2023/2413) requires renewables to reach at least 42.5% of final energy consumption by 2030, with a goal of 45%, across sectors such as electricity, heating, and transport.

The United States uses a decentralized system where state Renewable Portfolio Standards drive most renewable growth, while federal agencies must obtain at least 7.5% of their electricity from renewable sources [50].

In Southeast Asia, ASEAN countries aim for 23% renewable energy in total primary energy demand by 2025, with policies varying across nations [51].

Overall, these mandates aim to reduce fossil fuel dependence, improve energy security, and support the shift to sustainable energy systems.

9.3. Cooling Innovation Incentives

Cooling is becoming one of the biggest energy and climate pressure points, and it is also turning into a massive investment opportunity if policies push innovation in the right direction.

A major UNEP led report with the IFC says the cooling market in developing economies is expected to grow from about USD 300 billion to at least USD 600 billion per year by 2050. The fastest growth is projected in Africa, where the market could grow around seven times, and in South Asia, where it could roughly quadruple. The same report argues that developing economies already produce about two-thirds of global cooling related emissions and are set to double cooling demand by 2050 because of population growth, economic expansion, and urbanization, which is why shifting to sustainable cooling is treated as urgent, not optional [52].

Statistic	Value/Detail
Developing economies cooling market in 2024	~USD 300 billion per year
Projected by 2050	≥USD 600 billion per year
Growth multiplier-Africa	~7× increase
Growth multiplier-South Asia	~4× increase
Estimated economic benefits of sustainable cooling by 2050	~USD 8 trillion

Energy use for space cooling has more than tripled since 1990, increasing pressure on electricity systems. In response, over 90 countries now have building energy codes, and more than 95 countries use appliance labeling systems. Governments are also promoting innovation.

The U.S. Department of Energy funds projects through the ARPA-E COOLERCHIPS program to improve data center cooling efficiency, where cooling can account for up to 40 percent of electricity use [53]. **Figure 8**, titled "U.S. DOE COOLERCHIPS Program: Selected Cooling Innovation Funding," shows the funding awarded to various entities, including NVIDIA, Flexnode, and HP, for cooling innovations aimed at improving data center energy efficiency.

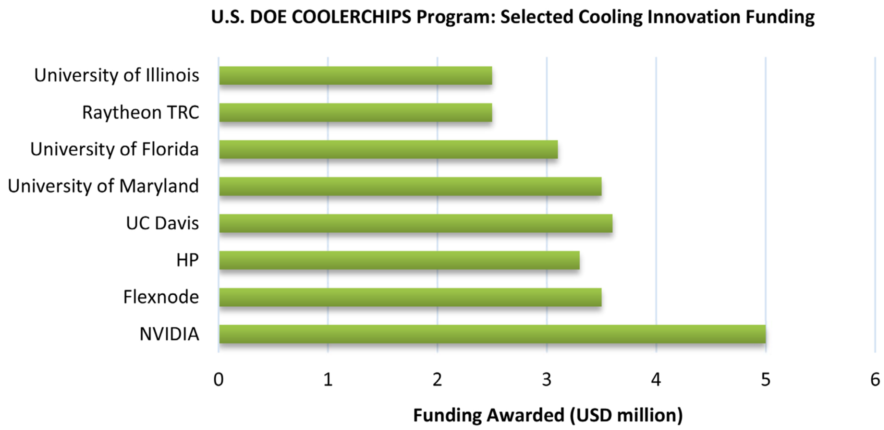

Figure 8. Adapted from U.S. Department of Energy (DOE). (2024). DOE announces $40 million for more efficient cooling in data centers. https://www.energy.gov/articles/doe-announces-40-million-more-efficient-cooling-data-centers.

In Europe, cooling policy is increasingly integrated into broader decarbonization strategies. The revised Renewable Energy Directive promotes renewable energy in heating and cooling systems and supports technologies such as high-efficiency heat pumps and district cooling networks. An EU Heating and Cooling Strategy expected in 2026 aims to further expand system integration, district energy networks, and waste heat recovery [54].

As cooling demand grows, innovation incentives are becoming an important part of climate policy. Through efficiency standards, building regulations, and funding new technologies, governments aim to meet rising cooling needs while limiting emissions and strengthening energy system resilience [55].

9.4. Geographic Planning Restrictions

Geographic planning restrictions influence where modern data centers can be built. These facilities require reliable electricity, water for cooling, transportation access, security, and strong digital connectivity. Planners often use Geographic Information Systems to map infrastructure, zoning rules, and environmental risks to identify suitable locations [56].

Infrastructure capacity is usually the main constraint. Large data centers can consume as much electricity as a mid sized town and may require significant water resources, making coordination with utilities essential. Transportation access and emergency services must also be considered [15].

Zoning regulations sometimes lag behind technological change, with many data centers still classified as warehouses or light industrial facilities. Some regions are introducing specific planning categories with standards for noise, height, and infrastructure requirements. GIS helps planners locate suitable sites while avoiding environmentally sensitive areas.

Rising demand for computing infrastructure is also making site selection more difficult. Data center capacity could triple by 2030 due to AI growth, and some hubs such as Ireland have already limited new projects because of grid constraints. Developers are therefore exploring strategies such as converting existing buildings in cities or building facilities in cooler regions with better access to renewable energy [57].

10. Future Outlook

10.1. What Happens If AI Continues Growing

If artificial intelligence continues to grow at its current pace, one major consequence will be a sharp rise in data center energy consumption. These facilities already use a notable share of global electricity, and demand is expected to increase as AI workloads require more computing power. Data centers currently account for a few percent of global electricity use, but projections suggest that by 2030 their consumption could more than double to about 945 terawatt hours per year, representing nearly 3 percent of global electricity demand. Much of this growth is linked to AI optimized data centers, whose electricity use is expected to more than quadruple over the same period [58] [59].

This increase is placing pressure on power systems, especially in advanced economies. In the United States, data centers could account for almost half of the growth in national electricity demand by 2030. AI-driven data processing is projected to consume more electricity than energy-intensive industries such as steel, cement, aluminum, and chemicals combined.

Across advanced economies, data centers are expected to contribute over 20 percent of electricity demand growth by 2030. Some hyperscale AI data centers already use electricity comparable to tens of thousands or even millions of households, highlighting the scale of infrastructure upgrades needed for power generation and grid reliability [60]. **Figure 9**, titled "Share of electricity demand growth driven by data centers by 2030," shows the projected share of electricity demand growth driven by data centers across different regions, including the United States, Japan, advanced economies, and Malaysia.

The environmental implications of this surge are significant. Without rapid growth in clean energy, rising electricity demand from data centers could increase greenhouse gas emissions and slow climate progress. Some analyses suggest AI driven data center growth could add millions of tons of carbon emissions by 2030, comparable to putting millions of gasoline powered cars on the road. Although renewables and natural gas are expected to provide much of the additional electricity, regions still dependent on fossil fuels face higher emissions risks. Data centers also require large amounts of water for cooling, which may strain local water resources.

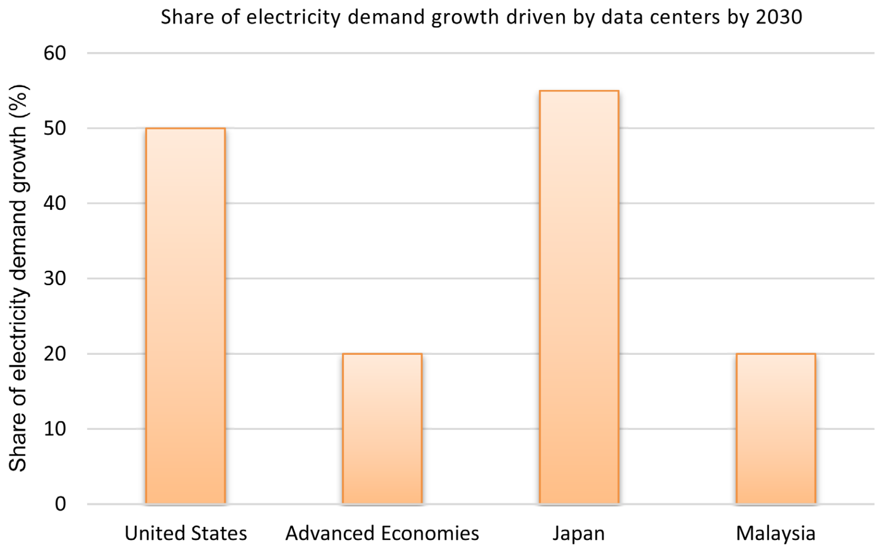

Figure 9. Adapted from International Energy Agency (IEA). (2025). AI is set to drive surging electricity demand from data centres while offering the potential to transform how the energy sector works. https://www.iea.org/news/ai-is-set-to-drive-surging-electricity-demand-from-data-centres-while-offering-the-potential-to-transform-how-the-energy-sector-works

At the same time, AI expansion is rapidly increasing demand for data centers. The generative AI market is growing at about 40 percent per year, projected to rise from 43.9 billion dollars in 2023 to nearly 1 trillion dollars by 2032. Supporting this growth requires hyperscale facilities with thousands of servers operating at high power density. In 2023, data centers consumed about 4.4 percent of electricity in the United States, and projections suggest this could rise to 6.7 to 12 percent by 2028. Some global estimates indicate AI related energy use could reach 21 percent of total electricity demand by 2030 if current trends continue.

Overall, AI growth suggests data centers may become one of the main drivers of electricity demand and infrastructure investment worldwide. While renewable energy, efficiency improvements, and better cooling technologies may help reduce impacts, the speed and scale of AI expansion mean managing energy demand will remain a major challenge.

10.2. Can Innovation Fix This?

Rising AI driven energy demand is accelerating innovation in data center infrastructure, although technology alone cannot fully solve the challenge. By 2030, companies are expected to invest nearly 7 trillion dollars in data centers, including more than 4 trillion dollars for computing hardware. Over 40 percent of this spending is projected to occur in the United States, encouraging faster construction methods and more flexible facility design.

Power delivery and grid capacity remain major constraints. Data centers require electricity at a scale existing grids were not designed to support, and building new high voltage transmission lines can take up to a decade. Research into advanced conductors, AI based grid management, and closed loop cooling systems aims to improve efficiency and reduce water use.

Energy supply strategies also vary by region. In the United States, natural gas may provide about half of the additional electricity needed by 2030 due to its low cost and existing infrastructure. In China, coal currently supplies around 70 percent of data center electricity but could fall below 50 percent by 2035 as renewables expand. In Europe, stricter climate policies mean solar, wind, and nuclear energy are expected to provide about 85 percent of data center electricity by 2030. These differences show that AI infrastructure development is shaped not only by technology but also by regional energy markets and policies [61].

Region	Key Statistic
United States	Natural gas supplies ~50% of extra data center power by 2030
China (today)	~70% of data center power from coal
China (2035)	Coal share < 50%
Europe (2030)	Renewables + nuclear ~85% of data center electricity

Looking ahead, data centers are becoming active drivers of clean energy innovation rather than just electricity consumers. Major technology companies are signing long-term power purchase agreements and investing in emerging energy technologies such as small modular reactors, advanced geothermal systems, and next-generation nuclear power, helping accelerate research and scale new energy solutions [62].

However, innovation alone cannot fully offset the environmental and infrastructure pressures created by AI-driven data center growth. While advances in cooling, efficiency, grid management, and clean energy are essential, the rapid increase in demand also requires coordinated planning, energy infrastructure investment, and supportive policy frameworks to keep pace with AI's expanding energy footprint.

10.3. AI+ Renewable Integration

As global demand for clean energy grows, artificial intelligence is improving how renewable energy systems are designed and managed. Machine learning models improve forecasting for solar and wind generation by analyzing weather data and historical patterns, making renewable energy more predictable and easier to integrate into electricity grids. AI also increases efficiency across the renewable energy system, helping renewables play a larger role in future power supply [63].

One key benefit is predictive maintenance. AI monitoring systems detect early signs of equipment wear, such as abnormal vibrations or temperature changes, allowing maintenance to be scheduled before failures occur. This approach has reduced unplanned wind turbine outages by up to 20% and extended equipment lifetimes by about 15%. Similar systems are used in solar installations to detect faults in panels and inverters early, lowering operating costs [64].

AI also helps manage electricity distribution and grid stability. Smart grids use AI to balance supply and demand in real time and improve the integration of intermittent energy sources like wind and solar. AI can also optimize energy storage by predicting demand patterns and managing battery charging cycles.

These improvements have produced measurable gains. Between 2020 and 2023, solar energy efficiency increased from 2.5% to 4.5%, wind energy efficiency rose from 2% to 3.7%, and energy storage efficiency improved from 5% to 9.5%. Grid stability increased from 2.8% to 4%, while hybrid renewable systems doubled their efficiency gains from 3% to 6% [65]. **Figure 10**, titled "Efficiency gains in renewable energy technologies enabled by AI (2020-2023)," illustrates these improvements, comparing the efficiency gains across various renewable energy technologies from 2020 to 2023.

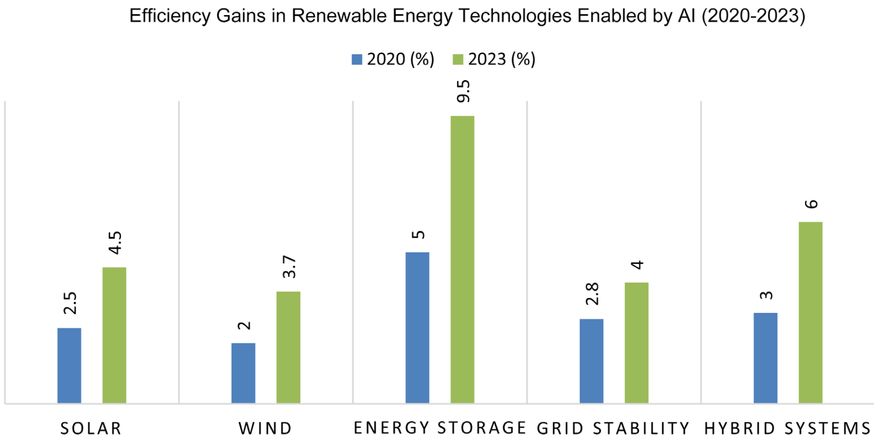

Figure 10. Adapted from ScienceDirect. (2025). Article title from S2666519025000950. https://www.sciencedirect.com/science/article/pii/S2666519025000950

Major technology companies are becoming some of the largest corporate purchasers of renewable energy as data center power demand grows. Firms such as Google, Microsoft, and Amazon are expanding renewable use through long term power purchase agreements, co locating data centers with renewable generation, and investing in energy storage and grid modernization. Amazon alone operates more than 30 TWh of clean energy across its facilities and aims to reach 100 percent renewable power in the coming years.

Technological advances are also helping renewables meet the constant energy needs of AI driven data centers. Large scale battery storage helps bridge gaps between variable renewable generation and continuous computing demand, while hybrid systems combining solar, wind, and storage provide a more stable supply. Microgrids can allow facilities to operate independently during outages and may reduce energy use by 20 - 30 percent compared to traditional grid dependent systems.

Data centers are also adopting a wider range of renewable technologies, including on site solar, wind turbines, biomass combined heat and power, geothermal heat pumps with 300 - 600 percent efficiency, and micro hydropower where available. Together, corporate investment and technological innovation are improving renewable integration, though continued infrastructure development and coordinated planning will be necessary to support sustainable AI growth [66].

10.4. Future Infrastructure Requirements

Future data center infrastructure is being reshaped by rapid AI growth and increasing energy constraints. One emerging approach is modular and distributed data centers located near renewable energy sources such as solar and wind. These facilities can scale gradually and adjust operations based on energy availability, increasing activity during periods of high renewable generation and reducing demand during grid stress [67]-[70].

Hybrid power systems are also expanding. Many data centers now combine renewables with natural gas, battery storage, carbon capture, and potentially small modular reactors to provide reliable electricity while supporting decarbonization. At the same time, AI workloads are increasing computing intensity and power density, raising cooling needs and pushing facilities toward liquid and hybrid cooling systems. Some advanced racks already exceed 160 kW [71]-[73].

Power supply is becoming a major constraint. Data centers increasingly request hundreds of megawatts of capacity, and in the United States their share of electricity use could rise from about 4.4% in 2023 to between 6.7% and 12% by 2028. As a result, future data centers will need to be more flexible and energy aware, integrating modular design, hybrid energy sources, advanced cooling, and closer coordination with power systems [74].

11. Conclusions

The rapid expansion of artificial intelligence has transformed data centers into one of the fastest growing sources of global electricity demand, reshaping the relationship between digital innovation and energy systems. Unlike earlier generations of computing infrastructure, AI-driven data centers require continuous, high-density power and intensive cooling, resulting in electricity consumption that scales directly with model complexity and usage. As demonstrated by current data and growth projections, AI is no longer a marginal contributor to global energy demand but a central factor that must be considered in long term energy planning and infrastructure development.

This paper has shown that the sustainability of AI infrastructure is highly dependent on regional conditions. In regions such as the Middle East, extreme heat and chronic water scarcity significantly amplify cooling energy requirements and create competition for limited resources, raising concerns about long-term environmental and operational viability. Similarly, in many developing countries, insufficient electricity infrastructure, unreliable grids, and limited access to power constrain the deployment of energy-intensive AI data centers, reinforcing existing global inequalities in technological capacity and economic opportunity. These regional disparities highlight that AI growth is not universally sustainable and that its benefits are unevenly distributed.

Ultimately, addressing the energy challenges associated with AI requires coordinated technological, infrastructural, and policy responses. Without improvements in energy efficiency, cooling technologies, and grid resilience, continued expansion of AI data centers risks placing unsustainable pressure on power systems and critical resources. As artificial intelligence becomes increasingly embedded in global economic and social systems, ensuring that its growth aligns with energy and sustainability constraints will be essential to achieving equitable and responsible technological development.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] IBM (n.d.) What are AI workloads? https://www.ibm.com/think/topics/ai-workloads

[2] Hitachi Energy (2025) AI Load Impact on Data Centers: Adapting to the Future of Infrastructure. https://www.hitachienergy.com/news-and-events/blogs/2025/03/ai-load-impact-on-data-centers-adapting-to-the-future-of-infrastructure

[3] S&P Global (2025) Global Data Center Power Demand Expected to Almost Double by 2030. https://www.spglobal.com/energy/en/news-research/latest-news/electric-power/110525-global-data-center-power-demand-expected-to-almost-double-by-2030

[4] Environmental and Energy Study Institute (EESI) (2025) Data Center Energy Needs Are Upending Power Grids and Threatening the Climate. https://www.eesi.org/articles/view/data-center-energy-needs-are-upending-power-grids-and-threatening-the-climate

[5] Environmental and Energy Study Institute (EESI) (2025) Data Centers and Water Consumption. https://www.eesi.org/articles/view/data-centers-and-water-consumption

[6] Environmental Law Institute (ELI) (2025) AI's Cooling Problem: How Data Centers Are Transforming Water Use. https://www.eli.org/vibrant-environment-blog/ais-cooling-problem-how-data-centers-are-transforming-water-use

[7] GlobeScan (2025) Regional Views on AI and Sustainability. https://globescan.com/2025/08/13/insight-of-the-week-regional-views-on-ai-and-sustainability/

[8] The Hill (n.d.) The Cost of AI Is Being Dumped on You. https://thehill.com/opinion/technology/5671747-big-tech-energy-burden-communities/

[9] DataCamp (2024) CPU vs. GPU: What Are the Differences? https://www.datacamp.com/blog/cpu-vs-gpu

[10] ICDREX (2023) CPU vs. GPU: Which One Is Right for Your Workload? https://www.icdrex.com/cpu-vs-gpu-which-one-is-right-for-your-workload/

[11] Nebius (2025) Difference between AI Training and Inference.

https://nebius.com/blog/posts/difference-between-ai-training-and-inference

[12] McKinsey & Company (2024) Scaling Bigger, Faster, Cheaper Data Centers with Smarter Designs. https://www.mckinsey.com/industries/private-capital/our-insights/scaling-bigger-faster-cheaper-data-centers-with-smarter-designs

[13] IoT Analytics (n.d.) Data Center Infrastructure Market. https://iot-analytics.com/data-center-infrastructure-market/

[14] ScienceDirect (n.d.) Heat Generation: Physics and Astronomy. https://www.sciencedirect.com/topics/physics-and-astronomy/heat-generation

[15] Chemistry World 2025) Modelling Suggests Waste Heat from AI Data Centres Could Power Carbon Capture and Water Purification. https://www.chemistryworld.com/news/modelling-suggests-waste-heat-from-ai-data-centres-could-power-carbon-capture-and-water-purification/4022049.article

[16] RCR Wireless (2025) Five Reasons AI Data Centers Require Massive Amounts of Power. https://www.rcrwireless.com/20250318/featured/ai-data-centers-power

[17] International Energy Agency (IEA) (2024) Energy and AI: Energy Demand from AI. https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

[18] International Energy Agency (IEA) (2025) Global Data Centre Electricity Consumption by Equipment Base Case 2020-2030. https://www.iea.org/data-and-statistics/charts/global-data-centre-electricity-consumption-by-equipment-base-case-2020-2030

[19] arXiv (2025) Holistically Evaluating the Environmental Impact of Creating Language Models. https://doi.org/10.48550/arXiv.2503.05804

[20] Agility Recovery (n.d.) AI, Power Outages, and Grid Reliability. https://agilityrecovery.com/article/ai-power-outages-grid-reliability/

[21] National Association of Manufacturers (NAM) (2025) AI Data Center Buildout Boosts Blackout Disruption Risk. https://nam.org/ai-data-center-buildout-boosts-blackout-disruption-risk-35012/

[22] Carbon Credits (2024) AI's Energy Hunger Is Straining America's Power Grids and Your Home Appliances. https://carboncredits.com/ais-energy-hunger-is-straining-americas-power-grids-and-your-home-appliances/

[23] Exxact Corp (2025) The Costs of Deploying AI: Energy, Cooling, and Management. https://www.exxactcorp.com/blog/hpc/the-costs-of-deploying-ai-energy-cooling-management

[24] Nlyte (2025) Data Center Rack Power Costs: A Condensed Analysis. https://www.nlyte.com/blog/data-center-rack-power-costs-a-condensed-analysis/

[25] Pellera (2025) Carbon Costs and Computation: The Compelling Case for Liquid Cooling. https://pellera.com/blog/carbon-costs-and-computation-the-compelling-case-for-liquid-cooling/

[26] Noda (n.d.) Data Centers: Turning Waste Heat into an Asset. https://noda.ai/insights/data-centers-turning-waste-heat-asset

[27] APL Data Center (2024) Technical Guides: Recovering Waste Heat from Data Centers. https://www.apl-datacenter.com/en/technical-guides-recovering-waste-heat-from-data-centers/

[28] Net Zero Insights (2025) Data Centers' Environmental Cost. https://netzeroinsights.com/resources/data-centers-environmental-cost/

[29] Reuters (2025) America's Largest Power Grid Is Struggling to Meet Demand for AI. https://www.reuters.com/sustainability/boards-policy-regulation/americas-largest-power-grid-is-struggling-meet-demand-ai-2025-07-09/

[30] Univers (2025) Grids in Crisis: A Wake-Up Call for Europe. https://univers.com/articles/grids-in-crisis-a-wake-up-call-for-europe/

[31] Utblick (2025) Old Infrastructure, Future Crises: Europe's Power Grid Seriously Outdated. https://www.utblick.org/2025/12/07/old-infrastructure-future-crises-europes-power-grid-seriously-outdated/

[32] World Resources Institute (WRI) (n.d.) Aqueduct Water Risk Atlas. https://www.wri.org/aqueduct

[33] International Energy Agency (IEA) (n.d.) SDG7: Access to Electricity. https://www.iea.org/reports/sdg7-data-and-projections/access-to-electricity

[34] World Health Organization (WHO) (2025) 2025 Tracking SDG7 Report. https://www.who.int/publications/m/item/2025-tracking-sdg7-report

[35] Modern Diplomacy (2025) Why Is AI Adoption Slower in the Global South and How Can It Leap Forward? https://moderndiplomacy.eu/2025/03/04/why-is-ai-adoption-slower-in-the-global-south-and-how-can-it-leap-forward/

[36] Otim Fredrick (n.d.) 8 Key Barriers Slowing AI Adoption in Africa. Medium. https://medium.com/@otimfredrick/8-key-barriers-slowing-ai-adoption-in-africa-c6b0afc0d6a0

[37] United Nations (n.d.) SDG7 Progress and Info. https://www.un.org/en/energy/page/sdg7-progress-and-info

[38] AI Certs (n.d.) Regional AI Divide: Why Wealthy Nations Dominate Adoption While Others Lag. https://www.aicerts.ai/news/regional-ai-divide-why-wealthy-nations-dominate-adoption-while-others-lag/

[39] JLL (2026) Global Data Center Market Outlook. https://www.jll.com/en-de/insights/market-outlook/global-data-centers

[40] Competitions and Markets Authority (2024) AI Foundation Models: Technical Update Report. UK Government. https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf

[41] Heise Online (2025) Cloud Market: AWS, Azure and Google Further Expand Dominance. https://www.heise.de/en/news/Cloud-Market-AWS-Azure-and-Google-Further-Expand-Dominance-11087807.html

[42] SRG Research (2025) Cloud Market Share Trends: Big Three Together Hold 63% while Oracle and the Neoclouds Inch Higher. https://www.srgresearch.com/articles/cloud-market-share-trends-big-three-together-hold-63-while-oracle-and-the-neoclouds-inch-higher

[43] TechTarget (n.d.) The Big Three Grab Two-Thirds of $107B Cloud Market in Q3. https://www.techtarget.com/searchcloudcomputing/news/366634757/The-big-three-grab-two-thirds-of-107B-cloud-market-in-Q3

[44] OECD (2025) Competition in Artificial Intelligence Infrastructure. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/11/competi-

tion-in-artificial-intelligence-infrastructure_69319aee/623d1874-en.pdf

[45] UNCTAD (2025) AI Market Projected to Hit $4.8 Trillion by 2033, Emerging as Dominant Frontier Technology. https://unctad.org/news/ai-market-projected-hit-48-trillion-2033-emerging-dominant-frontier-technology

[46] UNCTAD (2025) Technology and Innovation Report 2025: Overview. https://unctad.org/system/files/official-document/tir2025overview_en.pdf

[47] UAB Institute for Human Rights (2025) Construction and Consequences: The Human Impacts of Artificial Intelligence Data Centers. https://sites.uab.edu/humanrights/2025/10/02/construction-and-consequences-the-human-impacts-of-artificial-intelligence-data-centers/

[48] Ecochain (n.d.) The Energy Efficiency Directive (EED). https://ecochain.com/blog/energy-efficiency-directive/

[49] U.S. Department of Energy (DOE) (n.d.) Building Standards and Test Procedures. https://www.energy.gov/eere/buildings/standards-and-test-procedures

[50] U.S. Department of Energy (DOE) (n.d.) Federal Renewable Energy Use Statutory Requirement. https://www.energy.gov/cmei/femp/federal-renewable-energy-use-statutory-requirement

[51] European Commission (n.d.) Renewable Energy Targets. https://energy.ec.europa.eu/topics/renewable-energy/renewable-energy-directive-targets-and-rules/renewable-energy-targets_en

[52] UN Environment Programme (UNEP) (n.d.) 8 Trillion Opportunity: Sustainable Cooling Solutions in Developing Nations. https://www.unep.org/news-and-stories/press-release/8-trillion-opportunity-sustainable-cooling-solutions-developing

[53] U.S. Department of Energy (DOE) (2024) DOE Announces $40 Million for More Efficient Cooling in Data Centers. https://www.energy.gov/articles/doe-announces-40-million-more-efficient-cooling-data-centers

[54] European Commission (n.d.) Heating and Cooling. https://energy.ec.europa.eu/topics/energy-efficiency/heating-and-cooling_en

[55] International Energy Agency (IEA) (n.d.) Space Cooling. https://www.iea.org/energy-system/buildings/space-cooling

[56] Esri (2025) The Hidden Complexity of Data Center Sites. https://www.esri.com/en-us/industries/blog/articles/hidden-complexity-data-center-sites

[57] Procore (2025) Data Center Site Selection. https://www.procore.com/library/data-center-site-selection

[58] Huawei Digital Power (2025) Understanding Data Center Energy Consumption: Challenges and Solutions. https://digitalpower.huawei.com/en/blogs/data-center-energy-consumption

[59] Pew Research Center (2025) What We Know about Energy Use at U.S. Data Centers Amid the AI Boom. https://www.pewresearch.org/short-reads/2025/10/24/what-we-know-about-energy-use-at-us-data-centers-amid-the-ai-boom/

[60] International Energy Agency (IEA) (2025) AI Is Set to Drive Surging Electricity Demand from Data Centres While Offering the Potential to Transform How the Energy

Sector Works.
https://www.iea.org/news/ai-is-set-to-drive-surging-electricity-demand-from-data-centres-while-offering-the-potential-to-transform-how-the-energy-sector-works

[61] Brookings Institution (n.d.) The Future of Data Centers.
https://www.brookings.edu/articles/the-future-of-data-centers/

[62] ING (2025) Data Centers Race to Innovate as Power Demand Surges.
https://think.ing.com/articles/data-centers-race-to-innovate-as-power-demand-surges/

[63] Renewable Institute (n.d.) How AI Is Making Renewable Energy Smarter.
https://www.renewableinstitute.org/how-ai-is-making-renewable-energy-smarter/

[64] Hanwha Data Centers (2025) Renewable Energy for AI Data Centers: A Complete Guide.
https://www.hanwhadatacenters.com/blog/renewable-energy-for-ai-data-centers-a-complete-guide/

[65] Algburi, S., *et al.* (2025) The Role of Artificial Intelligence in Accelerating Renewable Energy Adoption for Global Energy Transformation. *Unconventional Resources*, **8**, Article ID: 100229.
https://www.sciencedirect.com/science/article/pii/S2666519025000950

[66] DataBank (2024) Data Centers and Renewable Energy Integration.
https://www.databank.com/resources/blogs/data-centers-and-renewable-energy-integration/

[67] RDM (2025) What Data Centers Have to Face in 2025.
https://www.rdm.com/blog/realizing-the-future-of-dcs-with-integrated-solutions/

[68] Reuters (2026) Microsoft Launches Data Center Initiative to Limit Power Costs, Water Use.
https://www.reuters.com/business/microsoft-launches-data-center-initiative-limit-power-costs-water-use-2026-01-13/

[69] Reuters (2026) U.S. Power Use to Beat Record Highs in 2026-2027, EIA Says.
https://www.reuters.com/business/energy/us-power-use-beat-record-highs-2026-2027-eia-says-2026-01-13/

[70] Tech Monitor (2025) Data Centre Power Infrastructure.
https://www.techmonitor.ai/comment-2/data-centre-power-infrastructure/

[71] ASHRAE (2024) Climate-Conscious Data Center Cooling in an AI Era.
https://ashrae.org.vn/wp-content/uploads/2025/08/12-18_DataCenters_Scott-Jan24.pdf

[72] ASHRAE (n.d.) Alternative Water Cooling Sources for Data Centers.
https://www.ashrae.org/technical-resources/ashrae-journal/featured-articles/alternative-water-cooling-sources-for-data-centers

[73] ASHRAE (2021) Emergence and Expansion of Liquid Cooling in Mainstream Data Centers.
https://www.ashrae.org/file%20library/technical%20resources/bookstore/emergence-and-expansion-of-liquid-cooling-in-mainstream-data-centers_wp.pdf

[74] U.S. Department of Energy (DOE) (2025) DOE Releases New Report Evaluating Increase in Electricity Demand from Data Centers.
https://www.energy.gov/articles/doe-releases-new-report-evaluating-increase-electricity-demand-data-centers