Key Questions on Energy and AI

International Energy Agency

World Energy Outlook Special Report

INTERNATIONAL ENERGY AGENCY

The IEA examines the full spectrum of energy issues including oil, gas and coal supply and demand, renewable energy technologies, electricity markets, energy efficiency, access to energy, demand side management and much more. Through its work, the IEA advocates policies that will enhance the reliability, affordability and sustainability of energy in its 32 Member countries, 13 Association countries and beyond.

Source: IEA.
International Energy Agency
Website: www.iea.org

IEA Member countries:

Australia
Austria
Belgium
Canada
Czech Republic
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Ireland
Italy
Japan
Korea
Latvia
Lithuania
Luxembourg
Mexico
Netherlands
New Zealand
Norway
Poland
Portugal
Slovak Republic
Spain
Sweden
Switzerland
Republic of Türkiye
United Kingdom
United States

The European Commission also participates in the work of the IEA

IEA Accession countries:

Brazil
Chile
Colombia
Costa Rica
Israel
Romania

IEA Association countries:

Argentina
China
Egypt
India
Indonesia
Kenya
Morocco
Senegal
Singapore
South Africa
Thailand
Ukraine
Viet Nam

This report was designed and directed by **Laura Cozzi**, Director for Sustainability, Technology and Outlooks, of the International Energy Agency (IEA), in co-operation with other directorates and offices in the IEA. The lead authors and co-ordinators of the report were **Thomas Spencer** and **Siddharth Singh**.

The lead authors of the analysis were: **Davide D'Ambrosio** (data centre demand and efficiency), **Yasmine Arsalane** (GDP analysis), **James Bragg** (financial and investment analysis), **Julie Dallard** (onsite power), **Inhoi Heo** (geospatial analysis), **Vincent Jacamon** (data centre demand modelling), **Martin Kueppers** (GDP analysis), **Brieuc Nerincx** (electricity prices), **Quentin Paletta** (geospatial analysis), **Hans-Kristian Ringkjob** (electricity prices), **Vera O'Riordan** (barriers to AI uptake), **Nicholas Salmon** (physical AI), **Max Schoenfisch** (onsite power), **Ryota Taniguchi** (power procurement), and **Brent Wanner** (electricity systems).

Other contributors were Paul Barraqué, Eren Çam, Michael Drtil, Bruno Idini, Jiahn Kim, Nikolaos Papastefanakis, and David Wilkinson.

Julia Horowitz carried editorial responsibility. **Adam Majoe** was the copy-editor. **Wonjik Yang** and **Irina Paun** led on graphic design. **Dylan Marecak** provided essential support.

Valuable comments and feedback were provided by members of senior management and other colleagues within the IEA. In particular, Stephanie Bouckaert, Alessandro Blasi, Tim Gould, Timur Gül, Brian Motherway, Maria Sicilia, Dan Wetzel.

Essential support was provided by the IEA's Communications and Digital Office, notably Jethro Mullen, Julia Horowitz, Astrid Dumond, Sam Tarling, Rob Stone, and Lucile Wall. IEA's Office of the Legal Counsel, Office of Management and Administration and Energy Data Centre provided assistance throughout the preparation of the report.

The report also benefited from analysis on the impact of AI on productivity and GDP growth from the Organisation for Economic Co-operation and Development (OECD). Valuable data, comments, and insights were provided by Peter Gal, Katharina Laengle, Francesco Filippucci and Matthias Schief in the Structural Policy and Research Division of the Economics Department of the OECD.

Peer reviewers

Many senior government officials and international experts provided input and reviewed preliminary drafts of the report. Their comments and suggestions were of great value. They include:

Harmeet Bawa	Hitachi Energy
Lisa Berry	GE Vernova
Johannes Bohner	Tennet
Alexandre Catta	Natural Resources Canada

Christina Christopoulou	Amazon
Vlad Coroama	Roegen Centre for Sustainability
Hélène Costa de Beauregard	Ministry for Ecological Transition, Biodiversity, and International Negotiations on Climate and Nature, France
Robert G. Schwiers Jr.	Chevron
Peter Gal	Organisation for Economic Co-operation and Development
Antonia Gawel	Google
Craig Glazer	PJM Interconnection
Simon Hinterholzer	Borderstep Institute
George Kamiya	Independent consultant
Mariah Kennedy	Microsoft
Jonathan Koomey	Koomey Analytics
Timothée Lacroix	Mistral AI
Sam Lowe	Permanent Delegation of Australia to the OECD
Sasha Luccioni	Hugging face
Jens Malmodin	Ericsson
Jenny Martos	Global Energy Monitor
Vincent Minier	Schneider Electric
Motoko Ogawa	Permanent Delegation of Japan to the OECD
Joshua Parker	NVIDIA
Brendan Pierpont	Energy Innovation
Nicola Rossi	Enel
David Sandalow	Colombia University
Varun Sivaram	Emerald AI
Léo Souart	OVHcloud
Stavros Stamatoukos	Directorate-General for Energy, European Commission
Tom Wilson	EPRI

Comments and questions are welcome and should be addressed to:

Laura Cozzi

Directorate of Sustainability, Technology and Outlooks
International Energy Agency
9, rue de la Fédération
75739 Paris Cedex 15
France

E-mail: EnergyAI@iea.org

Factors influencing electricity demand by data centres

Powering data centres

Impact of AI on energy

Annexes

The AI and energy nexus continues to evolve rapidly

The largest technology companies are contributing to a surge in data centre investment, as their capital expenditure exceeded USD 400 billion in 2025 – and is expected to jump by another 75% in 2026. Capital expenditure of just five technology companies is now larger than global investment in oil and natural gas production. Many jurisdictions are seeing project pipelines accelerate dramatically, although not all projects will come to fruition. Those that are moving forward are doing so at pace: the IEA's unique satellite-based tracking shows that "artificial intelligence (AI) factories" – cutting-edge data centres specifically designed for AI – have more than tripled in capacity in the past 18 months. Meanwhile, the capabilities of AI are improving quickly, increasing the likelihood that it will reshape economic growth, innovation and competitiveness and disrupt established industries and jobs.

In April 2025, the IEA published its landmark *Energy and AI* report, which provided first-of-its-kind global analysis on the links between AI and energy. Since then, the field has evolved rapidly: new questions have emerged and new data has come to light. This report builds on the foundation of previous work, providing fresh analysis on key issues. It covers the outlook for data centre electricity demand considering recent market developments; innovations in powering data centres; and the implications of these trends for energy security, affordability, competitiveness and overall energy demand.

Energy consumption per AI query has declined massively, but much more energy-intensive use cases are becoming increasingly popular

Measured per individual task, the energy efficiency of AI is improving at a rate unprecedented in energy history. Software and hardware advances have resulted in the energy use per AI task dropping by at least an order of magnitude annually in recent years. Simple text queries now typically consume less electricity than running a television over the same period of time. If all conventional internet searches were performed with simple AI text queries, it would consume less than 4 terawatt-hours (TWh) of electricity annually, equivalent to less than 1% of total data centre consumption today.

However, new energy-intensive AI applications are increasingly being launched and used, such as those for video generation, reasoning and agentic tasks. These kinds of tasks can consume hundreds or thousands of times more energy per query than simple text generation. The energy demand of AI is therefore the result of three rapidly evolving and uncertain trends: improvements in efficiency, surging uptake, and changing model capabilities, which can unlock new and, in many instances, far more energy-intensive use cases. To improve the robustness of the outlook for AI's energy demand, close monitoring, frequent updates and cooperation with the tech sector, including more systematic energy consumption disclosures, will remain important.

The global electricity demand of data centres – the critical infrastructure for training and running AI models – grew by 17% in 2025, in line with IEA projections. Electricity consumption from AI-focused data centres grew even faster, surging 50% in 2025. While there are no comprehensive statistics on the frequency and depth of AI usage around the

world, major model providers reported a threefold increase in active users and a fivefold increase in revenue over the past year, highlighting the rapid growth of demand.

Across the AI value chain, a scramble for electricity, grid connections, manufacturing capacity, chips and capital has set in

The speed of the AI revolution is increasingly contrasting with the speed of the physical, social and economic systems that underpin it. Bottlenecks across energy supply chains and advanced chip manufacturing have tightened since our last report. Planning and regulatory systems are being stretched by the wave of project applications for data centres, amid a broader trend of rapid load growth and electrification. Social acceptability is also a growing issue, as communities push back against data centre projects, and concerns about affordability and environmental impacts rise. Essential elements within the IT industry are currently facing limitations; notably, a shortage of high-bandwidth memory – integral to AI chip production – has developed over the past six months and is anticipated to persist through at least the end of 2027.

Data centre investments have grown too large to be funded from company balance sheets alone, and large amounts of funding from capital markets will be critical for their buildout. As a result, the pace of data centre growth, and the resulting increase in energy consumption, will be sensitive to market sentiment, including expectations for returns on investment in data centres and AI deployment, as well as to broader macroeconomic and financing conditions. Understanding the energy implications of AI therefore also means following closely the technology's economic trajectory.

Our updated data centre electricity demand outlook sees near-term bottlenecks but longer-term upside

The central projection for electricity demand from data centres remains close to the trajectory set out in the IEA's 2025 report. Our updated projections see electricity consumption from data centres roughly doubling from 485 TWh in 2025 to 950 TWh in 2030, accounting for around 3% of global electricity demand by that date. Electricity consumption from AI-focused data centres grows much faster than overall data centre electricity consumption, tripling in this period. Bottlenecks across the value chain, however, are reducing the likelihood of more aggressive near-term scenarios, despite booming investment and surging project pipelines.

The mid- to longer-term outlook for data centre electricity demand sees a possible upside. Investments in relieving bottlenecks across energy equipment and chip manufacturing, and rapid uptake of energy-intensive use cases of AI, raise the prospect that there could be an even higher upside case after 2030. The IEA will continue to update its projections regularly.

AI is pushing data centre power density to the limits of today's technologies

An individual server rack within an advanced data centre is only the size of a large refrigerator, but by 2027 it could have peak power demand equivalent to that of 65 households. The speed of this shift is remarkable: between 2020 and 2025, the power

density of AI servers increased by 11 times; by 2027, it is set to see a further fourfold increase. This will test the capacity to ramp up supply chains for key electricity technologies such as power electronics and transformers. Some of these technologies also depend on critical inputs from a small number of producers, notably China. Care is therefore needed to ensure that supply chains for the critical emerging technologies going into data centres are diverse and resilient.

Still an energy taker, the tech sector is also increasingly an energy maker

The tech sector remains a major driver of renewables procurement, accounting for around 40% of all corporate renewables power purchase agreements (PPAs) signed globally in 2025. Renewables PPAs signed by data centre companies are equivalent to almost half the sector's current consumption. But procurement innovation extends beyond renewables. The tech sector has become a major driver of momentum behind both conventional and new nuclear plants, as well as next-generation geothermal. Since the IEA's 2025 report, the pipeline of data centre offtake agreements with small modular reactors (SMRs) has grown from around 25 GW at the end of 2024 to 45 GW by the end of 2025. Nonetheless, the first projects are not expected to come online until around 2030.

Unlike traditional data centre operations, AI training and model use induce large and rapid power swings, making energy storage critical to ensure that electricity is always supplied reliably. By 2030, around 20-25 GW of battery storage could be installed in data centres globally, potentially making them a grid asset if the incentives are right. Recent months have seen a data centre operator sign an agreement for the largest battery project ever by energy capacity (four times larger than the previous record-holder), helping to commercialise long-duration energy storage.

Constrained by slow grid connections, data centre developers in the United States are pushing forward projects with onsite natural gas-based power generation. Satellite tracking of these projects indicates that around one-fifth have started land clearing or construction. This highlights that onsite gas power is an emerging solution, with key design, supply chain, regulatory and financial questions remaining. New IEA analysis shows that providing reliable onsite gas-fired electricity to meet critical and variable data centre load requires overbuilding onsite generation infrastructure by 30% to 70% relative to demand. Yet in the context of a supply crunch for gas turbines, it is not clear that onsite generation necessarily promises a faster route to market for data centres at scale. Though uncertainties are high, around 15-27 GW of onsite natural gas may power data centres by 2030, mostly in the United States. However, this does not remove the urgency of addressing grid bottlenecks, as most data centres prefer connecting to the grid.

The AI boom could therefore accelerate deployment and innovation in the electricity sector, if spending continues and if government support is also aligned. In this report, we analysed share price movements of AI and energy companies to understand how financial markets expect AI's energy demand to impact energy companies. On the one hand, financial valuations since the launch of ChatGPT do not suggest that AI demand will provide a generalised uplift to the energy sector; it is simply too small in the context of the energy

sector as a whole. In contrast, manufacturers of gas turbines and electrical equipment, some nuclear companies, and some energy startups have seen their valuations become more strongly linked to AI, highlighting the pull that data centre demand is providing for innovation and growth in their businesses.

Even if AI boosts economic growth, the impact on energy demand will be lower

Initial signs indicate AI is boosting productivity in some sectors, which may push up overall economic growth. There are a wide range of projections of the impact of AI on GDP. Based on cooperation with the OECD and economic modelling of the possible task-by-task productivity boost coming from AI, this report provides a first-of-its-kind analysis of the implications of an AI-driven GDP boost on the energy sector.

Stronger economic growth from AI does not translate one-to-one into higher energy demand. It is concentrated in knowledge-intensive services and high-income countries, where the elasticity between energy demand and economic activity is lower. Estimates show that, depending on the scale of uptake, an AI-driven growth boost could raise the level of global energy demand by between 1-4% in 2035 compared with trends without this AI boost. Effects are concentrated in advanced economies, particularly the United States, although emerging and developing economies also benefit from increased economic activity.

Ultimately, what matters most for energy demand are energy policies and energy technologies. In our analysis, the impact of energy policies and energy technology developments on energy demand is much larger than the potential impact of an AI-driven economic growth boost.

In an unstable world, the links between AI and energy security have tightened

Over the past year, energy and technology supply chains have become further strained. Trade restrictions have targeted key components going into data centres, such as the critical minerals needed for advanced power electronics, or the batteries and battery components needed to smooth AI loads. A 70% surge in gas turbine orders in 2025 has highlighted chokepoints in energy technology supply chains. And data centres themselves have been targeted in conflict zones, underscoring their role as critical infrastructure. The broader implications of the energy crisis tied to the conflict in the Middle East are still unknown, but they could extend to the choices that countries and companies make about the fuels and technologies used to power data centres, and where sites are located.

AI has the potential to be an important tool to enhance energy security and sustainability. For example, AI technologies monitor grids, transformers and other energy equipment to reduce unexpected failures and outages, and AI and digital grid-enhancing technologies are key to optimising the use of existing grid capacity, helping to offset lengthy and costly grid expansions. At the same time, AI risks making cyberattacks easier and more powerful, and an increasingly digitalised energy sector presents new vulnerabilities.

With increasing deployment of AI-enabled robotics, automation and efficiency solutions, AI will be critical to industrial competitiveness. Advances in data, models and hardware now allow for broader automation in industrial design and production, speeding up development, boosting innovation and lowering costs. The AI-enabled optimisation of production processes could reduce energy costs by 3-10 percentage points in energy-intensive industries, where energy is a critical production input and margins tend to be low. Well-documented AI use cases have the potential to save over 13 exajoules (EJ) of energy by 2035, equivalent to 3% of global final energy consumption, if barriers to wider uptake are overcome. Longer-term, AI could unlock another wave of industrial innovation and productivity: the almost two-fold jump in venture capital going into this field in 2025 is a marker of the potential. The race to develop more powerful AI models is being run alongside another race to apply AI across the economy for innovation and productivity, with China particularly focused on the latter.

But the energy sector is not yet taking full advantage of AI opportunities

An IEA survey of energy companies reveals that the lack of digital skills is the single largest barrier to greater AI adoption in the energy sector. The adoption of AI is also limited by fragmented data and concerns related to data protection, privacy, and cybersecurity. For example, only 10% of global electricity consumption is covered by open electricity data policies. The lack of adequate digitalisation of equipment can also be a barrier. Globally, less than half of energy demand is covered by policy frameworks aiming to promote the uptake of AI in the energy sector.

Social concerns around AI have grown, focusing particularly on the environment and electricity prices…

Data centres are a highly visible flashpoint for concerns around energy prices and the environment. Energy prices have risen up the agenda since the inflationary shock of Covid-19 and the 2022 global energy crisis, with the energy crisis sparked by the conflict in the Middle East compounding concerns. In addition to the potential impact of data centres on electricity prices, surveys increasingly show that citizens are concerned about AI's effect on jobs, the economy, the environment, and society more broadly. The emissions associated with data centres double in IEA projections, reaching around 350 million tonnes in 2035, but still make up about 2% of global electricity sector emissions by that date.

…but with the right conditions, data centres do not necessarily raise electricity prices

The impact of electricity demand growth on electricity prices depends on a combination of fundamental factors and policy choices. In systems with tighter demand-supply balances, load growth can indeed trigger a need for new investments, potentially raising prices. On the other hand, in systems with excess electricity supply, load growth can result in more efficient capital utilisation and lower prices. The shape of the load growth matters too. In the same way that high and reliable passenger numbers can allow airlines to lower average ticket prices on established routes, predictable baseload electricity demand can increase the

efficiency of the use of capital-intensive power plants and grids, potentially lowering electricity prices.

Data centres can create special challenges for electricity affordability. Data centres are large, concentrated, rapidly developed infrastructure that are likely to trigger a need for new generation and grid investments in systems that host them. Actual peak load from data centres is often uncertain, as data centres are filled progressively with operating servers and often oversize their grid connections initially. The mismatch between fast-moving data centres and slow-moving energy investments, and the uncertainty over their load, creates risks of misalignment between electricity system investment and data centre demand that could raise electricity prices in some places if unmitigated.

Policy remains key to ensuring AI and data centres play a constructive role in energy systems

The IEA proposes three principles to help ensure AI deployment is leveraged for the benefit of the energy sector, and that data centres minimise any adverse impacts on electricity systems.

- **Proactive management of data centre project pipelines and electricity sector investment can support adequate and reliable electricity for the sector without adversely affecting prices.** This includes reforming the management of data centre connection queues and streamlining permitting timelines. Similarly, more robust demand projections, building on greater disclosure from technology companies and strong cooperation with system operators, can help ensure investments are well aligned. Finally, how costs are allocated in the electricity system is ultimately up to policymakers. Tools such as tariff schemes can support the fair allocation of the costs of grid upgrades and new generation capacity.
- **Approaches that promote electricity system flexibility can help accelerate grid connections and ensure electricity affordability.** This can help defer the need for large investments and improve the efficiency of expensive grid and generation investments. System operators can explore non-firm grid connections and incentivise data centre developers to provide demand response in return for faster connection processes. Increasingly sophisticated, grid-interactive onsite power assets, such as battery storage and gas-fired generators, can help data centres support grid operations, moving data centres from grid loads to grid resources.
- **Removing barriers to AI adoption in the energy sector can ensure AI is leveraged to enhance energy security and sustainability.** Comprehensive policy frameworks that address data availability, cybersecurity, skills, and interoperability are crucial for boosting AI uptake. As public concerns grow around the local impacts of data centres, and the broader implications of AI on jobs and fairness, social acceptability will hinge on demonstrating AI's benefits for affordable, secure and sustainable energy systems, including at the local level.

Key questions on energy and AI

AI at the intersection of energy security, innovation and competitiveness

Introduction

The IEA published its first special report on energy and artificial intelligence (AI) in April 2025 (IEA, 2025a). The report explored three questions. Firstly, how much energy does artificial intelligence, and the data centres that underpin the technology, need? Secondly, how can the energy sector meet growing demand from AI and data centres? And finally, to what extent can AI help optimise the energy sector and accelerate energy innovation? The report provided the first global and system-wide analysis of these questions, building on the IEA's unique models and data. The report helped contextualise the role of AI in the broader energy system – and continues to remain relevant at the time of publication of this new report.

AI, however, is a fast-moving field. There have been several key developments since the publication of the first special report. AI models have continued to improve and new use cases, such as multi-step tasks (agentic AI), have become commonplace. As AI advances, so does its potential to transform the economy, raising questions around whether an AI-driven growth boost could trigger a surge in energy demand beyond AI's direct energy use.

Energy supply chain and permitting bottlenecks have intensified, while rapidly changing data centre designs are altering the power needs of data centres. In response, data centre developers are searching for innovative power solutions, including onsite gas generation, batteries, and demand response. Outside the energy sector, constraints in supply chains for manufacturing the cutting-edge chips and servers going into data centres also need to be factored into the outlook, and these supply chains are shifting rapidly in response to developments within AI.

At the same time, there are growing concerns over the impact of data centres on the environment and electricity prices. This has led to increased public scrutiny, and to policy responses to address these issues. In parallel, new datasets have been released over the past year, which throw further light on the intersection of energy and AI. For example, AI companies have disclosed more information about the per-task energy consumption of AI and the efficiency improvements being achieved in AI software and hardware.

These fast-moving developments necessitate new analysis to advance and deepen the understanding of AI's links with the energy sector and support evolving policy responses. This new report builds on our 2025 publication and is structured in three sections:

- Factors influencing electricity demand by data centres
- Powering data centres
- Impact of AI on energy

In each of these sections, the material is organised around key questions, which pinpoint new and important developments since our last report. Given that this report builds on our 2025 publication, the remainder of this introduction provides a high-level summary of the key findings of our 2025 analysis.

Key findings from the IEA's Energy and AI special report of 2025

Our 2025 report provided comprehensive estimates of data centre electricity consumption historically and projections to 2030 and 2035. These projections were structured into four sensitivities, or cases. Our Base Case saw electricity consumption of data centres more than double by 2030, while in our highest case, Lift-Off, it tripled in the same timeframe. The wide range of outcomes was due to uncertainties on the uptake and economics of AI; the outlook for efficiency improvements of models and chips; and in how quickly energy sector bottlenecks can be resolved.

Our 2025 report projected that data centres will be powered by a range of sources, led by renewables and natural gas. Most data centre developers have specific sustainability goals and are major procurers of renewables. At the same time, the scale of data centre load growth, taking place in a broader context of rising electrification, means that utilities are seeking to procure additional, dispatchable capacities to meet overall load growth reliably. In several markets, particularly the United States and Middle East, much of this is set to come from natural gas. At the same time, the challenge of grid integration of data centres is also creating incentives for innovative solutions, which our 2025 report analysed. These include incentivising data centres to provide various forms of demand flexibility to facilitate smoother integration into electricity grids.

Our 2025 report also analysed the applications of AI across the energy system. It documented numerous use cases of AI in optimising energy sector processes, from managing power systems, to optimising industrial processes, to supporting the exploration and production of oil and gas. In addition, the report analysed the potential for AI to accelerate energy technology innovation. Overall, our analysis found that while AI has substantial potential to improve the operation of energy systems, and accelerate the discovery of new technologies, it is not a "silver bullet" for challenges like environmental sustainability or affordability.

Finally, our report analysed emerging cross cutting issues, such as the total investments needed in data centres and related energy infrastructure, the supply chain risks arising from the build-out of data centres for components such as critical minerals or transformers, and the specific challenges faced by emerging market and developing economies.

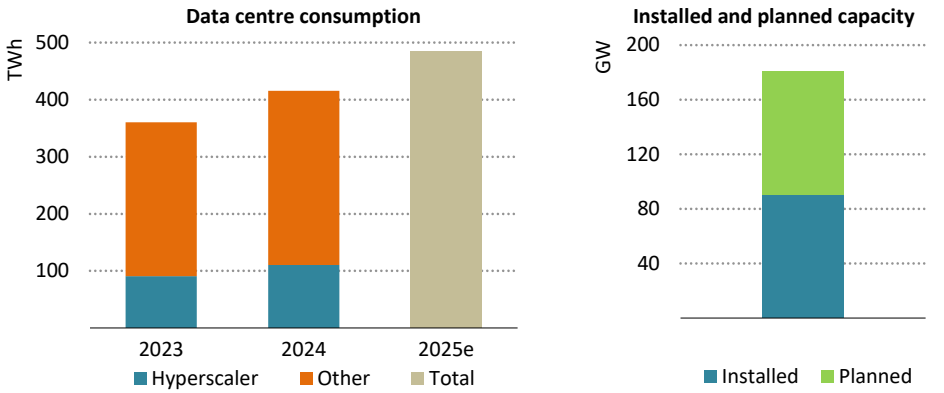
This analysis forms the basis of the present report, which aims to provide a timely update of a field that is moving at an unprecedented pace compared to the typical speed of developments in the energy sector.

Factors influencing electricity demand by data centres

1 In light of recent market trends, has the IEA adjusted its data centre demand outlook?

Electricity consumption by data centres grew by more than 15% between 2024 and 2025, adding around 70 terawatt-hours (TWh) of electricity demand and taking the sector's annual consumption to almost 500 TWh, or slightly more than 1.5% of global electricity demand. The combined electricity consumption of four major hyperscalers grew by 20% in 2024, the most recent full year for which data are available, reaching 110 TWh. Data on the planned global project pipeline for data centres as of the beginning of 2026 imply that, if all projects are realised, installed capacity from the sector would at least double in the coming years (Figure 1.1).

Figure 1.1 ▷ Electricity consumption by hyperscalers and all data centres, and global project pipeline for commercial data centres



There is significant momentum behind data centre electricity consumption growth

Notes: hyperscalers = Amazon Web Services (AWS), Microsoft, Alphabet and Meta. AWS's data centre electricity demand is derived from water demand figures. The data centre project pipeline refers to commercial data centres only (hyperscale and colocation). Total = total estimated data centre consumption; no hyperscaler reported data is available for 2025 at the time of writing.

Sources: IEA analysis based on company data from company reports: Amazon (2025), Microsoft (2025), Google (2025a), Meta (2025); and project pipeline from OMDIA (2025).

The IEA produced comprehensive global and regional projections of data centre electricity consumption in its 2025 report, *Energy and AI* (IEA, 2025a). The report highlighted several of the challenges involved in projecting data centre electricity consumption. Data on the sector remain patchy, which necessitates careful analysis of a range of sources to ensure robustness. The uptake of artificial intelligence (AI) is proceeding very rapidly, but technological, financial and commercial uncertainties remain. Moreover, the sector is integrated into complex value chains, ranging from chips to transformers, electrical switchgear, turbines and local electricity markets, and constraints in these value chains play out in complex ways. Against this backdrop, this section sets out some key considerations shaping the evolving outlook.

1.1 Big numbers, complex signals: Interpreting the message from financial markets

AI today is highly capital intensive. Cutting-edge data centres for AI can have capital costs, including information technology (IT) equipment, of USD 40 000 to 50 000 per kilowatt (kW) (five to ten times more than a nuclear power plant on a per-kW basis). The IEA's estimates for data centre growth imply cumulative investment in data centres of USD 3.9 trillion between 2026 and 2030. This is too large to be funded solely from the balance sheets of AI companies and will require capital from debt and equity markets. As a result, financial market perceptions of the return on investment in data centres and AI will be a material factor shaping the sector's development.

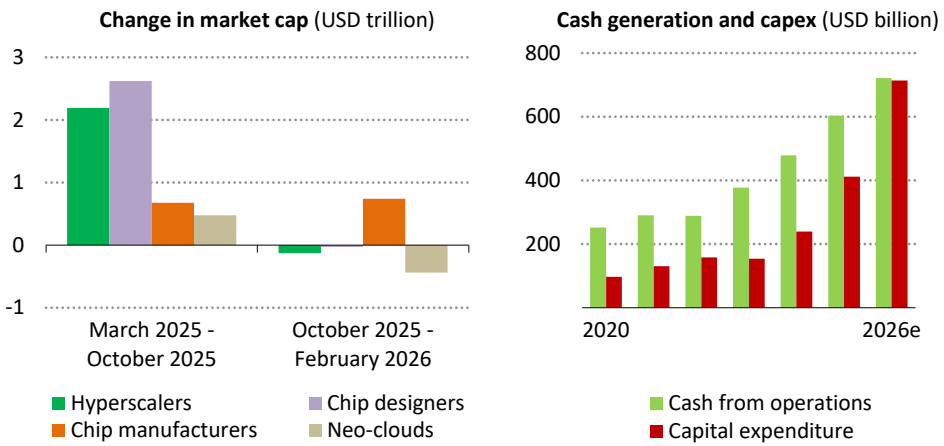
In the second half of 2025, the sentiment among investors in financial markets shifted from generalised optimism regarding AI towards a more cautious stance in light of the huge investments required and uncertain prospects for monetisation. Hyperscalers saw their market capitalisation increase by around USD 2 trillion from March 2025 to October 2025, while the value of leading chip designers increased by USD 2.5 trillion (Figure 1.2). New entrants into the data centre sector, so-called "neo-clouds", saw a 110% increase in market capitalisation across the same period, representing an aggregate gain of USD 0.5 trillion. Chip manufacturers increased their market capitalisation by around USD 0.7 trillion.

Since then, growth in market capitalisation has ground to a halt across the AI value chain. Neo-clouds have given up almost all of their gains since March 2025, while the market capitalisation of established hyperscalers and chip designers has stagnated. While there are pockets of growth, the shift is quite broad-based. One exception is chip manufacturers, which have continued to see strong share price growth, notably due to the shortage of memory chip manufacturing capacity.

A major driver of this shift in market sentiment has been concern over the huge growth in capital expenditure for data centres. Leading hyperscalers and neo-clouds have announced a 75% increase in capital expenditure for 2026, amounting to USD 715 billion. This is substantially more than annual investment in the entire energy sector of the United States, which totalled slightly less than USD 600 billion in 2024. In recent years, hyperscaler

investments have largely been funded through cash from operations. However, the upswing in capital expenditure has been so significant that it will absorb almost all cash generation in 2026, leading companies to turn increasingly to debt markets (Figure 1.2). In 2025, normally cash-rich AI companies issued around USD 200 billion in debt, and they are set to become among the largest issuers of corporate debt going forward (Bloomberg, 2025).

Figure 1.2 ▷ Change in market capitalisation across the AI value chain, and cash from operations and capital expenditure by major hyperscalers and neo-clouds



Markets are sending mixed signals: hyperscaler capital expenditure, largely on data centres, has surged, while equity market investors have become more cautious

Notes: Hyperscalers = Alphabet, Meta, Microsoft and Amazon Web Services; chip designers = Nvidia, AMD, Arm and Broadcom; chip manufacturers = TSMC, Samsung Electronics and SK hynix; neo-clouds = Oracle, CoreWeave and Nebius; Capex = capital expenditures. The right-hand chart shows hyperscalers and Oracle. Capital expenditure for 2026 comes from company guidance, except in the case of Microsoft, where it is based on consensus estimates. Cash from operations for 2026 is based on consensus estimates.

Source: IEA analysis based on data from S&P Global (2026).

Markets are therefore sending mixed signals about the outlook for data centre growth. Capital expenditure, largely on data centres, is surging, signalling both optimism about the demand outlook among AI companies and stiff competition to capture it. At the same time, equity and debt markets are showing increasing signs of concern that this surge in spending is running ahead of AI monetisation. The continued growth in the market capitalisation of chip manufacturers highlights the bottlenecks in scaling cutting-edge manufacturing capacity and the reality that the AI surge is increasingly coming up against physical constraints (see Section 1.3).

1.2 Public policy support

Data centres are increasingly viewed by national policy makers as essential infrastructure that is critical for digital sovereignty, security and economic competitiveness. As a result, a growing number of governments are putting in place policies to support data centre growth.

The world's leading data centre markets, the United States and China, both have supportive policies in place. In the United States, as of late 2025, at least 37 states (NCSL, 2025) had tax incentive measures targeting data centres, while the federal government has sought to accelerate the permitting and build-out of data centre infrastructure. In 2025, China continued to roll out AI infrastructure, registering 42 graphics processing unit (GPU) clusters with more than 10 000 chips (China Daily, 2026), and a large fibre optic network upgrade was commissioned to support the development of the "East Data, West Compute" initiative. At the same time, senior officials in China, including the president, have warned against overinvestment in data centres, and a large number of data centre projects have been cancelled in recent months (Data Centre Dynamics, 2025).

Over the past decade, the growth of data centre capacity in the European Union has been slower than the global average, resulting in a loss of global market share. However, through its AI Continent Action Plan framework (European Commission, 2025a), the European Commission has set out the goal of tripling data centre capacity in the next five to seven years. Recent months have seen a notable acceleration in the development of the bloc's project pipeline. Based on IEA analysis, the total capacity implied by the project pipeline is 2.3 times current installed capacity (IEA, 2025b), although overcoming energy constraints will be critical to achieving this. Other markets in Europe, such as the United Kingdom and Norway, have also seen significant growth in their project pipelines.

Other key markets, especially in Asia, have shown strong momentum. India recently adopted new tax incentives for data centres in its 2026 budget and has seen growth in its project pipeline. The government's position to expand data centre capacity in the country was emphasised during the AI Impact Summit, which it hosted in February 2026. Southeast Asia has also seen rapid growth in its project pipeline, with more than 6 gigawatts (GW) proposed. Malaysia is a key driver of this momentum (Baker McKenzie, 2025), with dedicated policy incentives to maintain an attractive environment for AI-focused data centre operators. In addition, Singapore has launched a call for applications for 200 megawatts (MW) of additional capacity (IMDA, 2025), reversing the pause on new projects that had been in place since 2019. Indonesia also has a significant pipeline of projects, although many are at an early stage (Ember, 2025).

The rise of "gigafactory" announcements for AI data centre projects was another key trend in 2025.[1] These huge projects, which in many cases exceed the total installed data centre capacity of the host countries today, make projecting data centre electricity consumption challenging for some regions. Most of these projects are at a very early stage of development

[1] In this context, gigafactory refers to gigawatt-scale data centres.

and, because of their size relative to existing infrastructure, the cancellation or delay of a single project could fundamentally change the outlook for data centre electricity consumption. For instance, in the Middle East, the pipeline of gigafactory data centres is an order of magnitude larger than the region's entire installed data centre capacity today (Epoch AI, 2026a). The total project pipeline in the Middle East is almost 10 GW.

1.3 Physical bottlenecks

In the IEA's 2025 report *Energy and AI*, we highlighted several physical constraints that may slow the development of data centres (IEA, 2025a). These included bottlenecks in supply chains for electrical equipment, grid bottlenecks and constraints across the IT value chain, such as the following:

- Key auxiliary equipment used in data centres is facing critical bottlenecks, with lead times increasing and prices rising. For instance, lead times for transformers average between two and three years. For gas turbines, deliveries could take around five years (Wood Mackenzie, 2025).
- Waiting times for grid connections can be as long as five to ten years in many jurisdictions. Policy makers and regulators are taking action to address this bottleneck, including through reforms to connection procedures and connection queue management (see Box 1.2), as well as investments to make grids more efficient, but progress will take time. Recently, numerous proposals have emerged for data centres powered by onsite generation, which may help ease this constraint in some instances, where feasible (see section 6).

Outside the energy sector, the data centre build-out is also facing constraints in the supply chain of key IT equipment. In 2025, limitations in the packaging capacity for high-end chips were one such constraining factor.[2] In the second half of 2025 and beginning of 2026, global production capacity for high-end memory also emerged as a binding constraint on the rate of AI server production. Specialised AI chips, such as GPUs, require enormous amounts of high-bandwidth memory. In 2023, the market-leading GPU systems included less than 150 gigabytes (GB) of memory. By 2027, this is set to grow by an order of magnitude to more than 1 000 GB (SemiAnalysis, 2025a). AI is highly memory-intensive and is therefore the key driver of this trend. Producing high-bandwidth memory requires three times the wafer capacity of conventional memory to achieve the same GB output (Micron, 2024). The result of this trend has been a sudden production bottleneck in high-end memory. Prices for memory increased by an order of magnitude between late 2024 and early 2026, and inventory declined from almost four months in late 2024 to around three weeks in late 2025 (Figure 1.3). Memory manufacturers are racing to build new factories, but the constraints will not be addressed overnight. It takes about two years to bring a new chip factory into

[2] Packaging refers to the final stage in the chip production process, where the silicon die is enclosed in a protective casing, providing power distribution and thermal management for the chip.

operation. As a result, the shortage of AI memory could last at least until late 2027 (IDC, 2025). Analysts estimate that current high-bandwidth memory production can support a maximum of around 25 GW of AI-ready servers per year through to 2027, broadly in line with the rate of capacity additions in the IEA's Base Case. More recently, the current disruptions to supply chains in the Middle East could impact critical production inputs for data centres, such as helium (Box 1.1).

In addition to supply chain constraints, concerns about the local impact of data centres, including their implications for the environment and energy affordability, have led to the cancellation or delay of data centre projects. In the second quarter of 2025, 20 projects with estimated investment of USD 98 billion were blocked or delayed in the United States (Data Center Watch, 2025). Growing local opposition could also affect the provision of policy, regulatory and fiscal incentives for data centres, such as tax exemptions.

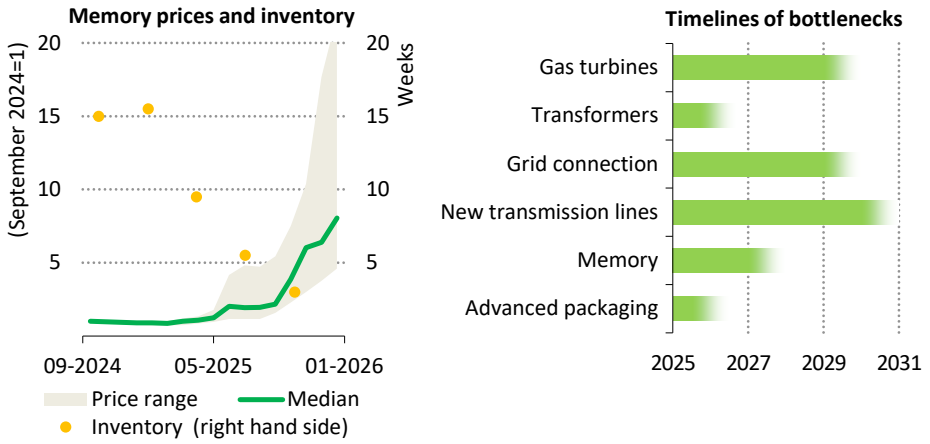
Box 1.1 ▷ What is the role of helium in AI chip production, and what is the impact of the 2026 Middle East conflict on its supply?

Helium is a critical enabler of advanced AI chip manufacturing. Helium's exceptional thermal conductivity allows it to efficiently cool manufacturing equipment, and its chemical inertness (inability to react with other chemicals) keeps contamination-free conditions stable. For the manufacturing of high-performance chips used in AI, there is currently no substitute that can match helium's cooling and cleaning properties. Without a steady supply, the precision machines owned by companies like TSMC and Intel simply cannot function.

The ongoing conflict in the Middle East has highlighted the critical role of this gas in AI chip manufacturing, and the impact of the conflict on chip shipments. Qatar, which supplied roughly one-third of global helium before the conflict, has seen production at Ras Laffan and related facilities significantly damaged, with reports of around 14% cuts to annual helium exports as of March 2026 (AP, 2026) and an effective temporary loss of supply equivalent to approximately one-third of world production (USGS, 2026). The strikes forced Qatar's state producer to declare force majeure, while the partial closure of the Strait of Hormuz has constrained shipping routes that move a significant share of global helium volumes, driving spot prices sharply higher. These disruptions disproportionately affect chip manufacturing hubs of Chinese Taipei and South Korea – which manufactured a large majority of all advanced semiconductors globally – and which sourced more than half of their helium from Gulf suppliers in recent years (Gas World, 2026).

Since the disruption of supply, semiconductor manufacturers have been leaning on inventories and diversified contracts to continue production. The IEA will continue to monitor the supply chains of key inputs to the manufacturing of chips to help inform its analysis on data centre server shipments and therefore electricity demand from them.

Figure 1.3 ▷ Price and inventory trends for memory and estimated timelines for supply chain bottlenecks



Physical constraints are limiting the rate at which data centres can be built and powered in the near term

Note: Memory refers to the manufacturing of high-bandwidth memory; advanced packaging refers to the final stage in the chip production process.

Sources: IEA analysis based on data from Bloomberg (2026), IDC (2025), IEA (2025a), Refinitiv (2025), Reuters (2025), and Wood Mackenzie (2025).

1.4 Key takeaways

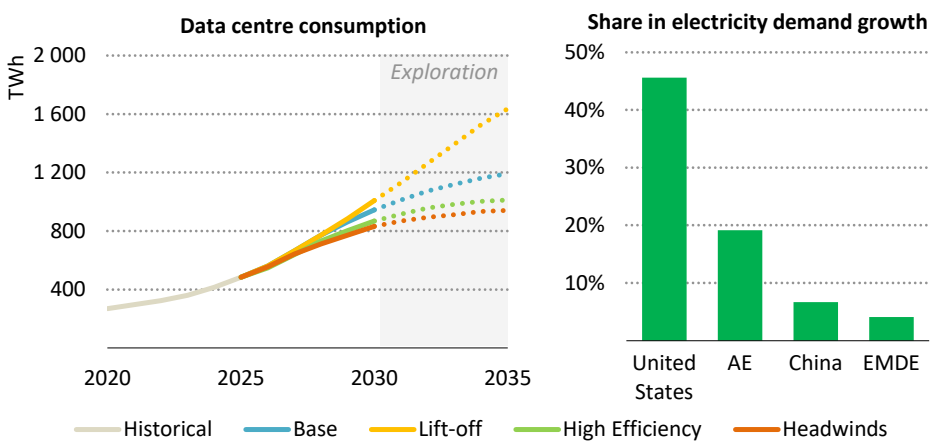
Projecting the electricity consumption of data centres is a complicated exercise that requires consideration of a range of factors. These include: the economics of and demand for AI itself; trends in energy efficiency and the development of more energy intensive AI modes; and how bottlenecks across the energy and IT equipment supply chain play out in the years to come. These uncertainties mean that a scenario framework is useful to provide different stakeholders with the tools to assess the implications of different possible trajectories. In our 2025 report, we developed a scenario framework based around four cases:

- The **Base Case** provides the central projection and considers trends in AI development, investment, and bottlenecks in energy and IT supply chains.
- The **Lift-Off Case** explores the impact of stronger AI adoption and increased global demand for digital services, leading to even stronger deployment of data centre facilities than in the Base Case. It assumes that bottlenecks can be effectively resolved over time, with increases in manufacturing capacity for chips and energy equipment, and policy progress and innovation easing constraints such as grid connection wait times.

- In the **Headwinds Case**, service demand and AI adoption grows more slowly. AI monetisation challenges reduce investment, while local constraints and electricity supply issues delay data centre development. Broader macroeconomic headwinds, such as higher interest rates or trade restrictions, could also push outcomes in the direction implied by the Headwinds Case.
- In the **High Efficiency Case** we assume that AI and digital services demand follow the same trajectory as in the Base Case. However, several efficiency strategies are implemented to counterbalance the increased energy demand resulting from the higher adoption of digital technologies, particularly AI. These include rightsizing models, continued improvements in model efficiency, and possible shifts towards more edge computing.

Taking recent developments into account, the balance of evidence suggests that the scenario framework we developed in our 2025 Energy and AI report continues to provide a sound frame of reference out to 2030. Now updated for this report, our Base Case still sees data centre electricity consumption growing from around 415 TWh in 2024 to around 950 TWh in 2030 (Figure 1.4). This means that data centre electricity consumption accounts for slightly less than 10% of total electricity demand growth globally to 2030, although some regions are much more affected. While some trends, such as booming investment and growing project pipelines (see Box 1.2) would appear to point to an upward revision in the Base Case, we believe this case continues to provide a balanced assessment of both the strong momentum in AI and data centres, and the real financial and physical frictions that limit the rate of near-term growth.

Figure 1.4 ▷ IEA cases for data centre electricity consumption, 2020-2035



IEA. CC BY 4.0.

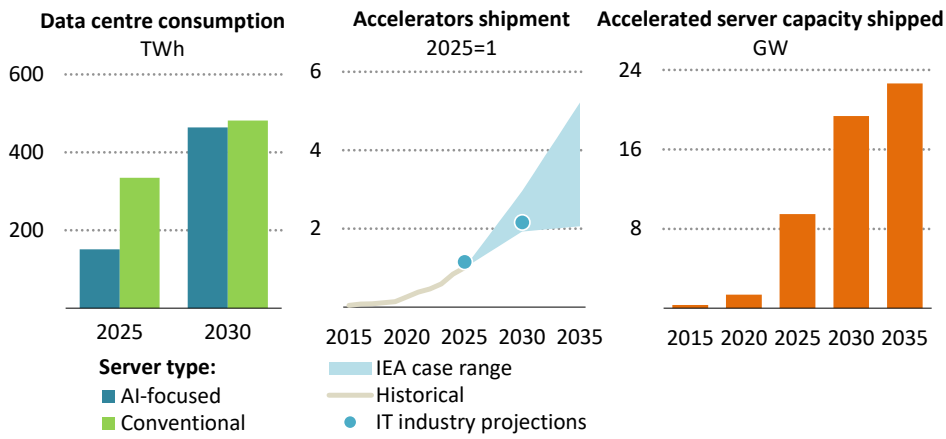
The IEA's Base Case continues to provide a sound frame of reference out to 2030, while uncertainty widens after 2030

Note: AE = advanced economies; EMDE = emerging market and developing economies. Share in electricity demand growth represents the share of data centres in total electricity demand growth by region to 2030.

Given these physical constraints and local concerns, our Lift-Off Case appears closer to our Base Case in the near-term. Beyond 2030, the outlook becomes more uncertain. On the one hand, physical constraints that shape the outlook in the near-term will start to relax over time. On the other hand, the incredibly rapid growth in AI capabilities and demand suggests that growth in electricity consumption could surprise on the upside over the longer-term after 2030. Our Headwinds Case and High Efficiency Case remain largely unchanged from our 2025 report, although they start to diverge from our Base Case somewhat later. This is due to the strong near-term momentum in data centre investment, and clear near-term trends towards more energy intensive AI modes, such as agentic AI, becoming widely adopted (see Section 2).

The IEA's data centre demand model is based on physical shipments of servers going into data centres, as these are the core electricity consuming equipment within data centres. Inputs to the IEA's model come from IT sector projections for the manufacturing and shipment of these servers historically and over the next five years. Given their high power intensity, a key variable in these projections is the annual manufacturing and shipment of so-called accelerators – specialised chips such as graphics processing units (GPUs) and tensor processing units (TPUs) that perform the parallel computations required for AI model training and use.

Figure 1.5 ▷ Key indicators for data centre electricity consumption in the Base Case



Electricity consumption by AI-focused data centres more than triples by 2030, driven by a more than doubling in accelerators shipped

Source: IEA analysis and modelling based on data inputs from OMDIA (2026).

In the IEA's Base Case, total data centre electricity consumption grows by two times by 2030. However, this aggregate number masks a much more significant increase in the electricity

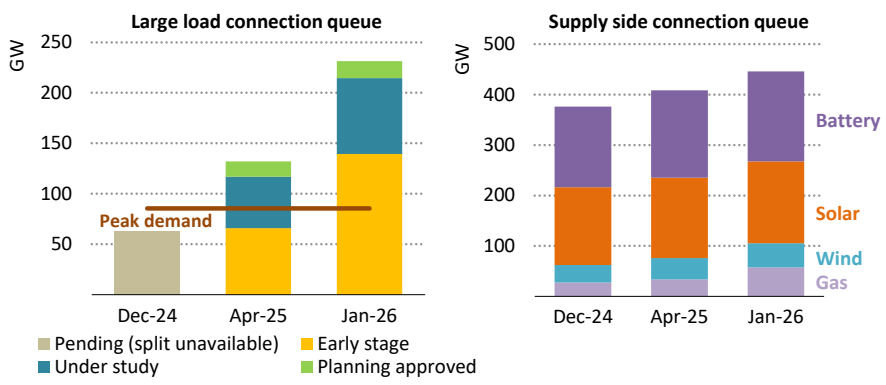
consumption of AI-focused data centres, i.e. those with large numbers of servers equipped with specialised accelerators designed for training and running AI models. In the Base Case, the electricity consumption of these AI-focused data centres increases by more than threefold to 2030, reaching around 465 TWh by 2030 (Figure 1.5).

The key driver of the electricity consumption of AI focused data centres is the shipment of AI accelerators. In the Base Case, these increase by a factor of 2.2 from 2025 to 2030, in line with the projections of shipments from the IT sector (OMDIA, 2026). In the Base Case, around 20 GW of accelerated servers are added per year by 2030. This does not include auxiliary consumption for functions like cooling, nor the addition of conventional server capacities.

Box 1.1 ▷ Understanding the boom in connection queues

Data centre connection queues have increased dramatically in many jurisdictions. However, while the speed and size of this increase are unique to data centres, large connection queues are not. They are often also large and convoluted for other types of projects. For both generation and demand projects, the system operator undertakes detailed connection studies to ensure that a project can be absorbed securely by the grid and to determine what grid upgrades may be required. Long connection queues therefore put pressure on the same regulatory system, regardless of whether the requests come from demand or generation projects.

Figure 1.6 ▷ Connection queue in ERCOT



The connection queue for large loads, mainly data centres, has grown by over 3.5 times between 2024 and 2026, but not all these projects will be realised

Source: IEA analysis based on data from ERCOT.

The Electric Reliability Council of Texas (ERCOT), Texas' electricity system operator, provides an interesting case study of this phenomenon, because it provides clear data on connection queues for both generation and load projects. In December 2024, the

connection queue for large loads, around three-quarters of which were data centres, was around 63 GW. Just four months later, it had more than doubled to 130 GW. By January 2026, it had nearly doubled again to over 230 GW. For comparison, the state's all-time high for peak demand is 85 GW. The generation connection queue has also increased, albeit much more slowly, and is dominated by non-dispatchable resources (225 GW of wind and solar). Batteries are the single largest technology in the connection queue, and natural gas projects in the queue have doubled since April 2025 to 57 GW (Figure 1.6).

Many jurisdictions are undertaking reforms to rationalise connection queues. These often involve stricter readiness tests, including larger upfront financial requirements and tighter project timelines. They may also involve "first ready, first connected" approaches or offers of non-firm grid connections. Others are experimenting with prioritisation frameworks, giving faster treatment to projects that support system needs, for example those that avoid siting near congested nodes. Several jurisdictions are also tightening timelines and accountability, with clearer deadlines for feasibility studies, greater data transparency on available capacity and, in some cases, performance incentives or penalties for both developers and system operators to keep queues moving.

These approaches are critical to enabling better planning for real load growth. The surge in project applications for data centres also highlights why project pipelines alone are not a reliable indicator of future demand, as only a limited and uncertain share of the pipeline may ultimately be realised. This creates challenges for system operators.

2 To what extent can efficiency improvements moderate AI electricity growth?

Rapid growth in data centres and AI is driving rising electricity demand in some regions. A central question is whether, and to what extent, efficiency improvements can moderate this trend. Gains are being delivered through two main channels. Software-level advances have so far yielded some of the largest and fastest reductions in energy per unit of compute, often without requiring new capital investment. Hardware and infrastructure improvements are equally essential over the medium to long term, though they typically materialise more slowly due to capital cycles and physical constraints.

Yet efficiency is only one side of the equation. Several concurrent trends in AI development are raising the compute intensity of individual tasks: advanced reasoning models generate longer chains of inference; multimodal systems integrating text, images and video demand greater compute and memory bandwidth; and agentic AI systems that plan and execute multi-step tasks autonomously can consume orders of magnitude more tokens per interaction than a simple query. By lowering costs and making AI tools more capable and accessible, efficiency gains are also fuelling broader adoption of AI services, a dynamic known as the Jevons paradox. As a result, even with continued rapid improvement in energy efficiency, total electricity consumption from AI is set to increase in the near term.

2.1 Software improvements

Software-level improvements have so far delivered some of the largest reductions in energy use per AI task, often without requiring new capital investment. Advances in model and algorithm efficiency have significantly reduced compute requirements while delivering comparable performance. Research estimates that the compute needed to reach a given level of language model performance has halved roughly every eight months (Ho et al., 2024). In production environments, combined improvements across models, software and serving infrastructure have delivered even larger gains (Elsworth et al., 2025). Smaller, more specialised models increasingly match or surpass the performance of older, large general-purpose systems while using far less energy, particularly for narrowly defined tasks (Stanford, 2025).

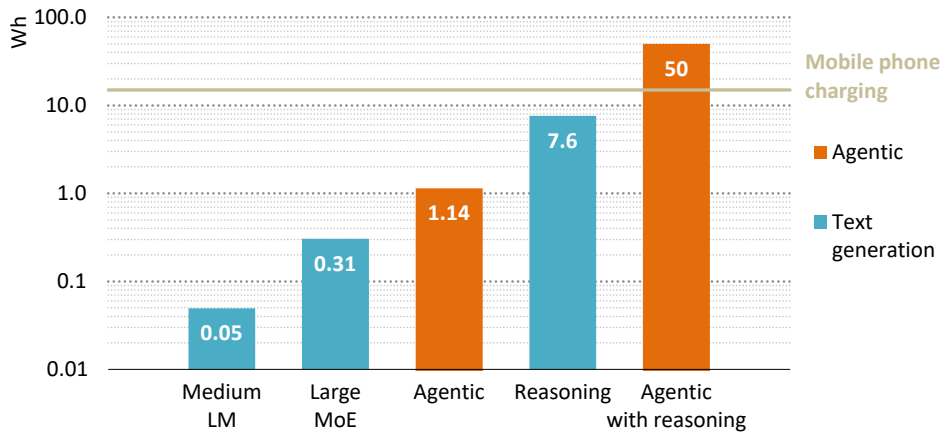
Further gains come from inference[3] optimisation and deployment choices. Techniques such as batching, caching, speculative decoding, quantised inference (running a trained model using lower-precision numeric representations than those the model was originally trained with) and hardware-aware compilation can substantially lower energy use per query. Industry reporting indicates that the cost per token of inference has fallen by orders of magnitude over the past several years through a combination of model improvements, system optimisation, and newer and more specialised hardware. Improvements in workload orchestration, including better GPU scheduling, autoscaling and higher utilisation rates, can

[3] Inference refers to the day-to-day use of deployed AI models.

reduce total data centre electricity consumption per workload, though reported savings depend heavily on the starting baseline and server types.

Energy use also varies significantly by task type. Short text queries can consume fractions of a watt-hour on modern systems, whereas generating high-resolution images may require ten times more energy per request. Video generation can require hundreds to thousands of times more energy, depending on duration and resolution. More complex reasoning models that produce longer outputs or use iterative problem-solving also consume materially more energy per query than brief text completions. Context-rich queries, where users supply documents, images or large datasets alongside a prompt, further increase per-request energy use.

Figure 2.1 ▷ Indicative inference GPU electricity consumption across different model types for text generation and agentic tasks



Model design and model choice have large impacts on electricity intensity

Notes: LM = language model; MoE = mixture of experts. Medium LM = DeepSeek-R1-Distill-Llama-70B. Large MOE = Qwen3-235B-A22B. Reasoning = DeepSeek-R1-Distill-Llama-70B with reasoning turned on. Only GPU electricity consumption is shown in this graph, as this is the metric for which measurement is most reliable. Text generation values are drawn from the AI Energy Score leaderboard, which benchmarks open-weight models under standardised conditions designed to ensure cross-model comparability rather than to represent optimised production deployments. Agentic estimates represent a moderate task profile of four to six sequential LLM calls, derived by applying token-scaled decomposition to GPU electricity consumption of DeepSeek-R1-Distill-Llama-70B with and without reasoning. "Agentic" encompasses a wide and evolving range of architectures; estimates here should be interpreted as indicative of the order of magnitude rather than as precise figures for any specific system. Reasoning and non-reasoning estimates are shown separately to illustrate the sensitivity of the results to the reasoning mode.

Source: IEA analysis based on Hugging Face (2025).

Agentic AI represents a further step change. Autonomous agents that plan and execute multi-step tasks and call on external tools can consume orders of magnitude more tokens

per interaction than a simple text completion query, significantly increasing absolute inference energy demand (Figure 2.1). Unlike conventional chat-based interactions, agentic AI applications can also decouple compute demand from direct human attention, meaning that systems may run continuously without a user actively waiting for a response.

Task-aware deployment – selecting appropriately sized models and limiting high-intensity generation to cases where it adds clear value – can therefore substantially reduce operational energy demand. In practice, however, competitive pressures and user expectations tend to favour overprovisioning of model capability, creating a persistent gap between what is technically efficient and functionally sufficient and what is actually deployed.

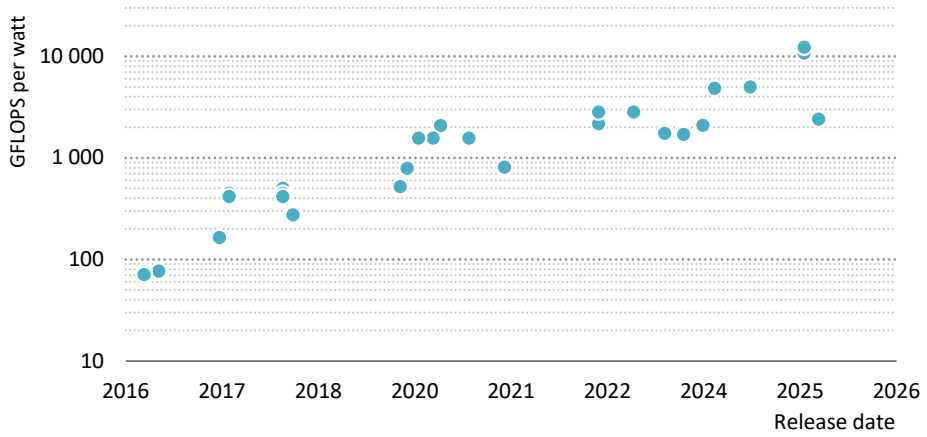
2.2 Hardware improvements

Hardware improvements remain essential for scaling AI systems and data centres over the medium to long term. Over the past decade, leading AI processors and accelerators have improved performance per watt by roughly 30-40% per year on average (Figure 2.2), although gains have been uneven and are gradually slowing as transistor scaling matures. Successive generations of GPUs and AI-specific chips have delivered large jumps in energy efficiency through architectural innovation, advanced process nodes and high-bandwidth memory integration. In addition, task-specific accelerators, such as tensor processing units and inference-optimised application-specific integrated circuits, significantly outperform general-purpose central processing units in machine learning workloads, and for certain targeted workloads, they can achieve higher performance per watt than contemporary AI GPUs. This can reduce energy use per operation by several multiples in many cases.

As core compute becomes more efficient, system-level factors increasingly determine total energy consumption. Memory access, data movement and interconnects now account for a growing share of AI system energy use, particularly in large, distributed training clusters. By improving how a model's work is divided across processors, streamlining data exchange and managing memory more efficiently, it is possible to achieve substantial performance and energy savings. Studies in high-performance computing and hyperscale infrastructure suggest that system-level optimisation, spanning scheduling, data movement, memory hierarchy, and power management, can meaningfully reduce AI workload energy consumption, depending on the baseline and scope of measurement (Patterson et al., 2022; Elsworth et al., 2025).

The efficiency of data centre infrastructure further shapes overall electricity demand. Best-in-class hyperscale facilities now achieve power usage effectiveness (PUE) ratios of approximately 1.1 to 1.2, compared with 1.5 to 1.7 in older or enterprise-operated sites. Moving from a PUE of 1.6 to 1.2 can reduce non-IT energy consumption, such as cooling and power distribution, by roughly two-thirds, with total reductions in energy use of around one-quarter. Major cloud providers report sustained low PUE performance across hyperscale fleets.

Figure 2.2 ▷ Energy efficiency of leading machine learning hardware, 2016-2025



Leading AI processors and accelerators have improved performance per watt by roughly 30-40% per year on average

Notes: GFLOPS = billion floating-point operations per second. Leading machine learning hardware includes NVIDIA's flagship data centre GPUs (P100, V100, A100, H100, B100, B300) and Google's tensor processing units (TPU v2, v3, v4, v4i, v5e, v5p, v6e, v7).

Source: IEA analysis based on Epoch AI (2026b).

However, many of the most impactful efficiency levers are locked in at the design and planning stage and are difficult or costly to retrofit. These include cooling architecture, whether the facility supports liquid cooling, and the choice of location, which determines how much of the year a site can rely on outside air for free cooling. This means that the PUE performance of a facility is largely locked in at construction, and the installed base of less efficient data centres will take years to turn over. More broadly, hardware and infrastructure efficiency improvements, while foundational to long-term sustainability, are constrained by capital investment cycles, supply chains for advanced semiconductors, and physical limits such as power density and thermal management. Consequently, hardware efficiency gains typically materialise more slowly at system scale than software-based improvements, even though both remain critical to moderating AI-driven electricity demand.

2.3 Key takeaways

The balance of AI-related energy consumption has already shifted decisively from training to inference. Yet even as per-unit efficiency continues to improve at a very rapid pace, several concurrent trends in AI development are likely to drive growing energy consumption across the sector. Rapid growth in overall AI usage could continue to offset efficiency gains. This is compounded by the growing prevalence of advanced reasoning models, now widely

deployed across major platforms, which increase compute intensity per-task by generating longer chains of inference and running multiple solution paths, thereby raising energy use per query even as underlying hardware efficiency improves.

In the near term, this dynamic is likely to be amplified by the rapid evolution of AI systems themselves. Multimodal AI models that integrate text, images, audio and video require greater compute and memory bandwidth than single-modality systems. As these models become more common across search, productivity tools and media generation, aggregate inference demand per user interaction may rise even if per-operation efficiency improves.

At the same time, the rapid deployment of agentic AI – systems that are already available in major platforms – is increasing compute intensity per-task. Unlike single-response systems, agents plan, iterate, call tools and execute multi-step tasks autonomously in response to user prompts, generating multiple internal reasoning steps, running simulations and performing verification loops before producing an output. This increases total token generation, memory usage and tool calls, raising energy consumption per completed task even if hardware efficiency improves.

Finally, the integration of AI into physical systems, including robotics, autonomous vehicles, industrial control and edge devices, will shift electricity demand beyond data centres. Embodied and real-time systems require low-latency inference, continuous sensing and on-device processing, which can multiply total deployment volumes. While individual edge models may be optimised for efficiency, the sheer scale of deployment across vehicles, factories, logistics networks and consumer devices can drive substantial aggregate electricity demand, although the potential magnitude remains uncertain.

On the hardware side, emerging architectures may offer step-change improvements that go beyond incremental gains. Three-dimensional chip stacking, which integrates memory and compute layers vertically, can substantially reduce off-chip data movement, which accounts for a growing share of AI system energy use. However, it comes at a cost of higher thermal density and manufacturing complexity. Photonic interconnects, which replace electrical signalling with lower-energy optical links, target similar bottlenecks at the chip-to-chip and rack-to-rack levels. Co-packaged optics are already entering initial production for rack-to-rack AI networking switches, with chip-to-chip deployment expected to follow from around 2027-2028, though the scale of their efficiency impact at the fleet level remains uncertain.

Taken together, these trends are not merely additive. Agentic systems that invoke multimodal models and interface with physical systems create compounding compute demand per-task, meaning the aggregate effect on energy consumption may exceed what any single trend would suggest in isolation. As a result, the IEA's electricity consumption projections for data centres incorporate substantial improvements in energy efficiency over time but still point to sharply rising electricity demand from AI in the near term.

Box 2.1 ▻ How much compute does AI actually need?

The first official per-query energy disclosures from major AI companies paint a surprisingly modest picture. Google reports that the median Gemini text prompt consumes 0.24 Wh (Google, 2025b); OpenAI puts the average ChatGPT query at 0.34 Wh (Altman, 2025); and a bottom-up estimation by Microsoft Research arrives at a median of 0.34 Wh for frontier-scale models with over 200 billion parameters (Felipe Oviedo, 2025). These figures date from 2025 and are likely already lower. Google itself reported a thirty-three-fold reduction in energy per prompt over the 12 months to May 2025, driven by advances in model architecture and hardware. As inference optimisation and next-generation accelerators continue to improve, per-query consumption can be expected to fall further.

Even taking the 2025 figures at face value and allowing for a generous blended average of 1 Wh to account for heavier workloads such as reasoning and multimodal generation, scaling to 10 billion queries per day would imply annual electricity consumption of roughly 3.6 TWh. To put this in perspective, 10 billion queries per day is comparable to the estimated volume of all daily Internet searches worldwide (Felipe Oviedo, 2025) and roughly four times ChatGPT's reported 2.5 billion daily prompts as of mid-2025 (TechCrunch, 2025). For context, this is less than 1% of the 485 TWh that data centres consume today and would represent only a small draw on the tens of gigawatts of new IT capacity that leading AI companies are individually seeking to secure.

This arithmetic raises a question that the available disclosures do not yet answer. If text-based inference at massive scale would account for such a thin slice of projected electricity demand, the bulk of planned capacity must be destined for other workloads, such as large-scale model training, video and image generation, autonomous agents running multi-step pipelines, or enterprise deployments not yet reflected in public usage figures. Yet none of the companies that have disclosed per-query metrics have offered a comparable breakdown of how their anticipated capacity will be allocated across these categories, nor how per-query consumption is expected to evolve as agentic and reasoning-intensive use cases become the norm. In practice, AI inference is diffusing rapidly across the full spectrum of digital services, from search and recommendation engines to customer support, public administration and autonomous multi-step agent workflows. Understanding the composition of future AI electricity demand, not only the cost of a single prompt today but the full mix of increasingly complex and embedded workloads driving tomorrow’s infrastructure build-out, would give energy planners and policy makers a much stronger basis for anticipating what lies ahead.

3 Are new data centres moving to locations beyond existing clusters?

The world's data centre capacity is heavily concentrated in clusters, driven by factors such as local demand for IT services, the availability of digital and energy infrastructure, and tax and other incentives. However, the clustering of data centres creates significant challenges for utilities and grid operators. While data centres accounted for 1.5% of global electricity consumption in 2025, they accounted for as much as 20-30% of electricity demand in the states, metropolitan clusters or regions where these clusters exist, compounding the challenge of meeting electricity demand reliably, quickly and affordably.

For example, in Alberta, Canada, government efforts to attract data centres led to connection requests totalling 16 GW, which exceeded the province's peak demand. This in turn prompted the grid operator to impose a temporary 1.2 GW cap to protect system reliability (AESO, 2025).

It is therefore vital to explore how the geographic distribution of data centres is evolving over time. Specifically, is new data centre capacity set to continue expanding within existing hubs, or will it become more widely distributed geographically?

Our analysis reveals that clusters are set to remain a key feature of the data centre landscape. Existing hubs are poised to keep growing, while new data centre clusters are also emerging. Trends differ by region, with capacity additions in advanced economies taking place in both existing and new clusters, while emerging and developing economies are experiencing their first wave of clustering.

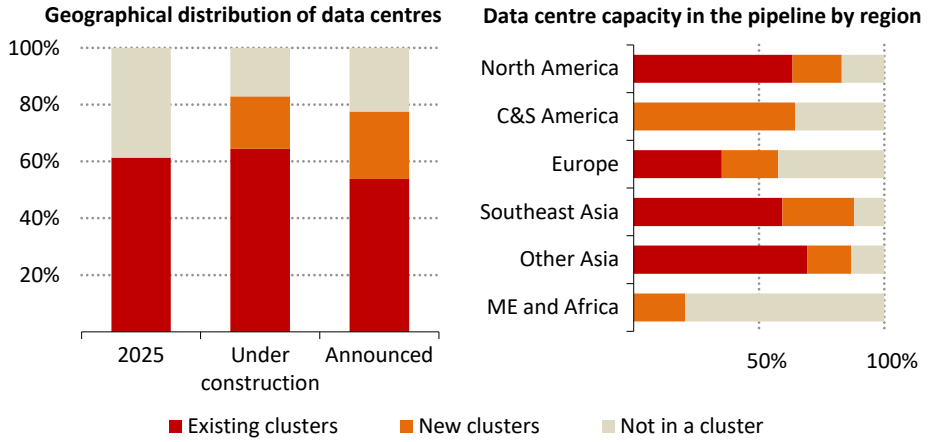
3.1 Global data centre clustering patterns

In analysing clustering patterns, we assessed both data centre capacity that is already under construction and capacity that has been announced but is not yet under construction, which together we refer to as capacity in the pipeline. Announced projects do not always break ground or lead to completed projects. Pipeline estimates are derived from third-party project databases and include only projects that have progressed to a credible stage.

Globally, existing data centre capacity is primarily concentrated in major hubs. Established clusters in the United States, Europe and China, where data centres benefit from robust grid infrastructure, reliable power supply, skilled labour and proximity to demand, represent more than half of global capacity today. For the purposes of this analysis, data centre clusters refer to collections of facilities with a combined total designed capacity of at least 500 MW within any 100 km radius.

The analysis reveals that, at the global level, over half of all data centres in the pipeline are situated in existing clusters, including around 65% of the data centres under construction and nearly 55% of those announced but not yet under construction. At the same time, a growing share of new capacity will form entirely new clusters (Figure 3.1).

Figure 3.1 ▷ **Clustering trends of data centres globally and across key regions**



IEA. CC BY 4.0.

Future capacity additions will primarily be located in existing or new clusters

Notes: C&S America = Central and South America; ME = Middle East. Data centre clusters here refer to collections of facilities with a total designed capacity of 500 MW and grouped according to the DBSCAN clustering algorithm using a neighbourhood radius of 100 km.

Source: IEA analysis based on BNEF (2025).

Taken together, the share of data centre capacity located in clusters, both existing and new, is therefore set to rise from around 60% today to over 65% by the time facilities currently under construction come online. Factoring in all announced projects, that share rises further to more than 70%, although it is again important to highlight that not all announced projects actually come to fruition.

Regions where data centres are already an established sector (mostly advanced economies and China) are seeing growth in existing hubs and the establishment of new ones. Meanwhile, in developing economies, new clusters are beginning to emerge. Across regions, clustering is a common feature, with the exception of the Middle East and Africa.

Europe's pipeline shows the greatest geographic diversification. Current data centre clusters in Europe include Frankfurt, London, Amsterdam, Paris and Dublin, often referred to as FLAP-D. Within the data centre pipeline, capacity is set to be relatively less clustered in Europe than in the United States and Asia. Nonetheless, the share of total data centre capacity in clusters in Europe is set to exceed 50%, including new clusters that will be created by future capacity additions.

Box 3.1 ▻ Physical security as an emerging determinant of data centre location

As AI becomes increasingly integral to economies, data centres are now regarded as vital strategic assets with security implications. Conventional security risks to data centres range from cybersecurity breaches to supply chain vulnerabilities – including those related to energy infrastructure and critical mineral supply. This was explored in IEA's 2025 special report *Energy and AI*.

More recently, as the 2026 conflict in the Middle East has highlighted, conventional physical threats have emerged as a key risk to data centre infrastructure. Data centres have joined energy and telecommunications infrastructure as potential sources of vulnerability. This risk is especially acute for large data centres that involve capital expenditures to the tune of billions of dollars. For example, some of the largest data centres in construction today have USD 20-50 billion in planned investment outlay. This might have implications on the locational considerations of data centres – which in turn have implications on energy demand in those regions.

This issue might have additional considerations as 60% of the world population today resides in countries that account for just 9% of data centre capacity. As a result, countries have put in place policies such as tax incentives and data localisation rules to encourage investment into new data centre capacity in their jurisdictions. With physical security of data centres as a consideration, there could be implications for some regions seeking to attract new investment.

The IEA is tracking this issue closely and will continue to factor in traditional and non-traditional security threats into its analysis on data centre locations and the resulting energy demand.

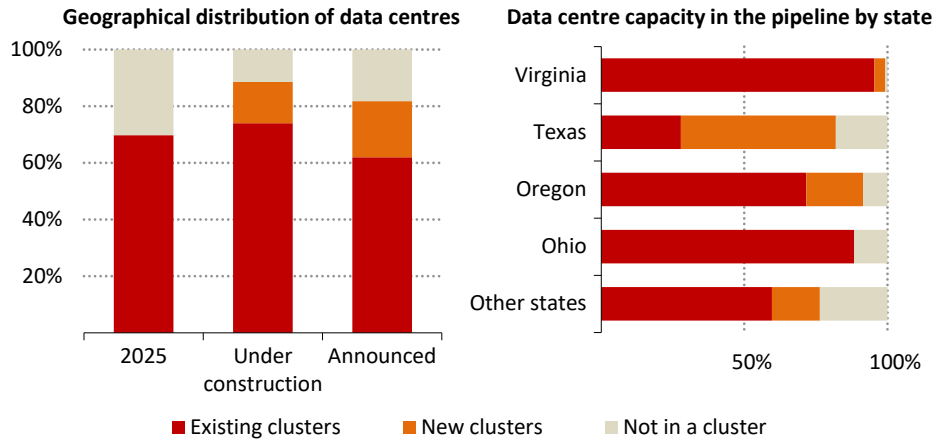
3.2 Data centre clustering in the United States

The United States is the world's largest data centre market, accounting for about half of total global installed capacity as well as announced global capacity additions. North America is the most clustered region, along with Asia. Factoring in the data centre capacity in the pipeline, the United States has the largest capacity in clusters, both existing and new. There are also reports of very large volumes of interconnection requests in existing clusters in the United States. For example, ERCOT, the grid operator in Texas, has received interconnection requests of 160 GW from data centres (Utility Dive, 2026), which is twice the existing global capacity of hyperscale and colocation data centres. However, it is worth noting that not all interconnection requests result in approvals and subsequent construction, and some of the requests may be speculative in nature (see Box 3.2).

Our analysis suggests that future capacity additions in the United States will be increasingly concentrated in existing and new cluster areas. However, national statistics hide significant regional differences, with a few states accounting for most of the capacity growth (Figure 3.2). For instance, in states such as Virginia and Ohio, nearly all future capacity will

expand existing clusters. In particular, Northern Virginia, the world's most significant area for data centre development, is set to see concentrated growth, with practically all projects in the pipeline located within or immediately adjacent to the established cluster, benefiting from the extensive communications infrastructure already in place. Texas, however, will see most additional capacity concentrated in new clusters of different sizes near cities such as Dallas, Houston, San Antonio and Austin, where land availability, competitive power supply and growing technology sectors are attracting investment. Emerging clusters in Texas also extend beyond population centres into areas close to energy resources and supportive infrastructure.

Figure 3.2 ▷ Clustering trends of data centres in the United States and in the main US states



IEA. CC BY 4.0.

Texas will experience significant growth in new clusters, whereas existing clusters are expected to capture most capacity additions in states such as Virginia and Ohio

Note: Data centre clusters here refer to collections of facilities with a total designed capacity of 500 MW and grouped according to the DBSCAN clustering algorithm using a neighbourhood radius of 100 km.

Source: IEA analysis and modelling based on BNEF (2025).

Box 3.2 ▷ Are large AI data centres in the United States being built further away from cities?

Geospatial analysis of data centres around the world conducted for the IEA's 2025 special report *Energy and AI* revealed that data centres tend to cluster in and around urban areas (IEA, 2025a). As urban areas are already major centres of energy demand, new data centres can present a challenge to utilities.

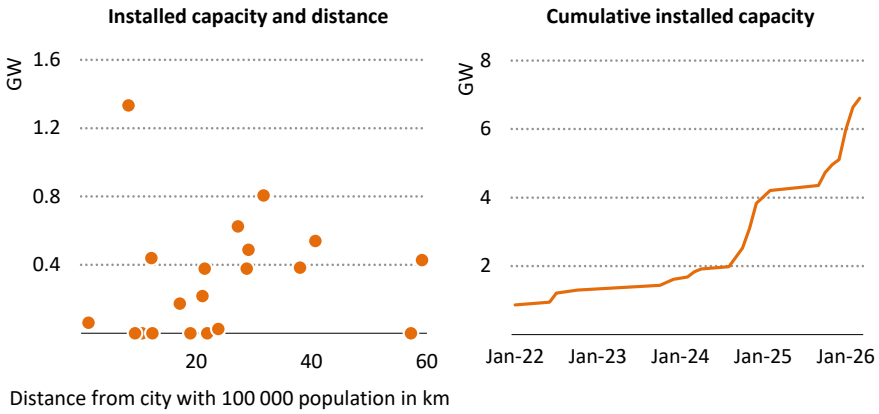
In recent years, a handful of very large AI-focused data centres have been announced by major technology companies in the United States. These data centres are designed to

train and run AI models. They tend to have a much larger associated capacity, and therefore electricity consumption, than typical colocation or enterprise data centres and are equipped with the latest hardware and technology (see section 4). Given the demands of competition in the AI industry, they are also expected to be constructed rapidly.

To understand their locational characteristics and the progress of their build-out, the IEA conducted a geospatial analysis of 21 such AI-focused large data centres. These data centres range 600 MW to 5 GW in nameplate capacity and are each estimated to consume as much electricity as 600 000 to 5 million households once complete.

IEA analysis revealed that 19 of these 21 hyperscale AI data centres in the United States are located within 65 km of the centre of a city with a population of at least 100 000 people. Their construction is also progressing rapidly, having tripled in floor space and estimated capacity within 18 months to exceed 6 GW cumulatively. The analysis highlights that while AI data centres display a greater degree of locational flexibility than traditional data centres, they still tend to cluster close to urban areas. Once completed, in many cases, the power demand of these data centres will dwarf that of the adjacent urban areas. This fact, combined with their rapid development timelines, can put pressure on local grids and utilities.

Figure 3.3 ▷ **Locational characteristics and construction progress of AI-focused hyperscale data centres in the United States**



IEA. CC BY 4.0.

19 of the 21 large AI data centres under construction in the United States are within 65 km of a city with a population of at least 100 000

Note: Installed capacity is estimated from the floor space that has completed construction at the locations of dedicated AI data centres in the United States.

3.3 Key takeaways

Based on this analysis, the geographic concentration of data centres appears set to persist and intensify, with existing clusters continuing to attract a large share of future capacity additions. Among projects located outside current hubs, an increasing share will form new clusters rather than being widely dispersed, reflecting the sector's tendency towards agglomeration. As the average size of data centres increases, fewer facilities are able to add large capacities near existing hubs, which supports the creation of new clusters. While patterns vary by region, the global trend points towards a world in which data centre capacity remains spatially concentrated, albeit across a growing number of hubs.

4 How is AI changing data centre design, and what does it mean for the mix of energy technologies in them?

There are many types of data centres, ranging from small enterprise facilities to those that provide digital services, such as cloud applications or web hosting, and large hyperscale facilities. Until now, very few data centres have been specifically built to serve AI model training and use. However, booming demand for AI services is leading to a new generation of data centres specifically designed and optimised for AI, so-called "AI factories". While these AI factories can vary in their technical specifications and business models, they face a shared set of technical challenges and rely on an emerging, highly specialised technology stack. Mapping their development is important for understanding how this type of data centre might integrate into the grid in the future, along with the implications for energy technology supply chains and possible innovation spillovers for the rest of the energy system.

AI factories are characterised by soaring power densities and sharp swings in IT load

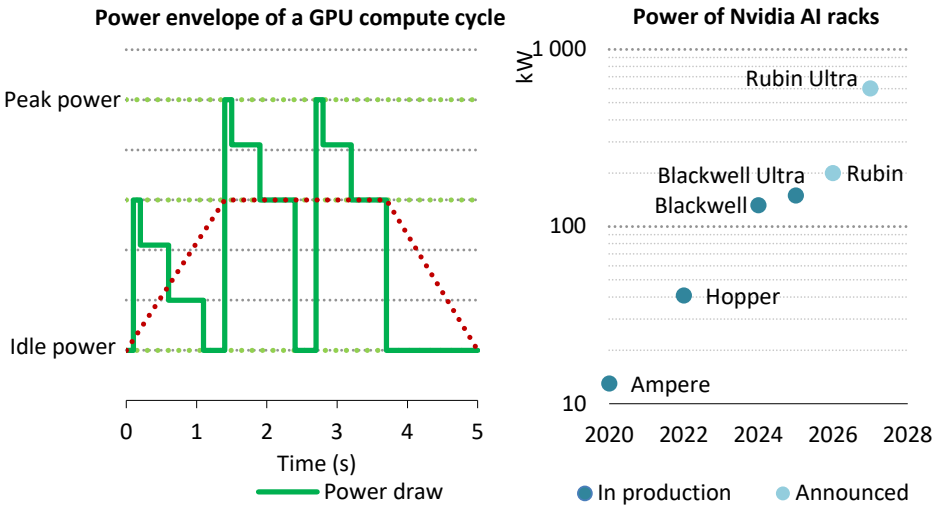
Two trends make AI factories different from traditional data centres. First, they are becoming increasingly power-dense, meaning they have higher power demand per unit of space. Second, they tend to house much more variable electrical loads. The first trend is driven by the growing power rating of chips (i.e. GPUs), as well as by the clustering of more chips into a single rack.[4] Multiple chips are networked tightly together to function effectively as a single computer. This trend is driven by the need to increase computing power and memory while minimising both electrical losses and data latency. The power requirements of the associated auxiliary equipment must also increase to maintain optimal conditions for the higher power density of the IT equipment. For instance, since more heat is generated at the rack level, cooling equipment must scale accordingly.

For example, Nvidia's Ampere architecture from 2020 had a rated power of around 400 W per chip and clustered 32 chips per rack, resulting in a power density of around 13 kW per rack, including central processing unit, networking and storage power. The current Blackwell architecture has a rated power of 1 000 W per chip and 72 chips per rack, with a power density of 130 kW per rack. The announced Rubin architecture will reach a power density of 600 kW per rack and will pave the way for 1 MW per rack density. To put this in perspective, with the announced Rubin architecture, a box the size of a household fridge would have a peak power draw equivalent to that of around 65 households. This rate of increase in power density has no historical precedent in electrical engineering.

[4] A rack is a standard-sized tray for IT equipment in a data centre and has a floorspace footprint of less than 1 m^2.

One consequence of GPUs being networked together and operated synchronously is that their workloads, and therefore their power consumption, are also synchronised. During AI model training and use, there are typically periods of intense computation, followed by periods of less computationally intensive data exchange and periods of computational downtime. Because computation requires power, this variability leads to large swings in IT power demand. These swings occur across timeframes ranging from microseconds to seconds to several minutes. As the power density of AI hardware increases, the absolute size of these swings also increases. Software-based mitigation strategies can help smooth these swings, but they may result in more wasteful electricity consumption. Depending on how data centres of this kind are configured and connected to the grid, these swings in power use may have implications for grid stability, potentially leading to voltage instability and harmonic distortion on local grids.

Figure 4.1 ▷ Power envelope of GPU workload and power density of data centres by first operational year



AI factories are characterised by soaring power densities and sharp swings in IT load

Note: Desired power profile refers to the time profile of electrical load that can be absorbed by power supply equipment.

Source: IEA analysis based on Nvidia (2025).

Implications for power delivery, storage and cooling technologies

This change in architecture has major implications for the technology stack needed by next-generation AI factories, which may in turn have broader spillovers for the energy sector.

For power delivery, the key challenge is to minimise the space and material footprint of power supply equipment, minimise capital costs and lower electricity conversion losses

power losses[5] and generate substantial amounts of heat, which would need to be managed in addition to the heat produced by the IT equipment. Power delivery infrastructure would also occupy as much space within the data centre as the compute infrastructure itself. This would result in facilities that are expensive, less energy efficient, wasteful of critical materials such as copper wiring, more complicated to build and operate, and less scalable.

Some of these challenges can be addressed by increasing the voltage from the standard 400 volts to 800 volts, and possibly even higher, and by shifting from alternating current (AC) to direct current (DC). Doubling the voltage means that the size of a copper cable for power distribution can be reduced by a factor of four while transmitting the same amount of power, while moving from AC to DC reduces the number of power conversion steps from the grid to the chip, limiting conversion losses. This architecture can build upon innovations in power conversion and control technologies pioneered by the electric vehicle and renewables industries. Critical technologies include:

- **Rectifiers and solid-state transformers**. These devices can convert grid AC power to DC power, control voltage for distribution across the data centre and provide power quality. Solid-state transformers have a 70% lower material footprint than traditional transformers, much higher flexibility and precision in terms of power control, and higher power conversion efficiency. However, current designs often face higher thermal management requirements and lower peak efficiency compared to high-efficiency liquid-filled transformers. Solid-state transformers are also still at an early phase of commercialisation.
- **Power electronics, or the use of semiconductors for efficient power conversion and control**. This technology enables voltage step-up or step-down, dynamic load balancing to maintain stable voltages, and microsecond-sensitive fault management, which is essential for safety in DC architectures and critical for protecting the extremely expensive IT equipment in AI factories. Critical technologies include silicon carbide and gallium nitride based power electronics.

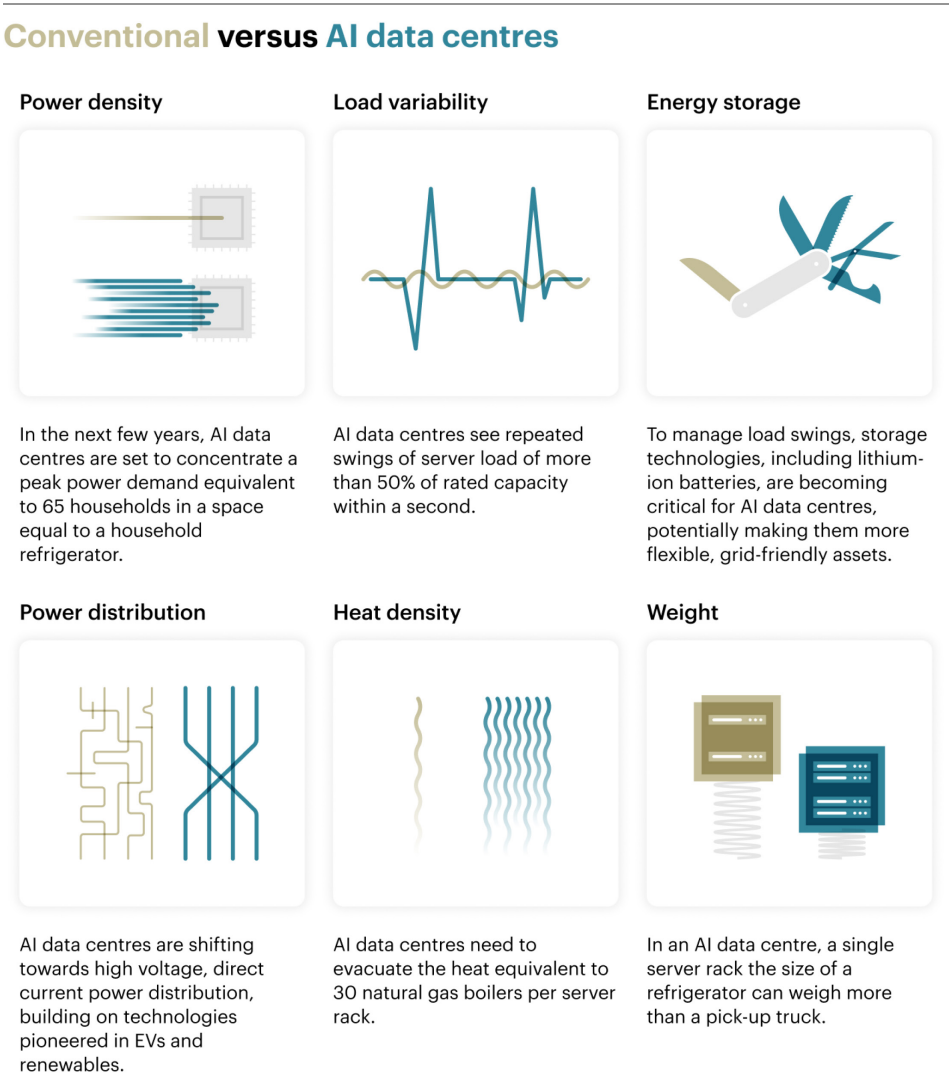
Power storage technologies are also critical for next-generation data centres due to the load variability of increasingly large and powerful GPU clusters noted above. For example, AI data centres have seen oscillations of tens of megawatts (around half of rated capacity) in less than a second (Elevate Energy Consulting, 2025). If left unmitigated, this can damage equipment within a data centre and cause voltage and harmonic disturbance at the grid level.

To mitigate these swings, storage is needed across multiple timeframes, from milliseconds to minutes or more. For millisecond timeframes, electrolytic capacitors can deliver or absorb power, smoothing the extreme cycling speeds of GPUs while protecting battery systems from the degradation caused by frequent charging and discharging. For longer timeframes, lithium-ion batteries, potentially complemented by sodium-ion batteries in the future, can

[5] Resistive power losses are the portion of electrical power that is converted into heat when current flows through any component with resistance.

smooth GPU ramp cycles, provide backup power in the event of an outage of the prime power source and enable a data centre to provide flexibility to the grid. This can help facilitate grid connections and make data centres more grid-friendly assets, protecting the grid from the voltage and harmonic disturbances that unmitigated load swings within a data centre could otherwise cause.

Figure 4.2 ▷ The changing architecture for AI factories



AI data centres are heavier, more power dense, have greater variability of loads, and emit more heat compared to conventional data centres

The increase in power density within AI factories inevitably results in additional heat that must be removed to ensure that servers remain within their operating temperature range. High-density AI racks increasingly require liquid cooling solutions, which leverage the fact that liquids can carry much more heat than air. Liquid cooling solutions could also make it easier to recover waste heat.

4.1 Key takeaways

The challenges and innovations described in this section relate to the next generation of AI-optimised data centres. They are not necessarily relevant for all data centres, many of which will continue to use more traditional architectures or adopt hybrid approaches, with some computing space dedicated to AI-optimised infrastructure and some occupied by more traditional computing infrastructure. Even for AI-optimised infrastructure, transitional architectures centred around the delivery of 400-volt AC power will continue to be used.

However, if future generations of ultra-high-power-density servers are deployed at scale in the next few years, the cost and technical limitations of traditional architectures will create strong incentives to change approaches. The size of AI factories means that these shifts in the technology stack could have broader implications for the energy sector. We explore several here, while highlighting that this is a fast-moving field that requires further analysis and closer co-operation between energy technology providers and the technology sector.

Table 4.1 ▷ Supply chain risks in the emerging power technology stack for AI data centres

Component	Relevance for data centres	China market share
Gallium nitride-based power electronics	Gallium nitride semiconductors are critical for stepping down voltages to the rack and chip levels and controlling voltages to protect IT equipment.	China provides 99% of refined gallium supply.
Silicon carbide-based power electronics	Silicon carbide semiconductors are used in rectifiers to convert grid AC to DC current and step it up to 800 V, and in DC-DC voltage step-down to the rack level.	China controls around 50% of supply. Price pressures driven by the rapid growth of Chinese producers led to the bankruptcy of the leading US producer in 2025. China controls 95% of the market for high-purity silicon.
Ferrite cores for solid-state transformers	Solid-state transformers are emerging as critical for managing the complex power conversions across data centres.	China controls around 60% of the global market for ferrite cores.
Lithium-ion batteries/ sodium-ion batteries	Batteries are becoming increasingly essential to manage power variability within AI factories.	China controls around 99% of the manufacturing capacity for lithium iron phosphate battery (LFP) batteries (by far the most common lithium-ion battery energy storage system technology for stationary storage). Manufacturing capacity for sodium-ion batteries is also almost exclusively in China.

Sources: IEA analysis based on HGBR (2025), IEA (2026), and (IEA, 2025c).

For several of the technological solutions outlined, there are supply chain concerns, whether in terms of technology readiness, supply chain bottlenecks or supply chain concentration. While the underlying technology of power electronics is well proven, large-scale solid-state transformers are not yet commercially available, and their deployment is limited to niche applications. Bottlenecks across the supply chain include grain-oriented steel; a severe global shortage of specialised power system engineers, site technicians and skilled labour, for example in transformer manufacturing; and limited production capacity for silicon carbide-based power electronics. Several aspects of the supply chain also show a high degree of geographic concentration, notably in China (see Table 4.1).

The extremely rapid shift towards new, more power-intensive data centre designs is set to test the ability of related supply chains to scale. At the same time, this transition in power architecture is likely to bring broader benefits to the energy system by accelerating the commercialisation and innovation of electrical technologies such as solid-state transformers. The shift towards actively managing variable power loads within the data centre through the deployment of battery storage systems with grid-interactive capabilities could also contribute to making data centres more grid-interactive and flexible in the future, providing both stability and flexibility to the network.

5 How are data centre operators procuring the electricity they need?

Growing electricity demand from data centres is pushing operators to secure rapid and cost-effective access to power while also meeting corporate emissions reduction and clean energy targets. To meet these objectives, operators are adopting a range of procurement strategies.

One of the most prominent approaches is the corporate power purchase agreement (PPA), which allows companies to contract electricity directly from generators and lock in long-term electricity supply. In the past few years, in addition to PPAs with wind, solar and hydro generators, major data centre operators have also signed PPAs to secure baseload low-emissions electricity from nuclear and geothermal sources. At the same time, natural gas is attracting greater attention as a dispatchable and cost-competitive source of electricity that can support rising power demand from data centres, particularly in the United States.

5.1 Procurement strategies

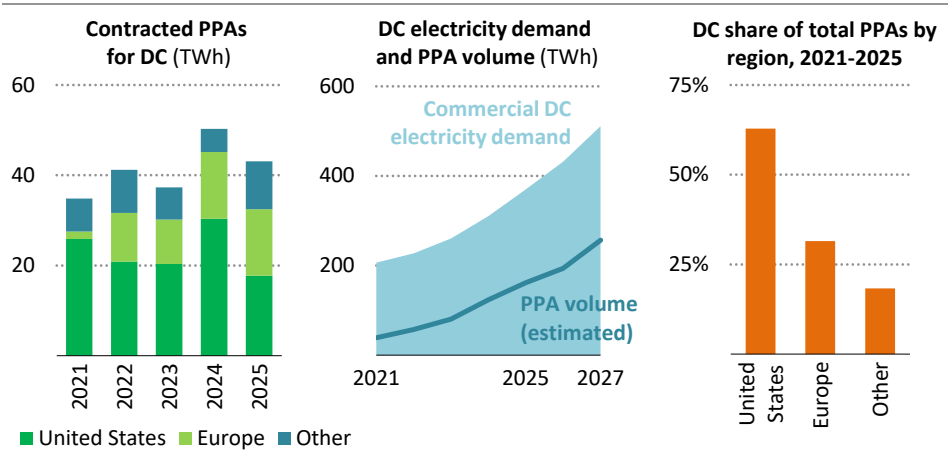
PPAs have become one of the main procurement strategies for data centre operators since they allow them not only to buy electricity at a fixed price to meet demand but also to procure electricity from specific sources to help companies achieve their clean energy goals. However, even when data centres purchase low-emissions electricity via PPAs, that electricity may not be located near the data centres and may not be generated at the specific times when it is needed. Therefore, the actual power use by data centres is often met by the grid using other sources, such as natural gas or coal. This results in a physical electricity mix that often differs from the procured, or "financial", electricity mix.

To account for this discrepancy, some data centre operators are moving from annual to hourly matching of low-emissions electricity. This approach seeks to ensure that electricity demand in each hour of the year is met with low-emissions generation, rather than offset through annual volumes. Achieving hourly matching often requires pairing variable renewables, such as wind and solar PV, with storage systems, typically battery storage, although other dispatchable low-emissions sources, such as hydropower, nuclear or geothermal, can also contribute. Technology companies such as Google and Microsoft have set targets to achieve hourly matched supply with low-emissions electricity by 2030, signalling momentum behind this approach. For example, Google developed partnerships with NV Energy and Xcel Energy to bring new low-emissions electricity under the Clean Transition Tariff and Clean Energy Accelerator Charge.

IEA analysis indicates that portfolios achieving up to 80% hourly matching using wind, solar PV and batteries can be broadly cost-competitive with annual matching portfolios (IEA, 2025a). While achieving nearly full hourly matching typically involves a cost premium compared with annual PPAs, mainly due to the need to incorporate storage technologies, it reduces exposure to electricity price volatility and offers protection for consumers against high market prices.

In recent years, data centres have increasingly adopted PPAs that specifically include renewable energy sources. Agreements covering more than 40 TWh were signed in 2025, with around 40% of this volume in the United States (Figure 5.1). Over the past five years, about 65% of renewable PPAs signed in the United States have been specifically dedicated to data centre operations. Renewable PPAs for data centres have also expanded in Europe, rising from about 1.5 TWh in 2021 to nearly 15 TWh in 2025, with data centre PPAs now accounting for more than 30% of the total volume of electricity contracted through corporate renewable PPAs. The trend is less prominent, but still significant, in the rest of the world, with about 20% of renewable PPAs dedicated for data centre operations. By 2027, electricity contracted through renewable PPAs is projected to be equivalent to 50% of total commercial data centre electricity demand, a significant increase from less than 20% in 2021.

Figure 5.1 ▷ Renewables PPAs for data centres



Renewables PPAs have been growing in recent years, with the United States accounting for around 40% of the share globally today

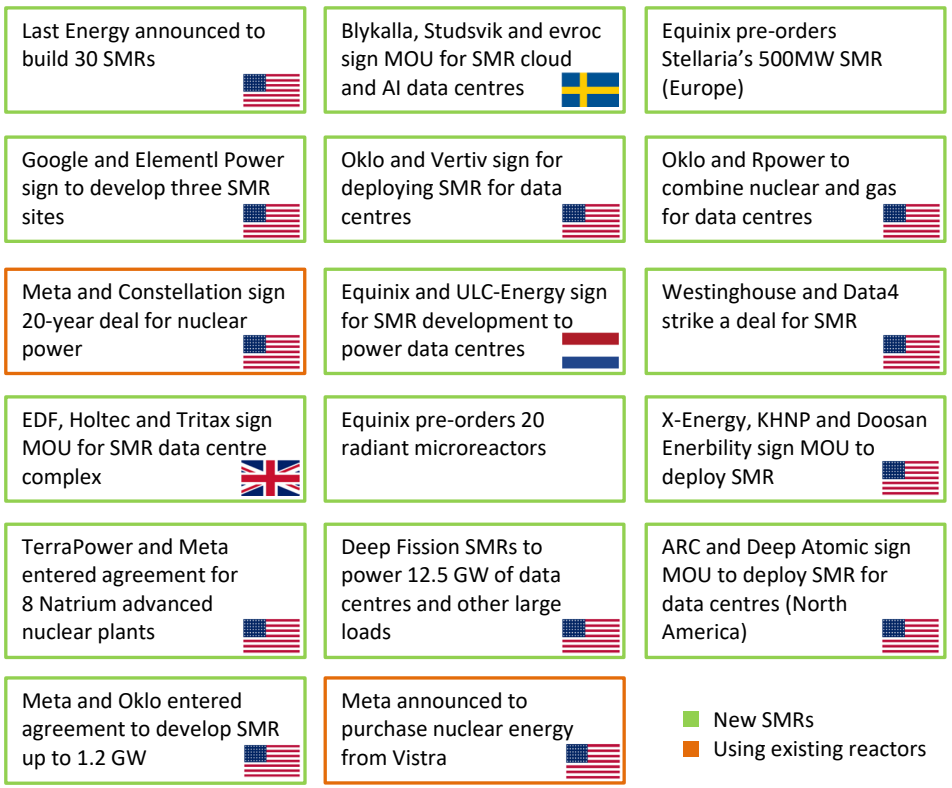
Notes: PPA = power purchase agreement; DC = data centre; PPA volumes in the left chart are shown by the year signed. DC electricity demand refers to commercial, i.e. hyperscale and colocation, data centres.

Source: IEA analysis based on S&P Global (2026).

As baseload low-emissions technologies, nuclear and geothermal power have gained increasing attention in recent years. For example, in 2024, Microsoft and Constellation announced a PPA to procure about 835 MW of nuclear capacity through the restart of Three Mile Island Unit 1, with the restart currently targeted for 2027. In 2025, Amazon Web Services entered into a PPA with Talen Energy to supply its data centres with up to 1 920 MW of power from Talen's Susquehanna nuclear power plant. Major technology companies are also supporting the development and commercialisation of innovative low-emissions technologies, particularly small modular reactors (SMRs) and next-generation geothermal. Although the type of agreement varies, ranging from corporate partnerships to PPAs, around

25 GW of SMR capacity had been announced by the end of 2024 to supply data centre operations (IEA, 2025a). A further 20 GW was announced in January 2026, raising the total planned capacity to 45 GW, while many other announced projects have yet to specify capacity. The first commercial SMRs are expected to start operating around 2030. Meta has signed multiple agreements with different suppliers, including Oklo, TerraPower and Vistra, totalling more than 6 GW of nuclear capacity to support its data centre operations. Regarding geothermal energy, Google and Ormat Technologies signed a PPA in 2026 aimed at supplying up to 150 MW of new geothermal capacity. Next-generation geothermal technology also has the potential to provide baseload electricity to data centres. Meta and Sage Geosystems have signed a PPA to procure up to 150 MW of next-generation geothermal capacity.

Figure 5.2 ▷ Recent announcements and agreements related to the procurement of nuclear energy for data centres



About 20 GW of SMR capacity was announced to supply data centre operations, bringing the total up to 45 GW

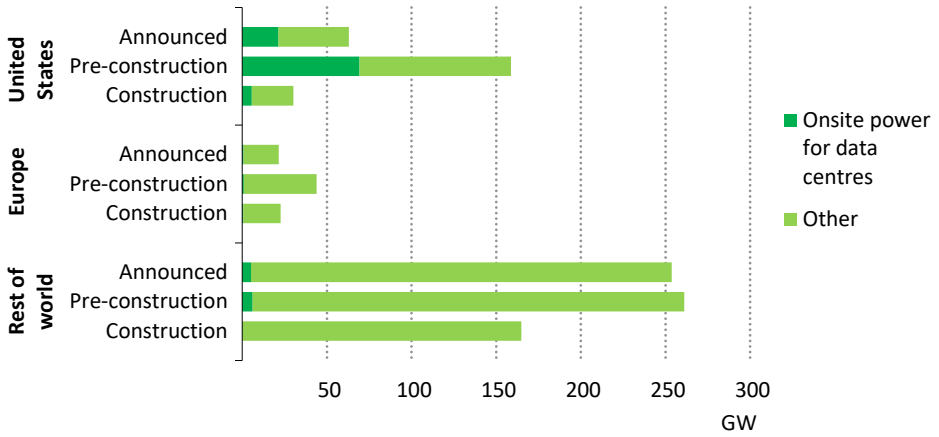
Notes: SMR = small modular reactor; MOU = memorandum of understanding. The figure includes announcements and agreements made from January 2025 to January 2026.

5.2 New natural gas capacity will help meet surging electricity demand

While the substantial growth in renewables and low-emissions electricity, such as nuclear and geothermal, reflects strong efforts by data centre operators to align growth with their clean energy objectives, fossil fuels continue to play a significant role in meeting the sector's rapidly rising electricity demand.

Over 100 GW of new natural gas capacity is currently planned as dedicated supply for data centres through onsite generation, including projects that have been announced or are in the pre-construction or construction stages (Figure 5.3). Most of this new capacity is in the United States, where more than one-third of all new global gas-fired capacity is intended to supply onsite electricity for data centre operations. The trend is different outside the United States, where only around 13 GW of onsite gas capacity has been announced or is under development. Even so, it remains uncertain how much of the global pipeline will ultimately be built (see section 6).

Figure 5.3 ▷ Global natural gas-fired capacity planned or under construction



More than one-third of new gas capacity is onsite power for data centres in the United States, accounting for about 10% of planned global gas capacity additions

Note: Gas projects in the figure include oil and gas dual fuel projects.

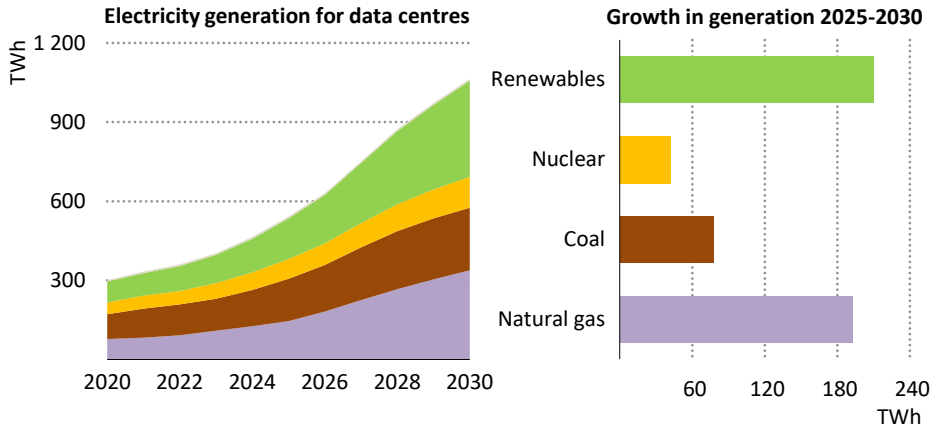
Source: IEA analysis based on GEM (2026).

5.3 Key takeaways

With global electricity generation for data centres projected to double to over 1 000 TWh by 2030, generation from all sources is set to increase. Renewables are poised to account for a substantial share of this growth, expanding by more than 200 TWh from current levels and reaching 360 TWh by 2030, at which point they would account for more than one-third of the sector's total generation. The growing pipeline of natural gas-fired projects means that

gas generation is also projected to increase strongly, particularly in the United States. It is set to more than double to around 340 TWh by 2030, pushing its share of the global data centre electricity mix to around 30%. Together, renewables and natural gas are set to account for over 65% of all electricity produced for data centres by 2030.

Figure 5.4 ▷ Global electricity generation for data centres, 2020-2030



Electricity generation for data centres is set to double to more than 1 000 TWh by 2030, driven by growth in renewables and a doubling of gas-fired generation

Nuclear power is also set to play an important role, providing low-emissions baseload generation and benefiting from increasing support from several major technology companies. Nuclear sources are projected to increase their generation for data centres from around 75 TWh today to nearly 120 TWh in 2030. Meanwhile, coal continues to supply electricity to data centres, primarily in China, and is projected to account for about 20% of the global data centre electricity supply in 2030.

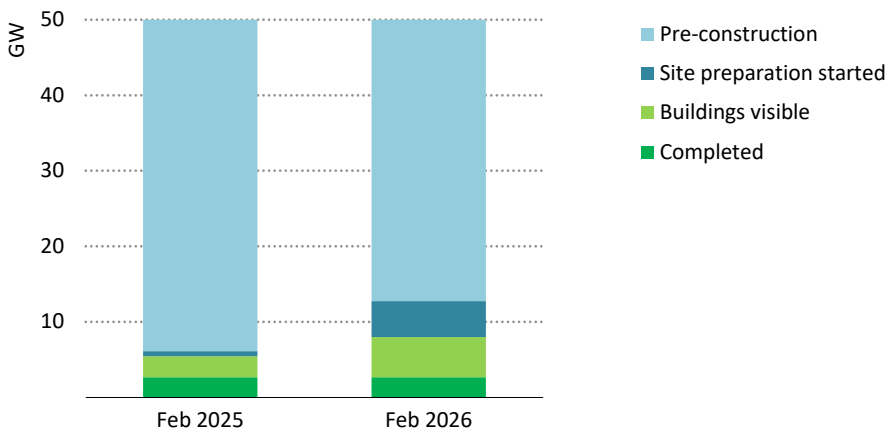
6 Why are some data centre developers considering onsite generation?

Today, most data centres are connected to the grid. However, given lengthy connection queues, onsite electricity generation is gaining interest. Onsite generation has the potential to provide faster access to power for data centres in regions with constrained grids. However, because of design complexity given the high requirements for reliability, as well as increasing component costs and growing backlogs in the gas turbine supply chain, onsite power can raise costs for developers and increase project timelines.

To date, the largest data centre project using onsite generation as its primary source of electricity is in Tennessee (United States), which initially deployed more than 400 MW of onsite natural gas turbines to bridge supply in 2024 and 2025. This was later scaled down as a 150 MW grid connection (with 150 MW of backup battery storage) came online. However, the pipeline of data centre projects with onsite generation is expanding rapidly, with larger, gigawatt-scale projects currently under development.

So far, onsite generation projects are concentrated almost entirely in the United States and, as with the Tennessee project, are based largely on natural gas-fired technology. Around 5.6 GW of onsite gas-fired capacity for data centres is currently under construction, more than 70 GW is in development and an additional 21 GW has been proposed (GEM, 2025).

Figure 6.1 ▻ Satellite-based tracking of 50 GW of onsite power plant capacity for data centres by construction stage, 2025–2026



Tracking of 50 GW of onsite power sites reveals that most projects remain in the pre-construction phase, although site preparation has been advancing

Source: IEA analysis based on GEM (2026).

Since the IEA's 2025 *Energy and AI* report, new analysis of onsite power projects for data centres using satellite imagery has shed further light on the evolution of this pipeline. In total, 63 power plants were trackable, accounting for 50 GW of the more than 75 GW pipeline. The analysis shows that construction has not yet started at most sites, although some projects have advanced relatively rapidly. In February 2025, nearly 90% of identified capacity remained at the pre-construction stage, while only 1% had begun site preparation, 6% had visible structures and 5% were nearing completion. By February 2026, pre-construction capacity had declined to 75%, while capacity under site preparation rose to 10% and capacity with visible structures to 11%. Completed capacity remained at the same level seen in early 2025.

6.1 Key considerations for onsite generation

In some areas, AI infrastructure is expanding faster than electricity networks can connect new load. Lengthy grid connection queues and grid expansion timelines are colliding with data centre delivery cycles of one to two years, turning electricity availability into a binding constraint on deployment.

As a result, an increasing number of developers are planning data centres with onsite power generation. Some are pursuing hybrid models, in which onsite generation and storage complement a grid connection. Others are proposing onsite generation as a temporary "bridge to power" while projects wait for a grid connection, and plan to later repurpose these assets for backup or flexibility services once full grid access becomes available. Another solution, as developers seek to shorten timelines, has been to co-locate new data centre loads near existing grid-connected power plants.

As noted, natural gas-fired technologies dominate the onsite generation pipeline, particularly aeroderivative gas turbines and large reciprocating engines, which combine fast start-up times, high ramp rates and modularity with high power density. Where possible, some developers are also deploying onsite fuel cells, particularly in areas with stricter limits on noise and air pollution, as well as renewables, including solar PV, wind and geothermal.

Data centres require highly reliable power, typically targeting 99.999% to 99.9999% uptime. However, the reliability of grid electricity varies significantly across regions. In advanced economies, annual average outages can amount to just a few minutes in the most reliable systems but can extend to 5 to 10 hours in areas with weaker grids or higher exposure to natural disasters. Even in this worst-case scenario, this equates to reliability of between 99.94% and 99.89%, and grid-connected data centres can reach their reliability targets by complementing grid power with backup storage, through uninterrupted power supply (UPS) systems, and backup generators.

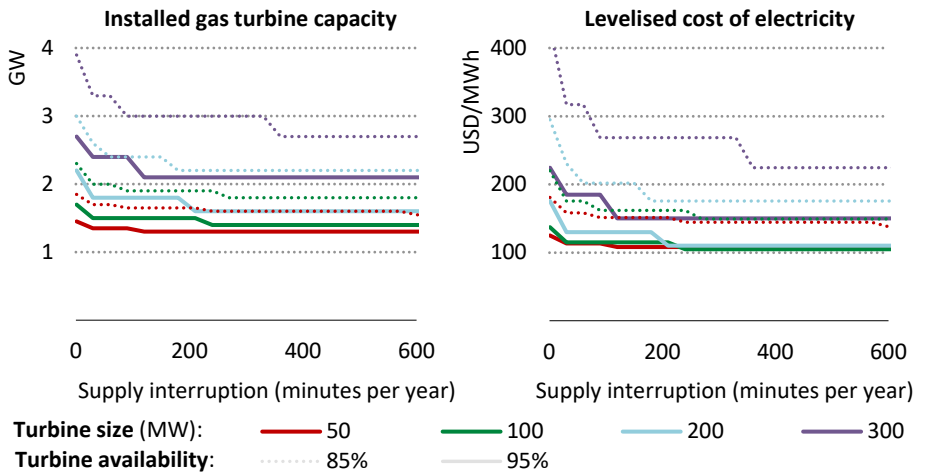
Achieving grid-like reliability through onsite power generation alone is more challenging. Operational data indicate that annual gas turbine reliability[6] generally ranges from 85% to

[6] This is the fraction of the year a unit is technically available to generate when called upon, considering unplanned forced outages and scheduled maintenance downtime.

95%. To ensure continuous power for a 1 GW data centre in an advanced economy at the necessary reliability levels, onsite generation capacity must be oversized to roughly 1.3 GW to 1.7 GW or more, allowing multiple units to be offline while still meeting the full load.

Reliability requirements favour the use of smaller, modular generators, as smaller unit sizes limit the impact of individual unit failures and reduce the reserve capacity needed to achieve a given reliability threshold. Turbine reliability is a key determinant of the required overbuild. Using turbines with an availability of 85% instead of 95%, for example, can increase necessary capacity by 25-45%, depending on the size of the unit and the target level of supply reliability (Figure 6.2). However, this overbuild primarily guards against turbine failure modes that are independent. Onsite units supplied by a single gas pipeline remain exposed to fuel supply interruptions that can affect all turbines simultaneously. Mitigating this risk may require additional measures, such as deploying turbines with dual-fuel capability and onsite liquid fuel storage, or dedicated additional diesel backup generation.

Figure 6.2 ▷ Installed gas turbine capacity and levelised cost of electricity for powering a 1 GW data centre at different reliability levels



IEA. CC BY 4.0.

Onsite generation can help bypass grid connection queues, but achieving grid-level reliability requires redundancy, making it more expensive than grid power in many cases

Notes: Assumes a turbine efficiency of 40%, a weighted average cost of capital of 7%, a data centre load factor of 70%, and operation and maintenance costs of 20 USD/kW/year. Capital costs are 2 500 USD/kW for 50 MW, 2 000 USD/kW for 100 MW, 1 650 USD/kW for 200 MW, and 1 450 USD/kW for 300 MW turbines.

The fixed costs associated with this overbuild strongly influence project economics. Capital expenditures and fixed operation and maintenance costs for seldom-used standby capacity are spread over the same energy output, pushing up the levelised cost of electricity (Figure 6.2). Recent increases in gas turbine capital costs, driven by rising component prices, full order books and growing backlogs, add further upward pressure. As a result, even with

low-cost natural gas of around USD 4/MBtu and very high turbine reliability of 95%, the levelised cost of electricity for onsite power would sit above typical grid electricity prices for large industrial loads in much of the United States. These can be as low as USD 60/MWh in some US states, including Louisiana and Texas.

Lower turbine reliability raises costs further. At 85% availability, levelised costs rise to roughly USD 180/MWh to USD 220/MWh, significantly above prevailing prices for grid electricity. In Europe, higher natural gas prices (TTF averaged around USD 12/MBtu in 2025) make gas-based onsite generation significantly more expensive than in the United States, with levelised costs approaching USD 190/MWh to USD 300/MWh for reliability levels that match those of local grids. This is well above the USD 150/MWh paid on average by industrial end users across the European Union. This is a key reason why onsite natural gas power projects for data centres are concentrated almost exclusively in the United States today.

Gas-based onsite generation is not only typically more expensive than grid electricity for data centres but also entails significantly greater planning, permitting and operational burdens. Moreover, long and increasing lead times for gas turbines (Box 6.1), combined with complex permitting and construction processes and, in some regions, constraints on available gas pipeline capacity, mean that deploying onsite generation may not ultimately be faster than waiting for a grid connection.

Box 6.1 ▻ The boom in natural gas turbine orders and growing waiting times

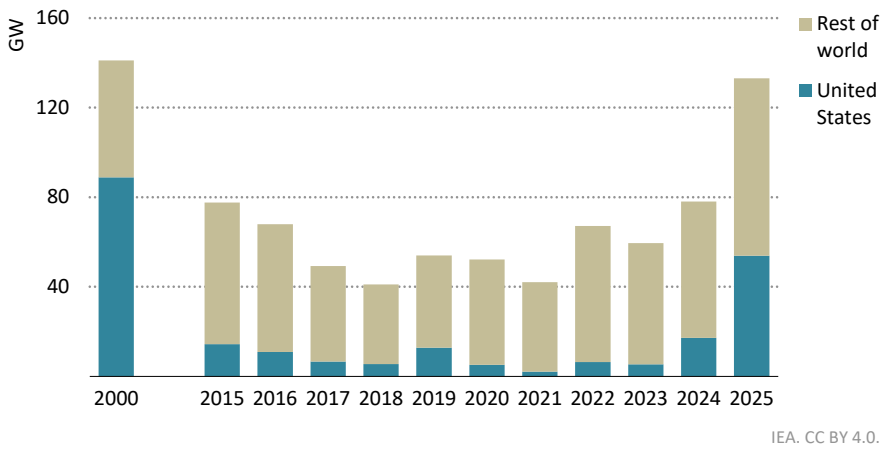
In 2025, global orders of natural gas turbines reached the second-highest level on record and their highest point in 25 years. AI and data centres were only one of the drivers of this surge.

Companies outside the United States accounted for nearly 60% of the global market and increased orders by one-third in 2025 relative to 2024. Except in a few countries, such as Ireland and Malaysia, data centres have not been a major driver in electricity demand growth among this group and therefore were not the direct cause of the jump in turbine orders.

Meanwhile, within the United States, data centres were an important contributor, both directly and indirectly, to the sharp increase in turbine orders. Utilities have been aiming to mitigate growing concerns over capacity adequacy as loads increase and reserve margins shrink, and in hotspot regions such as Texas and Virginia, data centre load growth is likely a major factor. At the same time, data centre developers looking to procure onsite prime or backup power have also been placing orders. Of the roughly 54 GW of turbines ordered by stakeholders in the United States in 2025, around 10% were tracked as intended to power data centres directly. An additional 24 GW of orders are in US states characterised by high existing or fast-growing data centre load. This suggests that data centre load growth may be a factor behind these orders, but it is also important to note that orders also rose sharply in US states that do not host significant data centre capacity. A broader trend of electrification, industrial reshoring and data centre load growth is at play.

A scarcity mentality has also taken hold, with companies rushing to lock in orders in advance as wait times extend to five years or more and prices jump. Currently, the leading producers have a combined manufacturing capacity of around 50 GW to 60 GW per year, implying a growing backlog at current order levels. While they are expanding manufacturing capacity, it will take time for this additional output to come online. Working through the current backlog and meeting future orders will therefore take time.

Figure 6.3 ▷ Global natural gas turbine orders by region



There was a surge in natural gas turbine orders in 2025, although data centres were not the main driver of this increase globally

Source: IEA analysis based on McCoy (2026).

6.2 Operational challenges of supplying electricity to data centres

Data centres are often viewed as stable, near-constant baseload electricity consumers. In many conventional facilities, this characterisation holds true. However, as computing needs evolve, particularly with the rapid expansion of AI, the operational requirements for supplying electricity to data centres are becoming more complex (see section 4). These shifts introduce new challenges for both grid-connected facilities and sites relying on onsite generation.

AI is changing data centre load profiles and making electricity storage an essential tool

A key emerging challenge is the increasing speed at which electricity demand can vary within modern data centres. AI and other compute-intensive workloads can trigger steep and sudden load ramps as clusters of GPUs transition synchronously between tasks. These transitions can cause power demand to swing within fractions of a second, far faster than traditional onsite generation technologies such as diesel generators or gas turbines can respond. One documented case showed that a data centre block focused on AI training

experienced a rapid load increase from 6 MW to 30 MW in just 0.25 seconds (Elevate Energy Consulting, 2025). It would be impossible for onsite generation to follow such fast load swings without the help of other technologies. A typical aeroderivative gas turbine requires around 5 minutes to reach full output after start-up and can then ramp at around 50% of nominal capacity per minute.

Where a data centre is connected to the grid, such load fluctuations, without sufficient buffering, could propagate to the grid and affect power quality at the distribution level, particularly in regions where data centres are clustered. In response to this challenge, grid operators have begun publishing ramp-rate requirements for large loads. For onsite generation, rapid load swings may strain equipment, reduce efficiency or be simply too fast for onsite units to respond in time.

As a result, operators are increasingly deploying hybrid control systems that combine UPS units, battery energy storage systems, battery backup units on rack and capacitors to absorb rapid load swings. These systems help prevent short-duration fluctuations from propagating back into the grid or to onsite generation equipment. They are likely to play a crucial role in enabling onsite systems to overcome the operational challenges of meeting data centre loads. Hardware and software solutions at the GPU and rack levels are also emerging as complementary approaches to managing AI load variability, although they may also involve wasteful "compute burn" to smooth rapid swings in compute demand, as noted in section 4.

Increasingly, battery energy storage systems are emerging as a key enabling technology for AI data centres. They are often described as the system's "multi-tool" due to their operational versatility as a complement to grid or onsite generation and as a lever for load-management. In data centres with onsite power, batteries can provide fast-acting reserve capacity by bridging periods when demand ramps up or down more quickly than onsite generation units can respond. This includes covering start-up delays, managing controlled shutdowns or absorbing short-duration fluctuations. The ability of battery systems to deliver near-instantaneous power allows them to smooth load profiles, reducing cycling stress and improving operational stability. In addition, these systems can enhance power quality by providing frequency support and harmonic mitigation, thereby limiting adverse impacts on gas turbines, reciprocating engines and other onsite power assets.

Overprovisioning and underutilised infrastructure

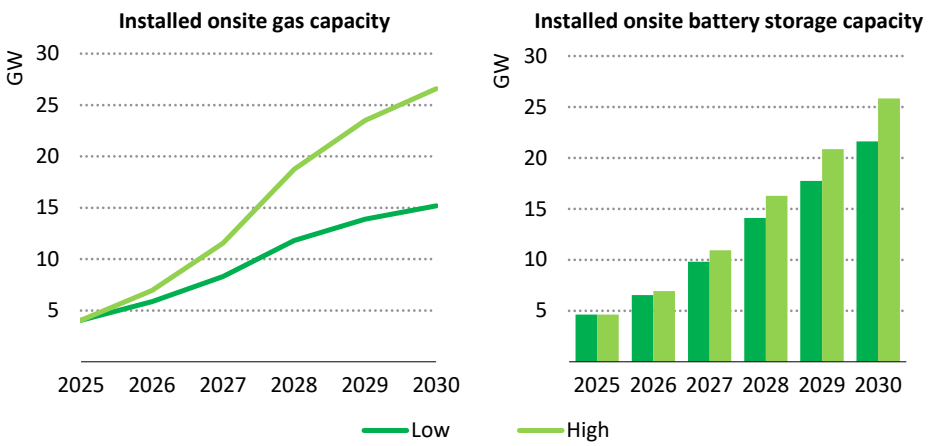
Another characteristic of data centres is the difference between the actual power consumption of IT and auxiliary equipment and their theoretical maximum design capacity. As data centres are designed to ensure high reliability, meet stringent redundancy requirements and accommodate future demand growth, most operate below their maximum design capacity for much of their lifetime. The Electric Power Research Institute reports that the hyperscale facilities it examined operated at an average annual capacity factor of 75% of their grid-connected capacity, compared with 57% across smaller co-location facilities (EPRI, 2026). While load shapes vary by workload and data centre tenant mix, peak demand often remains meaningfully below nameplate capacity.

While this underutilisation increases costs, it also provides valuable headroom to absorb sudden load changes without breaching equipment limits. From an operational perspective, this flexibility acts as an internal buffer, reducing the risk that rapid demand swings translate into power quality issues or grid disturbances. However, this also raises additional challenges for powering data centres with onsite power. Scaling onsite power equipment over time as the IT load ramps up may not be possible if, for example, supply chain bottlenecks require the upfront procurement of turbines. At the same time, overprovisioning onsite power infrastructure may further raise costs.

6.3 Key takeaways

In many regions, data centre expansion is outpacing power grid development, prompting some developers to consider onsite generation, primarily natural gas in the United States, to secure faster access to electricity and mitigate grid delays. However, many of these projects are still at an early stage of development. Many developers are pursuing hybrid solutions that combine onsite generation and storage with grid connectivity, possibly repurposing assets as backups once connected. Onsite generation adds technical and economic challenges. Project outcomes therefore remain uncertain, but reasonable estimates can help inform system operators and suppliers.

Figure 6.4 ▷ Global installed onsite gas and battery energy storage capacity for data centres, 2025-2030



Onsite gas-fired generation capacity and onsite battery storage capacity for data centres expand significantly to 2030

Sources: IEA analysis based on Benchmark (2026) and GEM (2025).

Looking ahead, we project that installed onsite gas-fired generation for data centres could expand rapidly to 2030, reaching 15 GW to 27 GW globally, equivalent to around 9 GW to

20 GW of data centre demand once reliability-driven overbuild requirements are considered (Figure 6.4). In the United States, where this growth is mostly concentrated, we project that onsite projects could cover 6-13% of total growth in data centre capacity to 2030.

Onsite battery systems are crucial for data centre power supply reliability, providing short-term backup through UPS units in both grid-connected and off-grid setups. By 2030, global data centre UPS capacity is expected to rise by more than 100 GW, broadly matching the growth in data centre infrastructure. Lithium-ion options are gaining ground due to longer lifespans, compact size and better performance.

Beyond UPS, longer-duration battery energy storage systems are increasingly being deployed to smooth loads within data centres and provide flexibility to the grid. As of 2025, around 5 GW of dedicated onsite battery storage capacity is operational at data centres worldwide. We project that global installed capacity of longer-duration battery energy storage in data centres will rise from around 5 GW today to 20-25 GW by 2030. The majority of this will be in the United States, where onsite battery storage in data centres is expected to reach 10-12 GW by 2030, compared with more than 90 GW of grid-connected utility-scale battery storage. This highlights that data centre battery storage could become an important grid asset if the incentives are right, helping both to smooth internal load swings and support the flexibility of the grid.

7 What are financial markets telling us about the impact of AI on the energy sector?

AI was a central theme in financial markets in 2025, with an early surge of capital market enthusiasm gradually shifting towards a more cautious and discerning outlook in the second half of the year. While the long-term trajectory of AI remains uncertain, continued uptake will require more energy for data centres, and some areas of the energy sector appear to be growing partly in response to expectations of higher data centre electricity demand. To this extent, financial markets may capture important information about the perceived impacts of AI and the build-out of data centres on the energy sector. This section examines the relationship between the equity market performance of energy and AI companies, how significant that relationship is and what it implies for the broader energy system.

7.1 Which energy market segments have benefited from the AI boom?

Market capitalisation reflects the collective assessment made by investors of a company's current business fundamentals and future prospects. Notably, since the release of public-facing AI applications in late 2022, which marked the emergence of this new era of modern AI, the shares of energy firms overall have increased for most sectors, though not at the same pace as the broader market or AI-related technology companies.

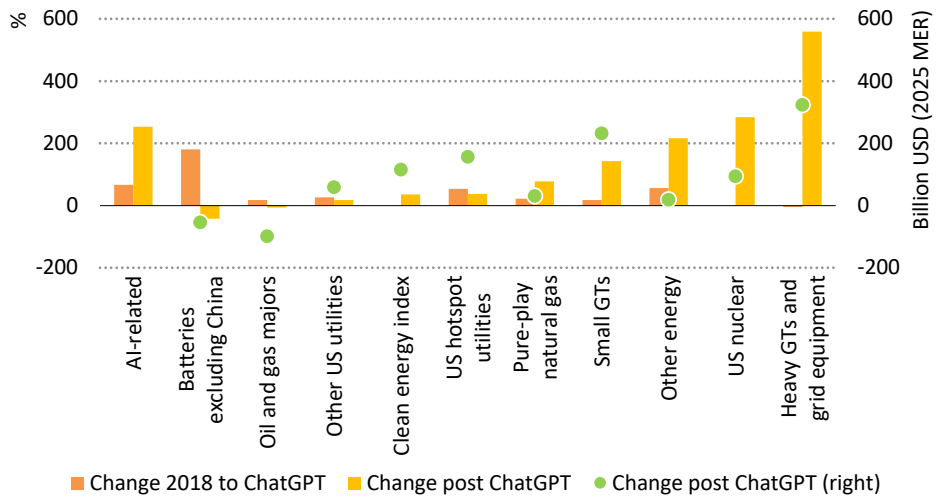
The combined market capitalisation of a broad basket of utilities, oil and gas companies, clean energy companies and battery storage manufacturers has grown moderately compared with AI-related technology companies (Figure 7.1). Moreover, growth across these different segments of the energy sector was generally stronger before the launch of ChatGPT than in the period since. Between December 2022 and January 2026, the market capitalisation of electric utilities located outside of data centre hotspots in the United States, for example, increased by 17%, compared with an increase of more than 70% in the benchmark S&P 500 Index. Utilities operating specifically in data centre hotspots in the United States saw their market capitalisation increase by around 38% over the same period.

That said, other pockets of the energy system, including gas turbine and grid equipment manufacturers, fuel cell and geothermal companies, United States-based nuclear companies and, to a lesser extent, pure-play natural gas companies, have seen their market capitalisations grow at a faster rate since the launch of ChatGPT. For certain subsectors, the increase has been so large that it exceeds, in percentage terms, the growth in market capitalisation for AI-related companies over the same period. Some of this growth is attributable to the data centre boom. For example, growing interest in the use of aeroderivative gas turbines[7] for captive power generation is a new phenomenon, as is the signing by technology companies of power purchase or other agreements for nuclear energy. At the same time, other factors, such as growing electricity demand in other end-use sectors,

[7] Aeroderivative gas turbines are lightweight, modular engines derived from aircraft jet engines and optimised for rapid start-up, high efficiency and flexible power generation.

macroeconomic trends and major policy shifts, cannot be discounted as also playing a significant role.

Figure 7.1 ▷ Changes in market capitalisation before and after the launch of ChatGPT, and absolute change after the launch of ChatGPT



Some energy sectors have seen their market capitalisation grow significantly since the release of ChatGPT, some even faster than AI stocks

Notes: Clean energy index = S&P Global Clean Energy Transition Index; GTs = gas turbines; US = United States; hot-spot utilities = utilities in states where data centre load is high and/or fast-growing as a percentage of total load growth. "Change 2018 to ChatGPT" measures the period between January 2018 and the public release of ChatGPT in November 2022. "Change post ChatGPT" measures the period between December 2022 and January 2026. Analysis is limited to publicly listed companies or private companies with valuations available. The market capitalisation change for AI-related companies is omitted from the chart due to its size (USD 12 trillion). Historical analysis is omitted for clean energy and US nuclear due to data gaps.

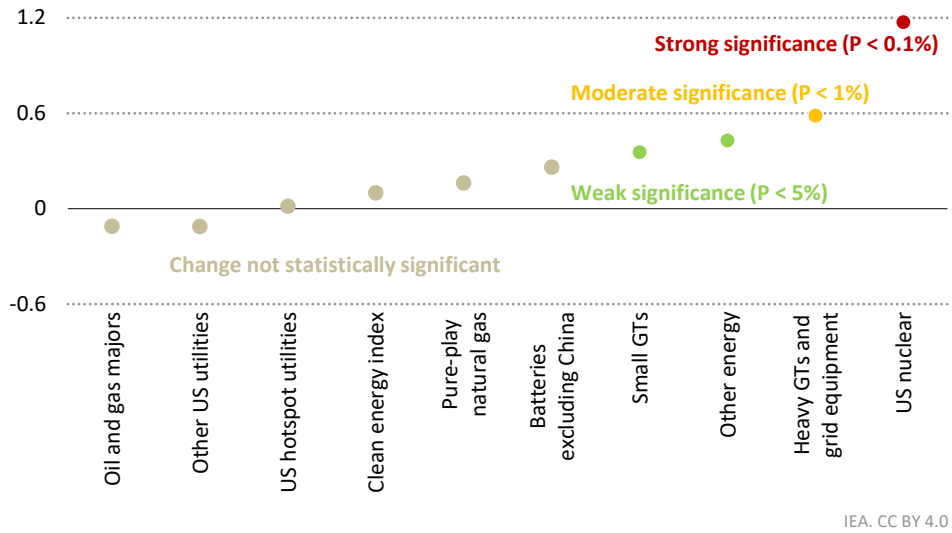
Source: IEA analysis based on data from S&P Global Capital IQ (2026).

Given these interlocking dynamics, we examined whether co-movement between AI and energy stocks had strengthened while controlling for some structural market variables and policy shifts (Figure 7.2). This analysis found no statistically significant change in the degree of co-movement between AI and energy stocks for most segments of the energy sector between the public launch of ChatGPT in late 2022 and January 2026, including for United States-based utilities. In contrast, the degree of co-movement among gas turbine and grid equipment manufacturers and US nuclear companies was the most pronounced.

The analysis provides evidence of a stronger linkage in market sentiment between certain segments of the energy sector and AI-related companies. However, it does not attribute their financial performance solely to the AI data centre boom, especially in light of concurrent

trends in electrification and, more generally, various other intangible drivers of market sentiment that could not be controlled for.

Figure 7.2 ▷ Change in strength of daily price co-movement between AI and energy stocks, before and after the launch of ChatGPT, by sector



IEA. CC BY 4.0

The price movements of some energy stocks have become more sensitive to movements in AI-related stocks since ChatGPT was released to the public in late 2022

Notes: Clean energy index = S&P Global Clean Energy Transition Index; GTs = gas turbines; US = United States; hot-spot utilities = utilities in states where data centre load is high and/or fast-growing as a percentage of total load growth. The y-axis measures how a unit-change in AI returns is associated with an X percentage-point stronger or weaker change in energy-sector returns in the post-ChatGPT period (December 2022 to January 2026) relative to the pre-ChatGPT period (January 2018 to November 2022). P = p-value, which tests whether the change in AI-energy co-movement is statistically different from zero. A smaller p-value means a lower likelihood that the change in co-movement is the result of chance (e.g. a 5%, 1%, or 0.1% chance or less). A positive value means co-movement has become stronger, but it does not provide any information on the direction of movement. The regression is run on daily price movement data for both periods and controls for policy changes like the Inflation Reduction Act and the One Big Beautiful Bill Act.

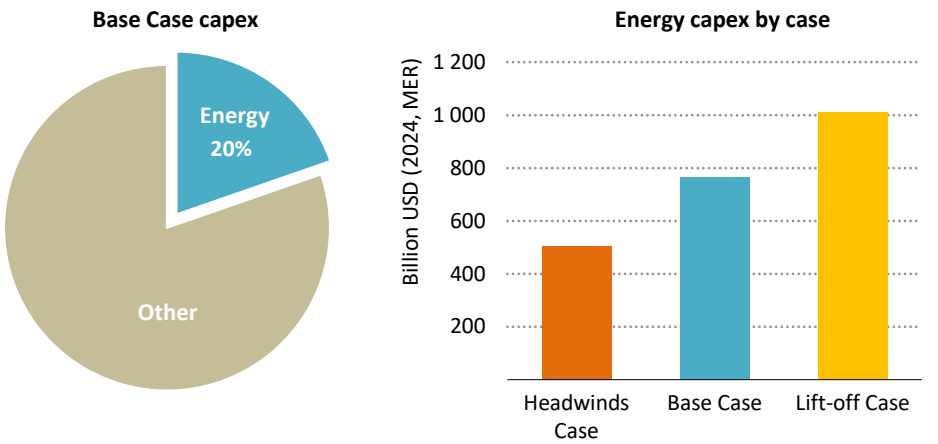
Sources: Regression analysis based on data from: S&P Global Capital IQ (2026); K. French (2026); Chicago Board Options Exchange (CBOE, 2026); Federal Reserve Bank of St. Louis (2026); U.S. Energy Information Administration (EIA, 2026).

7.2 How does increased exposure to market sentiment towards AI create opportunities or risks for the energy sector?

A company's valuation affects its ability to raise capital. A higher valuation makes it easier for companies to issue equity or borrow to finance new investments. Conversely, declining valuations make equity issuance more dilutive and worsen leverage ratios. This can trigger adverse debt covenants, raise borrowing costs and make refinancing more challenging.

Capital expenditures on energy equipment, such as backup power units, uninterrupted power supply units and grid connections, represent a relatively small share of data centre investment in any scenario (Figure 7.3). Investment in non-energy components, such as IT equipment, accounts for 80% of total data centre investment in the Base Case. Yet, with cumulative investments measured in trillions of dollars, even the small share of energy investment in the total is potentially significant. In absolute terms, the difference in energy-related investment between the highest and lowest IEA cases is more than USD 500 billion.

Figure 7.3 ▷ Data centre investment by type in the Base Case and total energy investment for data centres by case, 2026-2030



Of the USD 3.9 trillion in total data centre investment from 2026 to 2030 in the Base Case, only 20% is for energy

Notes: Capex = capital expenditure. Energy includes investment in electricity grids, generation, backup generators and uninterruptable power supply for data centres.

Source: IEA (2025a).

A pullback in the data centre build-out could, in theory, lead to stranded assets and refinancing risks for energy companies. At the same time, there is considerable upside potential, with the AI-led increase in demand for electricity and equipment presenting opportunities to commercialise new technologies or reinvigorate the supply chains and skilled workforce for existing technologies. How this plays out for energy companies is contingent not only on their exposure but also on company-specific strategies and approaches to risk mitigation.

7.3 How are companies responding to increased AI demand growth?

Large gas turbine manufacturers are expanding capacity with caution

Gas turbine manufacturers are preparing to expand manufacturing capacity over the next several years in response to rising demand (Table 7.1). Recent corporate announcements by manufacturers indicate a pivot towards the US market, due in part to the rapid expansion of data centres there, but also due to overall increases in domestic electricity demand following decades of stagnation (Siemens Energy, 2026). However, because this expansion follows a prolonged period of weak order intake, most companies appear to be taking a measured approach, focusing on reducing lead times and scaling output within existing facilities rather than pursuing aggressive greenfield growth. Only small turbine manufacturers with limited existing manufacturing capacity have signalled more ambitious plans.

Table 7.1 ▷ Expansion plans of selected turbine and fuel cell manufacturers

Company	Country of headquarters	Manufacturing expansion plans
GE Vernova	United States	Capacity expansion to 20 GW by mid-year 2026 and 24 GW by 2028 using existing facilities
Siemens Energy AG	Germany	Capacity expansion from 17 GW in 2024 to an average of 22 GW from 2025 to 2027, increasing to 30 GW by 2028/30 without increasing factory footprint
Mitsubishi Heavy Industries	Japan	Capacity expansion by 30% with a focus on short-term optimisation
Caterpillar	United States	1.2 GW peak production of Solar turbines in 2022; plans to increase turbine production 250% by 2030
Bloom Energy	United States	Capacity expansion to 2 GW by 2026
Boom Supersonic	United States	Capacity expansion to 2 GW by 2028

Sources: Siemens Energy (2025a); GE Vernova (2025); SemiAnalysis (2025b).

This capital discipline is also supported by a diverse set of long-term revenue drivers linked to electrification. For example, Siemens Energy has increased transformer manufacturing capacity by 20%, and the company is reporting an order backlog of EUR 42 billion for its grid technologies (Siemens Energy, 2025b). The electrification segment of GE Vernova similarly recorded year-on-year revenue growth of 25% and a 30% increase in its order backlog (GE Vernova, 2025). Improved margins in the turbine segment may even create positive spillovers for these adjacent business lines, where investment needs are accelerating.

New rate structures also aim to share infrastructure costs and risks

Large investments are needed to provide grid connections for data centres and meet their power demand. As noted above, uncertainty in the data centre demand outlook remains high, with energy investments ranging from USD 0.5 trillion to USD 1 trillion across different IEA cases. This potentially presents a risk to utilities if investments in power and grid capacity are made to meet demand that later falls short of expectations. At the same time, AI

companies have an interest in ensuring that power investments do take place to meet their demand. Utilities, technology companies and regulators are therefore pursuing innovative tariff models to share costs and mitigate risks.

These include collateral postings for new data centre connection requests in case a planned data centre is not built, stated capacity ramp periods, minimum billing demand commitments over periods of up to 20 years and early exit fees for data centre closures (Credit Sights, 2026). In the United States, regulators and/or utilities across 24 different states have approved or proposed some combination of these measures, covering many data centre hotspots (Credit Sights, 2026 and Goldman Sachs, 2026). In tandem, hyperscalers such as Microsoft have committed to covering the costs of adding and using electricity infrastructure for their data centres (Microsoft, 2026). Several other hyperscalers have signed a federal government-led pledge to prevent infrastructure costs from being passed on to US households (White House, 2026).

There are some exceptions where data centre load growth is high and protections have not been introduced, although data centre operators can still develop mechanisms to actively support and co-fund large grid investment in these jurisdictions. Moreover, even in locations where protections have been introduced, they may not cover all fixed costs arising from new large-load connections, including those that occurred before new policies took effect. Protections such as minimum billable demand and long-term contracts are conducive, but full fixed-cost recovery also depends on the tariff level and on whether network charges are shallow or deep.

Overall, recent measures should be effective in transferring some of the financial burden away from utilities, thereby reducing the downside risks for them. However, the extent to which consumers will be protected is less clear. To date, there are few examples of US states legislating full cost recovery by data centre operators (Goldman Sachs, 2026). For instance, proposed legislation in the US state of Georgia to explicitly ban electricity investment costs for data centres from being passed onto consumers was abandoned in favour of more flexible protections (Georgia General Assembly, 2026). More broadly, as discussed later in this report, the impact of data centres on electricity prices is contingent on several other factors. This means that protections for utilities still leave the door open to some form of cross-subsidisation in either direction, i.e. other ratepayers subsidising data centre costs, or data centre operators subsidising electricity costs for other ratepayers.

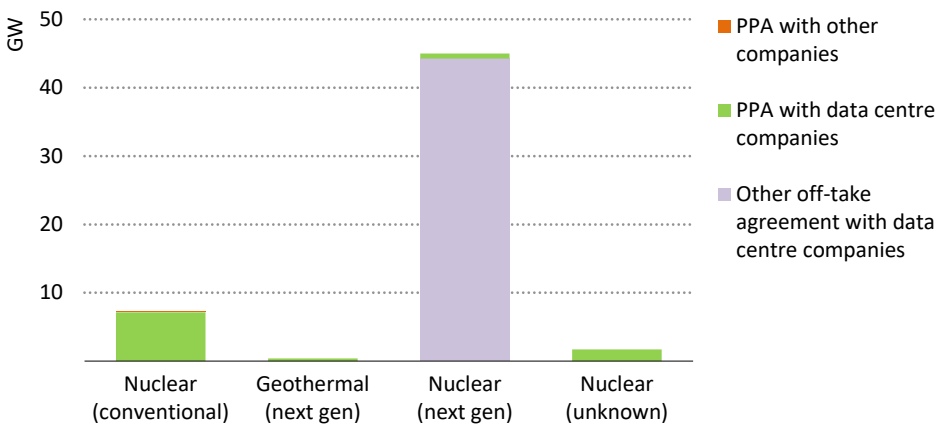
The success of smaller start-ups may be more contingent on the data centre build-out

Operators of existing nuclear power plants have been major beneficiaries of corporate power-purchase agreements (PPAs) from AI-related companies, which amounted to 7.1 GW from 2024 through the first quarter of 2026. A key quality of these PPAs is that most have received a positive final investment decision and represent a firm commitment to contract, restart, upgrade or extend the lifetime of existing generators. The details of these agreements are not always known, but off-takers are usually bound to their terms for up to 25 years, giving generators confidence to make new investments where necessary.

In contrast, the 45 GW of new small modular reactor (SMR) capacity contracted by AI-focused tech companies has largely taken the form of less formal off-take arrangements. These off-take agreements are much larger than firm PPAs for SMRs and, due to the immaturity of the technology, are not yet backed by firm financing commitments (Credit Sights, 2026). Consequently, any pullback from AI-related companies would be a serious setback. Continued demand growth from data centres for low-emissions, baseload power would provide a considerable boost to the commercialisation of these technologies. The same could also be true for next-generation geothermal, which has likewise benefited from the backing of large technology companies and the signing of several off-take agreements, with the first commercial project set to start supplying electricity as soon as 2026 or 2027 (Fervo Energy, 2023).

In either case, public support and strategic public capital will play an important role in reducing market risks for these early-stage projects while also helping to catalyse private capital.

Figure 7.4 ▷ Selected geothermal and nuclear generation corporate PPAs by off-taker type, total 2024-2026



Most PPAs for data centres have been for existing nuclear generators, but tech companies have signed 45 GW of other off-take agreements with next-generation nuclear

Source: IEA analysis based on BNEF (2026).

7.4 Key takeaways

Powering data centres will require cumulative investment of between USD 0.5 trillion and USD 1 trillion through to 2030. While large in absolute terms, this remains relatively small compared with overall investment in the energy sector, which is projected to be around USD 18 trillion from 2026 to 2030. Similarly, data centre electricity consumption is projected

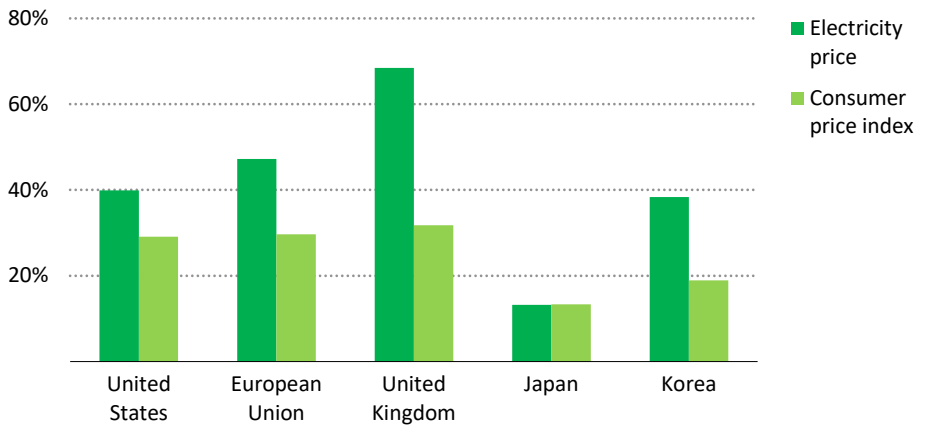
to rise from around 1.7% of electricity demand in 2025 to around 3% by 2030. For these reasons, the data centre boom is not showing up as unusual growth in market capitalisation at the level of the energy sector as a whole.

On the other hand, there are pockets of the energy sector where prospects do seem to be driven, at least in part, by the growth of data centres. These include gas turbine and grid equipment manufacturers, nuclear companies, some start-ups in fields such as fuel cells, and utilities in major data centre hotspots, although the empirical evidence of a link is somewhat weaker in the last case. In these areas, outcomes will depend not only on the outlook for data centre demand but also on the investment strategies adopted by companies and on regulatory measures to share risks, for example in the case of electricity tariffs. The data centre boom is providing significant impetus to a number of energy start-ups, but public support will also be required to ensure that they reach their full potential.

8 Are data centres raising electricity prices?

Wholesale and retail electricity prices have risen sharply across many major economies in recent years following the post-Covid-19 inflation spike and the global energy crisis of 2022. Between January 2019 and December 2025, residential electricity prices increased faster than headline inflation in markets such as the European Union, Korea, the United Kingdom and the United States. Although electricity prices eased somewhat from 2023 onwards as gas markets stabilised, they remain elevated relative to pre-crisis averages and continue to be exposed to market volatility in many markets, keeping affordability in focus.

Figure 8.1 ▷ Change in nominal residential electricity price and consumer price index by country or region, January 2019-December 2025



Residential electricity prices have risen faster than headline inflation in several markets in recent years

Sources: UK Department for Energy Security and Net Zero (2025), Eurostat (2025), Federal Reserve Bank of St. Louis (2025), Korean Statistical Information Service (2025), and Statistics of Japan (2025).

8.1 Surging data centre growth in some regions presents a unique challenge

In recent years, electricity demand from data centres has grown rapidly in some regions and is projected to continue rising, raising questions about how this growth could affect electricity prices.

Electricity demand growth from data centres affects power markets through several interrelated mechanisms. The challenge stems not only from the scale of demand but also from the speed of growth, which can outpace infrastructure planning and execution cycles. Understanding how these mechanisms operate, and how their effects vary in different markets, is essential to assessing both the risks and options for mitigation. Analysis of this

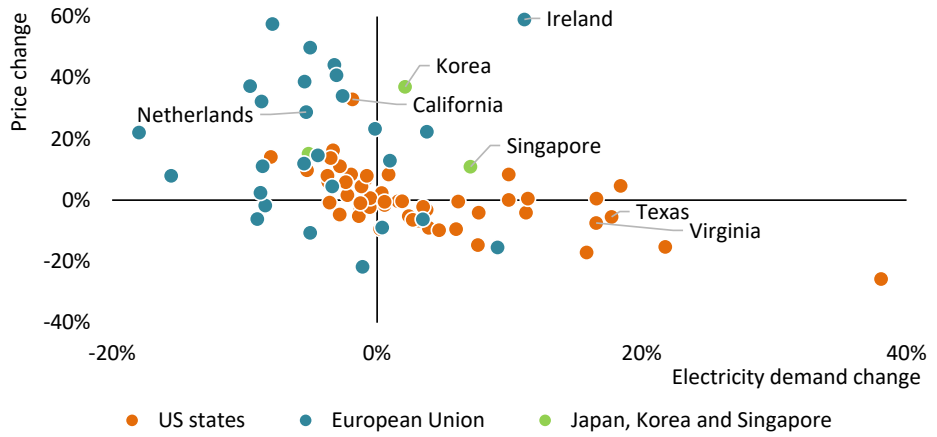
issue needs to distinguish between short-term and long-term effects, and between effects in systems with spare capacity and those with tighter demand-supply balances.

Short-term effects

In the near term, rapid growth in large and typically inflexible electricity loads, such as data centres, can increase reliance on higher-cost generation units that would otherwise run less frequently, putting upward pressure on wholesale prices. Where data centres cluster in specific locations, they can increase costs arising from transmission congestion. Fast-growing data centre clusters can also prompt the short-term procurement of emergency reserve generation to maintain adequacy.

By contrast, in systems with more relaxed balances of demand and supply, additional demand can improve the utilisation of existing assets and spread fixed costs over a larger customer base, lowering per-unit network and fixed-cost charges and, in some cases, even exerting downward pressure on per-unit system costs. These effects tend to be strongest in large, well-interconnected systems with diversified loads and supply, where incremental demand can be absorbed without triggering major network upgrades or new capacity requirements.

Figure 8.2 ▷ Change in real retail electricity price versus change in electricity demand for selected markets, 2019-2024



IEA. CC BY 4.0.

Changes in real retail electricity prices show no clear or systematic association with changes in electricity demand

Notes: This figure is descriptive only and does not control for other major drivers of electricity prices over the period, such as fuel price movements, taxation, retail price regulation, subsidies, network charges, market structure or the impacts of the energy crisis, all of which have played a major role in shaping electricity prices over the period and whose relative importance varies substantially across regions. Retail electricity prices are calculated as a consumption weighted average across residential and industrial sectors.

Sources: EIA (2025) and IEA (2025d).

Across selected markets, changes in retail electricity prices from 2019 to 2024 show only a weak relationship with changes in electricity demand (Figure 8.2). In the European Union, price increases have frequently occurred in the context of flat or declining demand, which may reflect demand destruction as a response to elevated prices in recent years. In the United States, electricity prices were less affected by the 2022 energy crisis, and states with greater load growth generally did not see higher price increases, consistent with earlier empirical findings (Wiser et al., 2025). The analysis shown in Figure 8.2 reflects national or state-level averages and may therefore mask localised price effects in areas with concentrated load growth. Ultimately, the data highlight that there is no simple, one-way relationship between load growth and electricity prices.

Longer-term effects

Beyond the near-term effects, a core challenge is that data centres are expanding on timelines far shorter than those of the power system. New data centre facilities can be deployed in one to two years, while adding major transmission infrastructure or new dispatchable generation capacity typically takes much longer. This mismatch means that sustained, concentrated demand growth can quickly overtake existing capacity, boosting investment needs for the energy system over the longer term.

Expectations of rapidly growing peak loads are already affecting capacity markets by tightening forward-looking capacity adequacy assessments. This was visible in recent auctions in the PJM[8] Interconnection, which covers Virginia's "Data Centre Alley", where clearing prices reached the regulatory price cap in successive rounds. The independent market monitor, Monitoring Analytics, identified projected large data centre load additions as a key contributing factor (Monitoring Analytics, 2025).

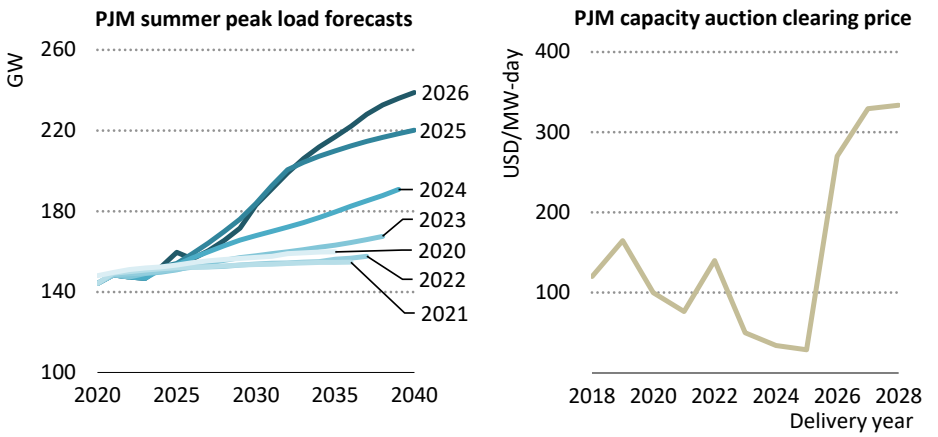
Capacity prices in PJM have been affected by supply side and administrative issues, such as delays from long interconnection queues and changes in resource accreditation that lowered capacity credits for parts of the gas fleet. Recent changes to PJM's capacity market have also drawn criticism. Monitoring Analytics found that the new design was implemented before it had been fully tested and that market outcomes are overly sensitive to modelling assumptions, leading to unstable price signals. In response, PJM has adopted interim measures, including tighter participation requirements across resource types, revisions to key cost assumptions and stricter price limits.

Uncertainty around the realisation of the load forecasts remains a key risk. PJM load forecasts may include duplicative or speculative requests from data centre developers pursuing multiple siting options. Recent updates show that near-term load projections for 2026 are lower than in the previous year's forecast, even as long-term projections continue to rise.

[8] PJM is a regional transmission organisation in the United States that coordinates and monitors the flow of electricity across the high voltage grid, and operates wholesale electricity and capacity markets. See: https://www.ferc.gov/introductory-guide-participation-pjm-processes.

Over the long term, sustained growth in data centre electricity demand can reshape power systems by raising peak loads and increasing the need for new generation capacity and network infrastructure. However, even if more investment is needed to maintain system adequacy, whether this leads to an increase in the per-unit cost of electricity depends on a range of factors.

Figure 8.3 ▷ PJM summer peak load forecasts and the clearing price in PJM's capacity auction



Summer peak load forecasts have increased significantly in recent years, while clearing prices have reached the market cap in the two latest capacity auctions

Notes: PJM Interconnection is a regional transmission organisation in the United States that coordinates the supply of electricity to key data centre clusters, among other consumers. The delivery year is defined as the period from 1 June of the preceding year to 31 May of the year referenced in the figure and is the period during which auctioned capacity must be available.

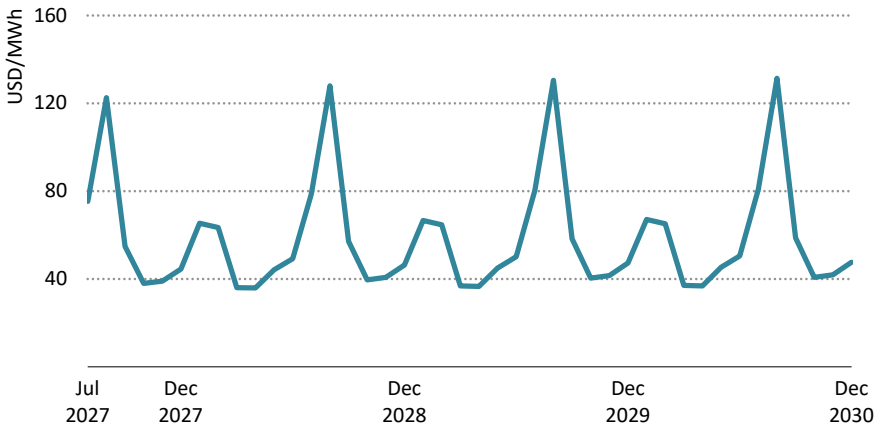
Sources: PJM (2020, 2021, 2022, 2023, 2024, 2025a, 2025b, 2026).

Box 8.1 ▷ What can Texas forward prices tell us about the effect of data centres on electricity prices?

Forward electricity prices are often used to gauge expectations about future market conditions. By reflecting the price agreed today for power delivered in the future, they incorporate market views on the overall balance between supply and demand, including factors such as fuel prices, capacity additions, policy changes and demand growth. Expectations about the expansion of data centres are only one of many drivers that may influence forward price curves. In principle, if data centre growth were expected to significantly tighten system balances or raise costs, this could be reflected in higher forward prices. However, forward prices also reflect risk premia, hedging dynamics and declining liquidity at longer horizons, which means they imperfectly capture the underlying fundamentals.

The Electric Reliability Council of Texas (ERCOT) offers an interesting case study in this regard. First, the state is a hotspot for data centre growth (Box 1.2). Second, ERCOT does not have a capacity market, so the energy-only market provides the only incentive for investment, with scarcity pricing used to signal tight conditions. A relatively flat forward curve suggests that market participants currently anticipate sufficient supply to meet rising demand, including from data centres, without sustained price pressure (Figure 8.4).

Figure 8.4 ▷ ERCOT North Hub on-peak power forward prices



Despite large seasonal variation, current ERCOT North hub forward prices are relatively flat through to 2030

Note: The forward prices reflect market expectations as of 5 March 2026 and may change as market conditions evolve.

Source: IEA analysis based on S&P Global (2026).

Several factors may help explain this. First, supply growth expectations remain strong, with rapid additions of solar, wind and battery storage continuing to expand capacity. Second, a large share of announced data centre demand remains uncertain, as interconnection queues for large loads often overstate the amount of demand that ultimately materialises. Third, fuel price assumptions, particularly for natural gas, which frequently sets the marginal price, continue to play a dominant role in shaping forward electricity prices. Finally, regulatory developments, including proposed state measures to better manage large-load connections, such as in Senate Bill 6, may also influence expectations about how quickly data centre demand will translate into realised system load.

Taken together, a relatively flat forward curve should therefore not be interpreted as evidence that data centre growth will have no impact on electricity prices, but rather as a reflection of current expectations that supply will keep pace with demand in ERCOT.

8.2 Understanding what drives system costs

To maintain electricity security, power systems have traditionally been designed to meet periods of peak demand, even if this results in the underutilisation of assets at other times. As a result, both generation capacity and transmission and distribution networks are designed to ensure adequacy during relatively short windows of time. Electricity systems are also highly capital intensive, with the capital recovery costs of generators and networks usually accounting for around 40-60% of total system costs. Since these costs are relatively insensitive to short-term variations in output, the extent to which assets are utilised over time is a key determinant of average electricity costs.

The ratio of average system demand to peak system demand, also called the system load factor, is a key indicator of asset utilisation and cost efficiency. Systems with low load factors, characterised by high peak demand relative to average load, must maintain large amounts of expensive, often idle generation and network equipment. Conversely, systems with higher load factors allow fixed costs to be spread over a larger volume of electricity consumption, improving asset utilisation and reducing average per-unit costs.

While the system load factor is a useful indicator of asset utilisation, it is not a fixed characteristic of electricity systems. It reflects the aggregation of demand profiles across different end-use sectors and can be influenced by technologies such as battery storage and mechanisms such as demand response. Industrial loads tend to be relatively stable over time and therefore contribute to higher system load factors, while the reverse is typically true of residential loads. A crucial question in assessing the impact of data centres on system costs over the longer term is whether data centre loads will raise or lower system load factors in affected regions. If data centre loads improve the utilisation of generation and network assets over time, they could contribute to lower average costs. If they instead increase peak demand and reduce system load factors, they risk increasing system costs by intensifying the need for additional, rarely utilised capacity.

To understand how these factors play out in the case of data centres, it is important to distinguish between the load factor of data centres and the capacity utilisation rate. The first refers to the ratio of average load to peak load. The second refers to the ratio of average load to nameplate capacity.

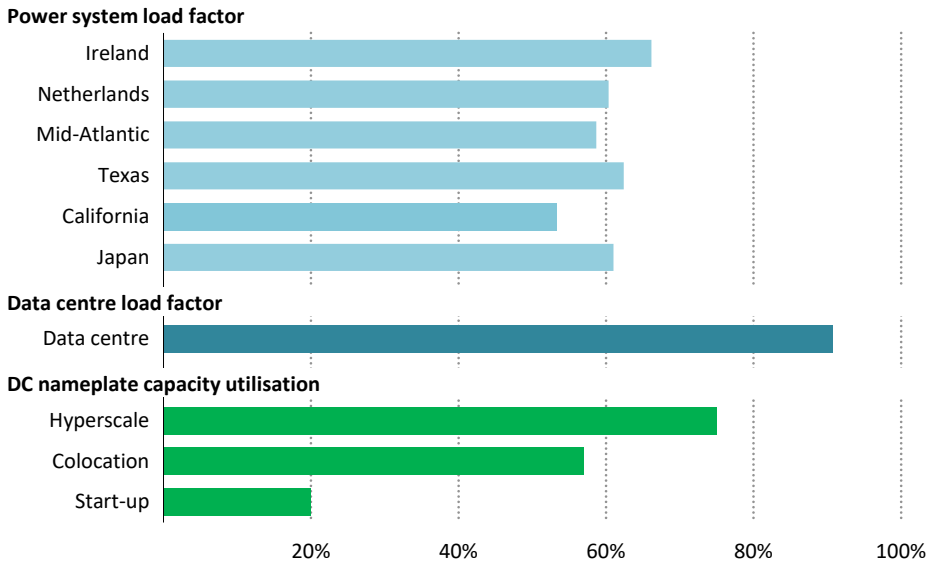
Data centres typically have relatively high load factors, in the order of 75-90%. This characteristic can support the efficient utilisation of generation and network assets over time. However, higher load factors do not automatically imply lower price pressures, as a data centre with a high load factor may still contribute significantly to peak demand, tightening capacity margins and increasing price pressures during peak periods.

At the same time, data centres often have lower capacity utilisation rates. This is not unique to data centres and can also occur with other large industrial loads or residential loads. However, a number of factors make this issue particularly relevant to data centres. Unlike large industrial facilities, they are modular. This means computing equipment is installed progressively, leading to a slow ramp-up of demand and low initial utilisation of grid

connections. Combined with uncertainty about future demand, and competition for scarce capacity, this encourages operators to over-contract capacity. The resulting underutilisation of available grid connections can lead to the oversizing of networks and reserve infrastructure, reducing capital efficiency. In congested systems, the over-contracting of grid capacity can also impose an opportunity cost by blocking capacity that could otherwise serve other loads.

The key point here is that, while data centres may have high load factors, which in theory support affordability by promoting efficient capital utilisation, they often have lower capacity utilisation rates, which may work in the opposite direction. Importantly, there is scope for policy makers to intervene by supporting more efficient integration of data centres into the grid.

Figure 8.5 ▷ Load and capacity utilisation factors of data centres and at power system level



Many data centres exhibit flat demand profiles while making limited use of nameplate capacity, with implications for both operations and capital efficiency in power systems

Notes: DC = Data centre. Load factor is defined as the ratio of average load to peak load. Nameplate capacity utilisation is defined as the ratio of average load to nameplate capacity. Start-up refers to the initial level at commissioning. Data centre load factors are based on a dataset covering large-scale data centres in the United Kingdom. Mid-Atlantic refers to the EIA MIDA electricity region (largely overlapping with PJM).

Sources: IEA analysis based on IEA (2025e), EPRI (2026), and UK Power Networks (2026).

8.3 Policy and technology levers offer scope to mitigate price impacts

Overall, rising electricity demand from data centres poses a growing challenge for power systems and energy affordability, but the scale of the impact depends on system conditions, the pace and location of growth, regulatory frameworks and market design, leaving policy makers and data centre developers with significant scope to shape outcomes.

For example, improving visibility on the location, scale and timing of data centre developments can reduce system costs by avoiding premature or unnecessary investment for power loads that may not ultimately materialise. Early co-ordination between developers, system operators and regulators – combined with stronger financial requirements for new large loads, "use it or lose it" measures for reserved grid capacity and better data transparency around load ramps for data centres – could help align network and capacity planning with the actual outlook for grid demand from data centres.

Policy and planning also shape where new data centre demand materialises and how local system impacts are managed. In highly constrained areas, temporary caps or moratoria on new connections, as seen in Singapore, can safeguard reliability and prevent inefficient investment. In contrast, in less constrained areas, clearer incentives for data centres to locate where spare grid capacity is available, together with faster permitting, can guide development more efficiently.

Operational flexibility could be another effective lever for mitigating system and price impacts. Data centres can provide flexibility by shifting non-urgent workloads over time (Box 8.2), reallocating computing tasks across multiple locations, modulating cooling demand, or using onsite energy storage or generation to manage peak grid demand. Emerging initiatives such as demand response programmes, dynamic contracts and accelerated grid access for non-firm connection agreements could reduce the need for additional reserves, network updates and peaking capacity, thereby mitigating both system-wide costs and local network stress.

Efficiency improvements in servers, cooling systems and overall power usage effectiveness can reduce the growth of electricity demand per unit of digital output. While efficiency gains cannot offset rapid expansion in data centre activity, they can slow demand growth and delay the need for new generation and network investment.

Expanding electricity supply and grid infrastructure is the most direct way to mitigate price and adequacy pressures arising from rising data centre demand. This can include accelerating new generation, co-locating data centres with new supply and introducing capacity remuneration mechanisms to bring forward additional capacity, including rapidly deployable options such as solar PV paired with battery storage. Optimising existing grid infrastructure utilisation through grid-enhancing technologies can also delay or reduce the need for additional network upgrades. In Ireland, for example, new large loads are increasingly required to co-locate with generation capacity, ensuring that demand growth is matched by additional supply and reducing transmission costs. The effectiveness of this lever depends on permitting timelines, grid access and whether new supply is delivered in locations where

demand growth is occurring. At the EU level, recent initiatives, such as the European Grids Package, aim to address these constraints by strengthening network planning and connection processes (European Commission, 2025b).

Policy makers should also ensure that co-location agreements, onsite generation or direct supply contracts do not result in the shifting of costs to other electricity consumers. For example, when a data centre co-locates with an existing plant and diverts its output, other consumers may lose access to that generation, potentially pushing up electricity prices. Similarly, if data centres install onsite generation and reduce their draw from the network, other consumers may face higher per-unit infrastructure costs, raising equity and affordability concerns.

Box 8.2 ▻ Incentivising more flexible data centres

More flexible data centre loads could reduce the need for costly network and generation capacity upgrades and therefore help limit the impact of data centre loads on system costs and prices. Our analysis finds that if data centre demand is flexible for between 0.1-1% of the hours in a year, there is enough room in current electricity systems to integrate all new data centre capacity projected to come online through to 2035 (IEA, 2025a).

There are three kinds of data centre flexibility. First, onsite power assets such as batteries or gas turbines can be used to reduce data centre loads at times of peak demand. Second, auxiliary power consumption, notably for cooling, can be reduced. Third, data centre operations (also known as workloads) could be re-routed geographically or deferred in time to lower electricity consumption during periods of high grid stress. However, numerous challenges must be overcome to make data centre operations more flexible, particularly for the third option. These include the following:

- Organisational challenges: Some data centre operators, notably co-location providers, may not have visibility over the workloads taking place inside the data centre, as these are scheduled and executed by the actors that lease the facility from the operator. This means the operator may not know which workloads could be rescheduled, and which are time-sensitive and cannot be deferred.
- Contractual challenges: Data centre operators may have service agreements with their customers that involve strict penalties for service interruptions. Unless frameworks are robust and incentives are sufficient, operators may be reluctant to take on the added risk of providing load flexibility in case it leads to a breach of their service obligations.
- Financial challenges: Electricity costs typically represent a small share of the total costs of a data centre, which are dominated by the capital costs of IT hardware. There is therefore a strong incentive for data centre customers to maximise IT hardware utilisation, as reduction in utilisation can have high opportunity costs.

Several initiatives and pilot projects are under way to test approaches for making data centres more flexible. Given that time to market is a key consideration, incentives related to faster grid connections are likely to be more attractive than payments for demand response. Frameworks that determine the parameters for curtailment, such as duration and the applicable conditions, could help data centre operators plan more effectively. Improving the exchange of information, including labelling schemes for workloads that define time sensitivity and agreed parameters for deferring workloads, would help achieve greater standardisation and visibility across different parts of the value chain.

Allocating grid costs to ensure fairness

Grid cost allocation, or the method by which the costs associated with investing in, operating and maintaining electrical grids are distributed among various users or customer groups, has become a central policy issue as large new electricity users, including data centres, raise questions about fairness. In some regions, concerns have emerged that data centres may not be paying their full share of the infrastructure required to support them, resulting in cross-subsidisation by other consumers. The process of grid cost allocation is designed to fairly apportion costs based on factors such as usage, capacity and connection type. While tensions among stakeholders on this topic are not new, the pace and scale of recent data centre load growth have brought them into sharper focus.

Traditionally, grid costs are recovered primarily through regulated network charges embedded in retail tariffs. These charges are typically structured as a combination of volumetric (per kilowatt-hour) and capacity-related (per kilowatt) components, often with time-of-use differentiation. In addition, new users usually face one-time connection charges. These may be shallow, covering only the direct connection to the grid, or deep, requiring the user to fund upstream network reinforcements triggered by their connection.

There are large variations across markets in how these charges are applied, as well as in how large load additions are treated (Table 8.1). Globally, large new loads such as data centres are generally integrated into existing grid tariff frameworks rather than being subject to dedicated tariff classes. However, there are exceptions. In the United States, several utilities have introduced, or are introducing, data centre-specific or large load-specific rate classes. One example is Virginia, where regulators have approved a new rate class for large loads over 25 MW and with load factors over 75%, which will take effect from January 2027. This fee includes minimum demand charge provisions, exit fees and long-term commitment requirements (Virginia State Corporation Commission, 2025). Another example is in Canada, where Hydro-Québec, which manages the region's electricity supply and distribution, is proposing a new rate for data centres above 5 MW that is approximately double what large consumers currently pay (Hydro-Québec, 2026).

Data centres are increasingly under scrutiny in this context. Tensions between ensuring that data centres pay their fair share of grid costs and maintaining their global competitiveness are at the heart of today's policy debate. Regulators are increasingly requiring large new

loads to contribute more directly to grid costs, for example through upfront connection fees or additional conditions attached to system access. Several large technology companies have recently signed a pledge with the US federal government (White House, 2026) to take measures to limit the impact of data centre demand on electricity prices for other consumers. Proposed measures include covering network upgrade costs or procuring additional generation capacity. While voluntary and non-binding, this pledge shows that the topic is high on the policy agenda.

Table 8.1 ▷ Grid cost allocation and treatment of new large loads

Region	Current cost allocation	New large load/ DC connection costs	Recent policy focus
PJM (US)	State-regulated tariffs by class, with some utilities introducing DC carve-outs	Direct connection costs plus applicable grid upgrade costs	FERC order to clarify co-located load treatment and cost responsibility (FERC, 2025)
CAISO (US)	Regulated tariffs by class, no DC carve-out	Direct connection costs plus applicable grid upgrade costs	Large-load integration assessment (CAISO, 2026)
Germany	Regulated tariffs by class, no DC carve-out	Direct connection costs plus possible capacity-related capital contribution	Review of system-wide grid tariff structure (BNetzA, 2025)
Ireland	Regulated tariffs by class, no DC carve-out	Direct connection costs	Conditional connection policy with onsite generation requirement (CRU, 2025)
Singapore	Regulated tariffs by class, no DC carve-out	Direct connection costs	Management of DC capacity via allocation rounds (EDB/IMDA, 2025)

Notes: DC = data centre; US = United States; FERC = Federal Energy Regulatory Commission; BNetzA = Bundesnetzagentur; CRU = Commission for Regulation of Utilities; EDB = Singapore Economic Development Board; IMDA = Infocomm Media Development Authority. Class is defined as sector, voltage level or connection capacity.

8.4 Key takeaways

The electricity price implications of data centre growth are highly context-specific and depend on factors such as the pace, location and system conditions under which new demand emerges, as well as on market design and regulatory frameworks. When large loads connect rapidly in constrained markets, they can place upward pressure on electricity prices. By contrast, when demand growth is anticipated and aligned with timely investment, operational flexibility and clear cost allocation, price impacts can be limited. Policy and planning choices by both data centre developers and system planners will therefore be central in determining whether data centre expansion translates into affordability pressures or is integrated efficiently into power systems.

9 How can AI enhance energy security and sustainability in energy, and what are the barriers to wider adoption?

AI is being deployed commercially in the energy sector for a range of objectives, including making energy systems more secure, competitive and sustainable. As the capabilities of AI systems have increased, so too have the opportunities. AI is now being used not only to process large datasets and predict what will happen but also to help enable real-time decision-making in the energy sector, such as by managing batteries and improving grid stability.

These improvements come at a time when the energy sector is facing new challenges. Concerns about energy security and supply chain resilience are high on the political agenda, as are energy affordability and industrial competitiveness. At the same time, the energy system is becoming more complex and dynamic. Electricity is increasingly used in energy consumption, and variable renewable sources are playing a bigger role in power generation. Batteries are also being rapidly installed to provide greater flexibility to both the grid and local users. Additionally, countries are working to achieve long-term climate, energy transition, efficiency and sustainability targets. While AI can contribute to these goals, there is still a lack of understanding of its full potential and the barriers to its wider use in the energy sector.

9.1 The role of AI in enhancing energy security and sustainability

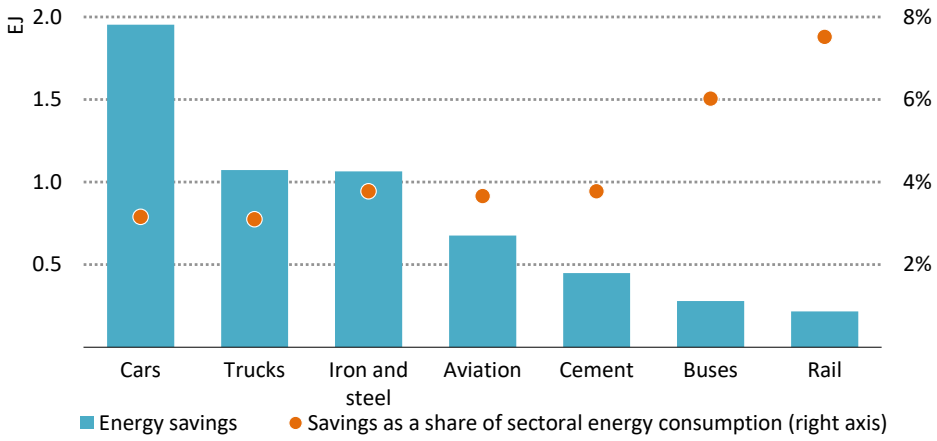
As geopolitical and supply chain risks mount, the role of AI in enhancing energy security, reliability, resilience and demand response has come into focus. AI is also being deployed to integrate renewables, enhance efficiency and accelerate innovation in new energy technologies. The IEA published real-world case studies of AI applications in the energy sector in its *Energy and AI Observatory*, as well as in a joint publication with India's Ministry of Electronics and Information Technology for the AI Impact Summit 2026. These use cases, many of which have been commercially deployed, illustrate the impact that AI is already having in the energy sector. Examples of these applications include:

- **Reliability and resilience**: AI-powered predictive maintenance reduces downtime in critical energy systems by enabling proactive interventions, unlike traditional schedule-based maintenance, which can result in unexpected failures and emergency repairs. In the United States, AI tools now monitor more than 10 000 transformers and 22 000 circuit breakers, significantly reducing outages (C3.ai, n.d.).

- **Demand response**: AI platforms are being deployed to help reduce electricity demand during periods of grid stress, thereby improving reliability at the system level. AI-enabled platforms can automate demand response. For example, AI has been deployed to automatically manage heating, ventilation and air conditioning devices based on real-time market conditions among more than 1.5 million devices in 250 000 buildings in eight European countries. This approach has unlocked 2 GW of demand response (IEA, 2025f).
- **Integration of variable renewable energy**: AI-enhanced weather forecasting is supporting the integration of variable renewables into grids and minimising curtailment. AI is also being deployed to automate the management of complex energy systems involving batteries and distributed solar resources. For example, in Chile, the use of an AI-enabled forecasting model that is up to 15% more accurate than previous industry standards has led to more effective grid management, greater integration of wind energy and reduced curtailment (IEA, 2025g). In India, an AI-enabled plant-specific solar forecasting platform has helped solar generators reduce regulatory penalties arising from inaccurate output forecasts (IEA, 2026a).
- **Energy efficiency and cost reductions**: AI solutions are increasingly being used to improve energy consumption in industry and buildings. In Brazil, for example, an AI tool used in iron and steel production draws on blast furnace data to forecast thermal conditions, enabling better decisions that lower energy and material use, as well as emissions per tonne of iron (IEA, 2026b).
- **Energy innovation**: AI can help accelerate parts of the innovation pipeline, which in turn can help make energy systems more secure and sustainable. For example, at a laboratory in the United States, AI helped accelerate the screening process for potential battery materials, predicting material properties 1 500 times faster than traditional density functional theory calculations. What would have typically required years of high-performance computing calculations and laboratory experimentation was reduced to just a few days using AI. One of the predicted candidates showed the potential to reduce lithium requirements in batteries by approximately 70% (IEA, 2025h). AI is also being used to support the development of nuclear fusion technologies (IRFM, 2025).

IEA analysis estimates that if existing AI applications in energy end-use sectors were scaled up to the sectoral level, they could unlock energy savings to the tune of 13.5 EJ by 2035, equivalent to Indonesia's annual energy demand today. Doing so, however, would require overcoming the various barriers that currently impede the deployment of AI at the sectoral level.

Figure 9.1 ▷ Global energy savings potential from AI in selected end-uses, 2035



Widespread adoption of known AI applications has the potential to save 13.5 EJ across all end-use sectors in 2035, greater than Indonesia's annual energy consumption today

Note: This energy savings potential is estimated using existing AI applications only and does not factor in solutions that might emerge in the future. A detailed discussion of this approach is available in IEA (2025a).

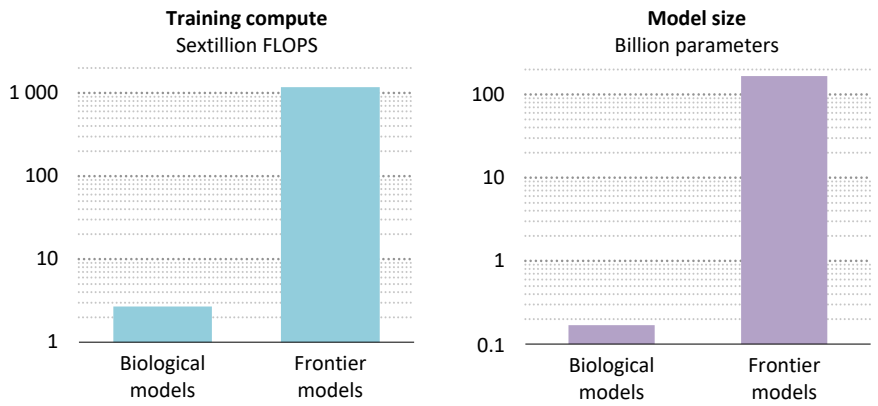
Box 9.1 ▷ What types of AI are used in innovation and energy sector optimisation?

Popular AI models used for generative tasks such as writing a few paragraphs or rendering an image are often called frontier models. These models are trained on vast databases of text, audio, videos and images. They are based on AI approaches such as deep learning. These models typically require large quantities of electricity for training and inference and are the main drivers of electricity consumption growth from data centres today.

In contrast, domain-specific models, such as those used in energy sector optimisation and scientific domains such as biology or engineering, are often smaller and trained on specific data. These models are deployed in scientific research, industrial design and the optimisation of complex economic systems such as factories or power grids. Their energy footprint is typically many orders of magnitude lower than that of frontier AI models. For example, the training of biological AI models requires only a small fraction of the compute used in frontier large language model training runs (Figure 9.2).

Today, many of the commercially available solutions in the energy sector still rely on classical AI approaches.[9] For example, several commercially available AI-led building energy management solutions leverage predictive analytics and automated heating, ventilation and air conditioning control through machine learning. However, AI solutions in the energy sector are increasingly benefiting from AI approaches such as neural networks and deep learning. For example, hyperlocal weather forecasting models that enable the integration of variable renewable energy into the grid may use a mixture of classical machine learning alongside deep learning.

Figure 9.2 ▷ Average training compute and average model size of frontier and biological AI models



Note: FLOPS = floating-point operations per second.

Sources: IEA analysis based on Epoch AI (2026a and 2026b).

While there are differences in training data, model size and energy consumption across such models, many innovations spill over between consumer-facing generative models and scientific models. For example, diffusion-based image generation models are finding applications in object recognition and robotics, accelerating progress in physical AI systems (see section 10). Core architectural innovations from generative AI, such as transformers, have also brought critical improvements to domains such as materials science by improving the prediction of material properties from material composition. Finally, frontier AI models, such as large language models, are also increasingly being integrated into scientific and energy sector workflows, helping researchers and system operators interact with large amounts of structured and unstructured data and text.

[9] IEA's 2025 *Energy and AI* report includes discussions on these AI approaches and archetypes.

9.2 Barriers to scaling up AI use in the energy sector

While the uptake of AI solutions in the energy sector is increasing, most successes so far have been individual case studies rather than system-wide deployments. Further scaling up of AI use requires overcoming barriers such as limited training data, concerns over cybersecurity and data privacy, insufficient policy support, the limited digitalisation of energy equipment and a shortage of AI-related skills in the energy sector.

Data availability, accessibility and quality

The adoption of AI in energy, and specifically within the electricity sector, is limited by data that are often fragmented, low-quality and inaccessible. Critical information on demand patterns, grid performance, outages and asset health is dispersed across legacy systems and organisational silos, with limited data sharing between public and private actors. Concerns over cybersecurity are often a contributing factor. Notably, grid operators and utilities face regulatory obligations concerning the security of operational technology, which limit data sharing and AI model deployment in real-time control systems.

Data across the energy value chain also lacks standardisation, making aggregation and interoperability for AI training difficult. Regulatory constraints and the absence of widely adopted open standards restrict the creation of clean, integrated datasets. Together, these factors form a structural bottleneck that prevents the wider deployment of AI for grid balancing, predictive maintenance, demand forecasting, fault detection and renewables integration, all of which would support broader energy goals.

In recent years, regulations have advanced to support new electricity data-sharing frameworks. These frameworks are increasingly being formalised, but they vary in scope and ambition. The European Union mandates smart metering, free access to consumption data via standardised interfaces, cross-border interoperability and AI-oriented energy data initiatives.

Policy, regulation and AI ambition

Policies and regulations designed to support data availability, interoperability and system modernisation can help enable the adoption of AI in the energy sector. For instance, energy efficiency regulations and mandates can incentivise energy consumers to pursue innovative approaches involving AI to make energy use more efficient across the buildings, industry and transport sectors. Frameworks such as EU energy efficiency reporting requirements create structured datasets on energy use that support AI model development, while in the United States, AI adoption in the energy sector is incentivised through federal initiatives that frame grid modernisation as essential to AI competitiveness, with the Department of Energy explicitly promoting AI for grid planning, permitting, operations and resilience (U.S. DOE, 2024).

Regulatory incentive structures in many electricity markets can also act as a barrier to the adoption of AI solutions in the energy sector. In many jurisdictions, regulated utilities earn

an approved rate of return on capital expenditure. This incentivises utilities to prioritise physical, asset-heavy investments over AI and software-based solutions, even where the latter could deliver comparable or lower-cost system benefits.

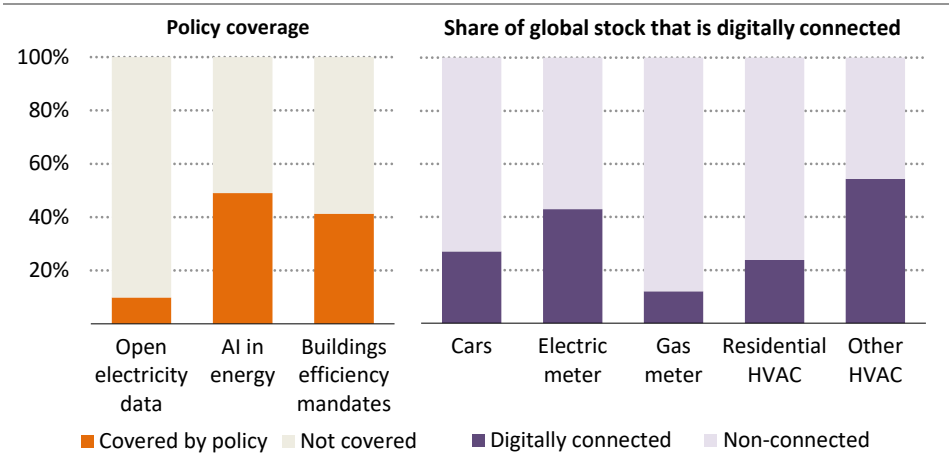
The central policy challenge is to ensure that AI governance, data policies, energy efficiency regulation, grid rules and financial incentives evolve in a co-ordinated way so that data mandates and digital standards reinforce one another and accelerate AI deployment at scale.

Digitalisation

AI models need high-quality and abundant data on which to train. Digitally connected devices and equipment are key enablers of the generation of such data. However, in the energy sector, the rollout of digitally connected devices, equipment, infrastructure, and diagnostic and monitoring tools has been uneven.

In buildings, smart meters, for example, can provide a valuable source of training data for use in AI applications. While smart metering for electricity is growing, accounting for an estimated 43% of all meters globally, metering for gas networks lags behind. Estimates indicate that smart gas meters account for between 10% and 15% of the total.

Figure 9.3 ▷ Indicators of policy coverage and digital connectivity for key energy consumption equipment and infrastructure



Policies supporting the adoption of AI applications in the energy sector are still lacking, while digital infrastructure coverage remains insufficient

Notes: HVAC = heating, ventilation and air conditioning. Policy coverage of open electricity data is the share of global electricity generation in countries where open electricity data policies exist. Policy coverage of AI in energy refers to the share of global energy demand in countries that have stated policies or ambitions for AI adoption in the energy sector. Policy coverage of buildings efficiency mandates refers to the share of global floor space that is compliant with building codes, including zero-carbon-ready buildings.

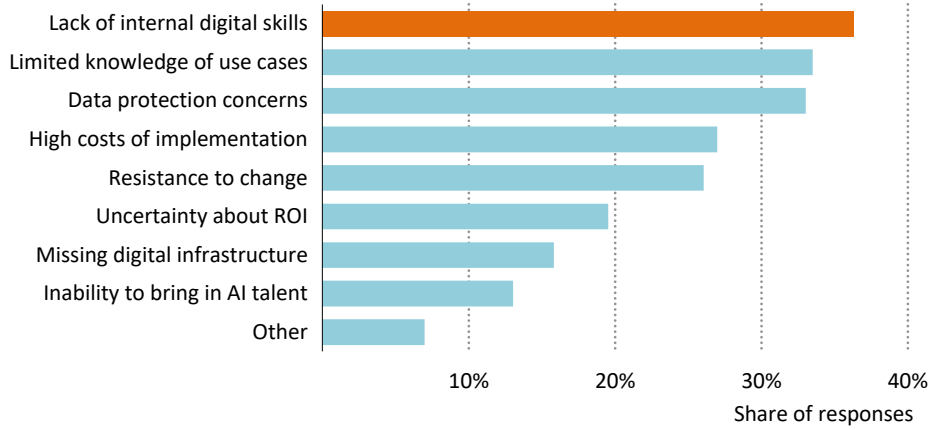
Sources: HVAC: IEA (2025a); Gas, electric meters: RCR Wireless (2024) and IoT Analytics (2024); Cars: IEA analysis based on Counterpoint Research (2025).

For cars, embedded telematic units are becoming more popular.[10] Estimates place the stock of digitally connected cars at over a quarter of the global fleet. The eCall mandate in the European Union, the European Free Trade Association and the United Kingdom has accelerated the adoption of on-board car diagnostics, and virtually all leading carmakers today have launched mass-market services for car diagnostics. This stands to further improve the quality and size of datasets available for car operations and maintenance diagnostics, which could further the applications of AI in transport, a sector which accounts for around 20% of global energy consumption.

Skills

Energy firms have identified the lack of digital skills as the single largest barrier preventing the greater adoption of AI-led optimisation. Compared with other industries, the energy sector lags behind in developing digital capabilities among its workforce. From 2018 to 2024, the proportion of AI professionals within utilities, oil, gas and mining was, on average, 40% lower than in industries such as education, financial services, technology, information and media (IEA, 2025i).

Figure 9.4 ▷ Responses in the IEA Industry Employment Survey 2025 to the question, "What is the greatest barrier to adopting AI and digital technologies in your day-to-day operations?"



The lack of digital skills has been identified as the single largest barrier to greater adoption of AI-led optimisation by energy companies

Note: ROI = return on investment.

Source: IEA (2025h).

[10] An embedded telematics unit is an onboard electronic device that connects a vehicle to external networks via cellular or other wireless links.

Several energy companies today rely on the consulting services of specialised AI companies to identify AI tools that can be used in their operations and help implement them. However, this might pose its own challenges, as specialised AI companies often lack knowledge of energy markets, regulation and other sector-specific factors. An alternative is to develop internal skills and upskill existing employees, although this approach also presents challenges, as AI is a fast-moving field that requires advanced and specialised technical expertise.

9.3 Key takeaways

To scale AI across the energy sector in ways that enhance energy security, boost system affordability and efficiency, and advance countries' sustainability goals, a unified approach to policy, regulation and market incentives is essential. Steps such as making data more accessible and interoperable, accelerating the digitalisation of energy assets, investing in workforce skills and developing policies to support AI applications are crucial. Without these conditions in place, AI use in the energy sector is likely to remain limited and restricted to small-scale pilot initiatives.

At the same time, the potential of AI applications needs to be kept in perspective. Near-term efficiency savings potential of around 13.5 EJ by 2035 represents substantial gains, but they are still small in the context of total energy consumption of around 450 EJ today. While AI is an important tool for achieving energy sector goals, it cannot eliminate the need for more traditional policies and approaches to improve efficiency, accelerate the adoption of new technologies, and advance affordability and sustainability.

10 What does the rise of physical AI and robotics mean for innovation and competitiveness in the energy sector?

Remarkable developments in AI software in recent years are beginning to translate into outcomes for physical AI systems that are designed to interact with the real world. Physical AI is embodied in hardware such as industrial robotic arms, robots used in warehouses, drones and autonomous vehicles. The rise of physical AI has implications for the energy sector, as it supports the rise of self-driving cars, can optimise industrial processes and has the potential to transform the production of key energy technologies, particularly solar PV and batteries, which are produced through mass manufacturing. This section explores the drivers behind growing interest in physical AI, historical and emerging trends in process automation and the growing potential of robots, along with possible consequences for the industrial and energy sectors.

10.1 How is physical AI evolving?

Several trends in AI and hardware are aligning to generate growing financial and scientific interest in the next generation of automation and robotics tools. These include:

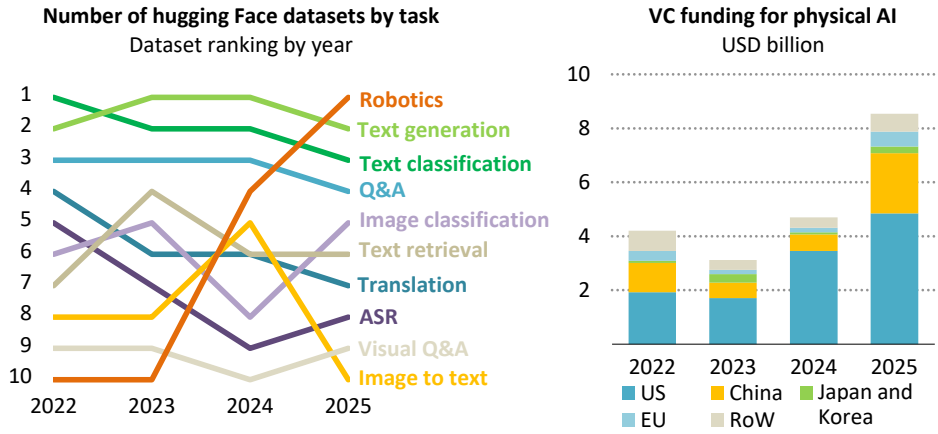
- **Falling costs for hardware and increasing capabilities of on-device computation**. Costs for critical components such as motors and sensors have fallen sharply, notably following the entry of Chinese manufacturers. Computational hardware, such as tensor processing units and neural processing units, is becoming increasingly available and efficient and is well suited for use in edge computing (i.e. computation performed on devices or on nearby servers). This is important for applications in which latency matters, such as autonomous driving. Software improvements have also enabled the use of smaller models that can run directly on devices.
- **Spillovers from generative AI that boost performance**. Innovations in generative AI are now enabling physical autonomous systems to understand the "semantics", or the meaning and logic, of the world they operate in. For example, an autonomous vehicle can use semantic understanding to predict the behaviour of others more effectively in novel situations (for example, a football rolling onto the street might be followed by a child running after it). Similarly, semantic understanding can be used to give high-level instructions to autonomous systems in natural language, without the need for these instructions to be encoded in a more detailed programming language.
- **Improved modelling of physics and integration of video, simulated and real-world data**. The development of world foundation models,[11] trained on video, simulated and physical data, is enabling autonomous systems to learn the laws of physics more intuitively. Better integration of simulated and video data with real-world data enables

[11] World foundation models are models designed to understand, simulate and predict the physical world. They aim to be generalisable across a range of physical environments by understanding the laws of physics and model the behaviour of agents (humans, animals or other autonomous systems) in the real world.

autonomous systems to be trained rapidly in simulated environments and then fine-tuned in the real world.

The combined effects of these developments can have important impacts on a range of industries. The automation of industrial processes, which are complex systems that cannot be represented by simple models, is becoming increasingly sophisticated, with the potential for high returns for early adopters that can make them more competitive. Robots are also attracting significant attention from investors and could transform several other industries, though the outlook for their deployment remains uncertain. Growing interest in this field can be seen in the rise of new datasets for AI training for robotics and the increase in venture capital in physical AI (Figure 10.1).

Figure 10.1 ▷ Creation of new datasets on Hugging Face by task and venture capital funding for physical AI by region, 2022-2025



IEA. CC BY 4.0.

Since the launch of ChatGPT, robotics and physical AI have attracted increasing levels of research and venture capital funding

Note: Q&A = question and answer; US = United States; EU = European Union; RoW = rest of world.

Sources: IEA analysis based on Praas, Sánchez, & Balland (2025) and Crunchbase (2026).

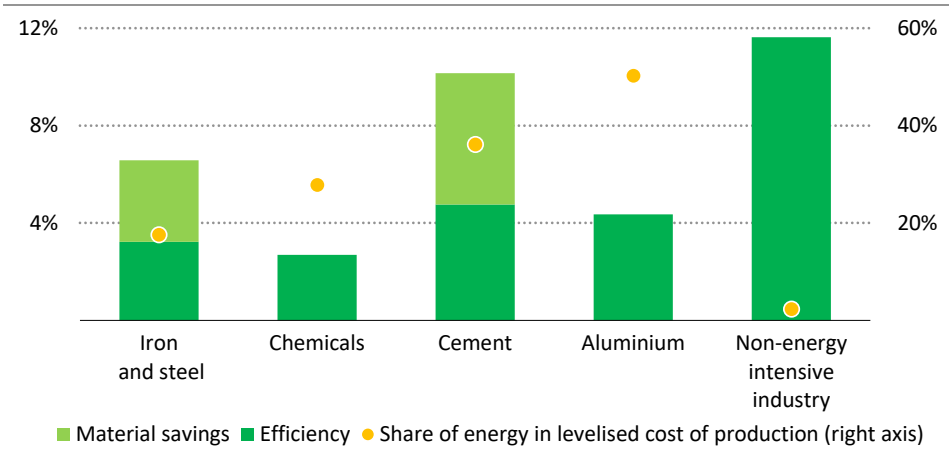
10.2 Can physical AI drive competitiveness by improving process automation?

For decades, the use of state-of-the-art computational technology to automate production processes has played a key role in maintaining industrial competitiveness. Early automation processes were based on simple actuators whose responses were defined in prescriptive ways in response to a small amount of external data. Over time, greater computational and hardware capabilities have enabled many industries to introduce more sophisticated predictive automation, including through early AI techniques like machine learning. This has

enabled a more advanced, predictive style of process automation in which multiple processes interact and learn from previous events.

These existing uses of AI have already demonstrated incremental opportunities for reducing energy demand, thereby improving industrial competitiveness. For instance, image detection can optimise material input and output quality and reduce waste. Machine learning applied to data collected from sensors across an industrial facility can also bring system-wide optimisation (IEA, 2025a). These incremental improvements can reduce energy costs by 3 to 10 percentage points in energy-intensive industries, where energy costs are an important factor and margins tend to be low. Because of short implementation times and payback periods, these tools offer opportunities to quickly improve competitiveness, especially in trade-exposed industries such as steel and aluminium. The opportunities to reduce energy costs are even greater in non-energy-intensive industries, although since energy costs contribute less to total production costs, the impact on competitiveness is less substantial. We estimate that total energy savings from the widespread application of AI tools across the industry sector could deliver 8 EJ of energy savings by 2035, although barriers remain (see section 9).

Figure 10.2 ▷ Energy cost reduction from existing AI applications, and share of energy in the levelised cost of production, 2024



AI can drive small but meaningful changes in energy costs, and these are largest for non-energy-intensive industry, but energy accounts for a smaller share of their production costs

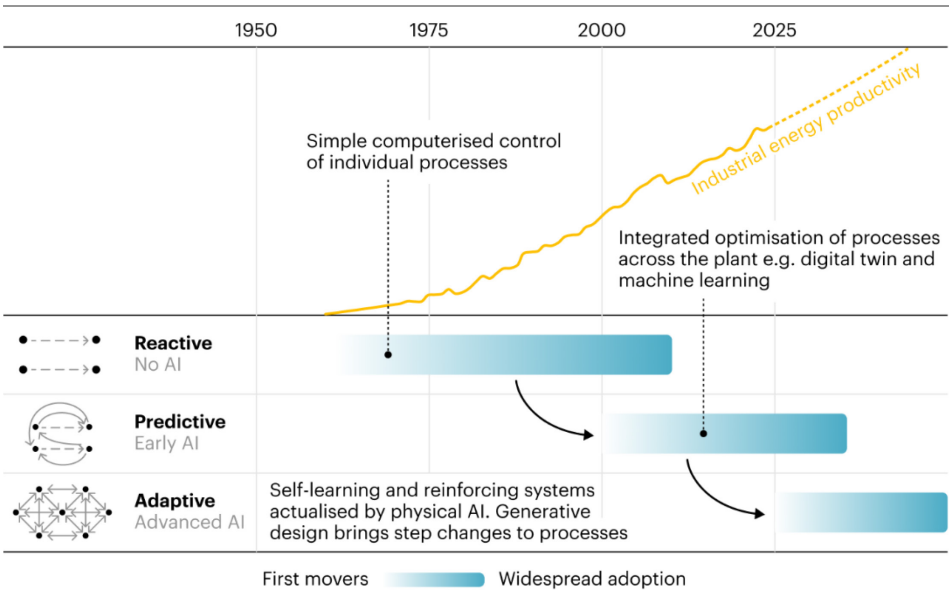
However, the growth of physical AI is increasingly leading to more flexible and high-performing process automation, driven by the growth in domain-specific models.[12] These

[12] Domain-specific models are trained on specific tasks, in specific environments. They contrast with models that are trained to generalise to a broader range of tasks. A domain-specific model would be one trained to optimise heating and cooling in a specific production environment, for example, based on historical data from the facility.

models are trained on context-specific data and often use physics-based models to improve accuracy. By processing terabytes of data generated every hour from thousands of sensors in industrial processes, far more than can be processed by simple systems or by onsite engineers, such models can learn to anticipate opportunities, bottlenecks and risks.

Once established, these domain models can bring industries into a new and more adaptive phase of automation. Tools that self-correct in real time now allow for the automation of complex tasks that were previously impossible. For example, light industries are increasingly electrifying heat to substitute or supplement their existing fossil fuel boilers. AI tools can already support electrification by optimising participation in electricity markets, but domain models mean they can integrate power market optimisation into the management of both flexible and inflexible onsite processes and respond to unexpected events when they occur. This adaptivity can make industries more resilient, more efficient in their use of energy, materials and labour, and ultimately more competitive.

Figure 10.3 ▷ Phases of industrial automation, 1950-2050



Industrial automation has a long history and increasingly uses AI, but new approaches could enable the next wave of productivity growth

Note: Industrial energy productivity is the ratio of industrial value added to industrial energy demand, shown here for advanced economies indexed to 1960.

Highly adaptive and automated processes could become increasingly important differentiators of industrial competitiveness across regions, especially in industries with high technological precision and extensive material processing. For instance, about half of the

difference between the cost of producing a battery in the European Union and in China is due to differences in production efficiency enabled by China's higher deployment of advanced automation techniques, which in turn lead to higher manufacturing yields (IEA, 2025j). Box 10.1 addresses China's AI strategy in more detail.

Despite the opportunities they offer for competitiveness, the deployment of advanced physical AI is neither straightforward nor inevitable (see section 9). Subject matter experts are needed to coordinate model training for domain-specific models, particularly because significant amounts of knowledge and data are not held in formats suitable for model training. The absence of fully digitised enabling hardware, the cost of developing domain models for individual sectors and plants, and a lack of awareness, trust and AI capability within some subsectors can also reduce appetite for these projects.

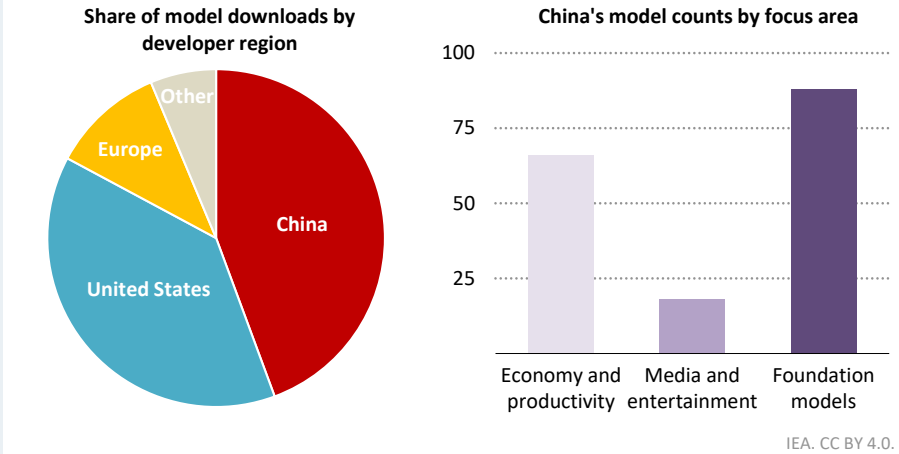
Box 10.1 ▷ How does China think about the role of physical AI in the energy and manufacturing sectors?

The United States remains the leader in so-called frontier AI models, as well as in the design of key AI hardware. However, after Chinese researchers surprised the AI world with the release of the highly capable model DeepSeek R1, there has been increased attention on China's AI strategy and progress, particularly regarding its application in manufacturing, given China's dominance in global manufacturing.

China's strategy can be characterised at a high level by three aspects:

- A strong focus on compute efficiency but lower hardware efficiency: China's model developers are compute-constrained due to a lack of access to high-end chips. They therefore focus strongly on techniques to reduce the use of compute during training and model use. At the same time, China's AI chips lag significantly behind cutting-edge chips from suppliers such as NVIDIA in terms of energy efficiency (SemiAnalysis, 2025b).
- A strong focus on open-source ecosystems: China's model developers tend to release their models as open source, enabling users to download and customise them as needed. This can be attractive for enterprise use in the services and industry sectors. In 2025, China's models overtook those from developers in the United States in terms of total downloads on open-source platforms.
- A focus on AI applications across the economy: China's national AI strategy focuses on increasing the use of AI across the economy to improve productivity and innovation. This can be seen in data on the primary focus of Chinese models registered with its telecoms regulator. To this extent, China does not view the "AI race" in terms of a race towards the best model or achieving artificial general intelligence (Chan, 2026). Rather, its strategy aims to drive the uptake of AI models in manufacturing, in science and innovation and across the economy more broadly. This can also be seen in the strong venture capital numbers for start-ups focused on physical AI in China.

Figure 10.4 ▷ Open source model downloads by region of developer, and China's AI model counts by focus area, 2025



China's AI strategy focuses on uptake across manufacturing, infrastructure and services in order to drive productivity and innovation

Source: IEA analysis based on data from Qian (2026).

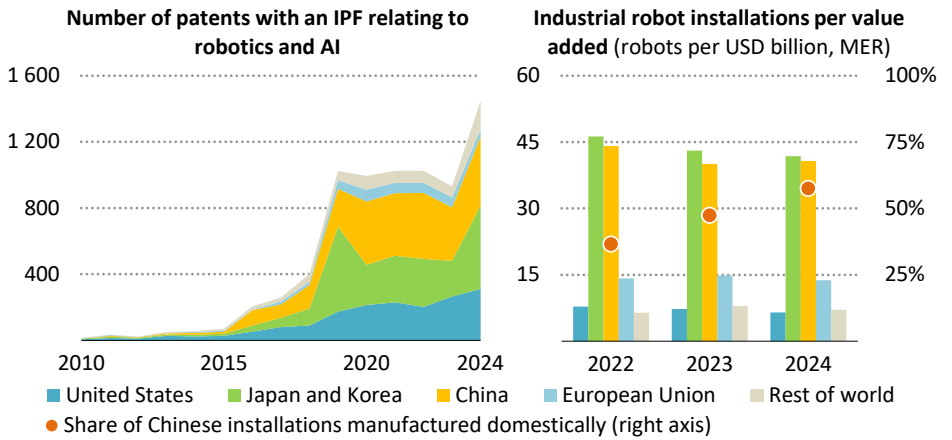
10.3 Will robots be the next technology to transform the industrial and energy sectors?

Industrial robotics and physical AI are attracting attention as a possible next frontier of AI deployment. Although the automation of physical tasks is already widespread in industry, AI could enable industrial robots to operate in a wider range of settings and with comparatively less supervision. Robotic dogs, for instance, have been deployed in cement facilities to perform routine inspections around plants in a repeatable and precise manner. Given the opportunities businesses see to enhance competitiveness, the adoption of robotics across the industrial and energy sectors could rise in the coming years.

Enthusiasm for robotics is reflected in increased venture capital funding for companies focused on physical AI, as well as in the rapid acceleration in the number of patents filed at the intersection of robotics and AI over the past decade (with a particular spike noted in 2024, the most recent full year for which data are available). However, this enthusiasm is not necessarily replicated in installations of industrial robots, which have stayed relatively stable. China is the largest market both for producing and deploying industrial robots. More than half of the 550 000 industrial robots installed in 2024 were in China, the majority of which were manufactured domestically. The economic conditions that gave rise to the mass manufacturing of energy technologies such as solar PV and electric vehicles position China well for a similar expansion into modular robot technologies. Beyond manufacturing robots, China is a leading supplier of robot parts, particularly actuators, which convert electrical

energy into physical motion and account for more than 50% of the content value of typical humanoid robots (Financial Times, 2025).

Figure 10.5 ▷ Patent applications for robots and AI, and industrial robot installations, by region, 2022-2025



IEA. CC BY 4.0.

Despite a significant increase in venture capital investment in physical AI in 2024, industrial robot installations are estimated to have remained steady

Notes: 2025e = 2025 estimate. IPF = international patent family, as determined by Cooperative Patent Classification codes combined with targeted keywords. Data for 2020 to 2024 are nowcast based on historical data with a variable factor for the main patenting countries and a fixed factor for the rest of the countries.

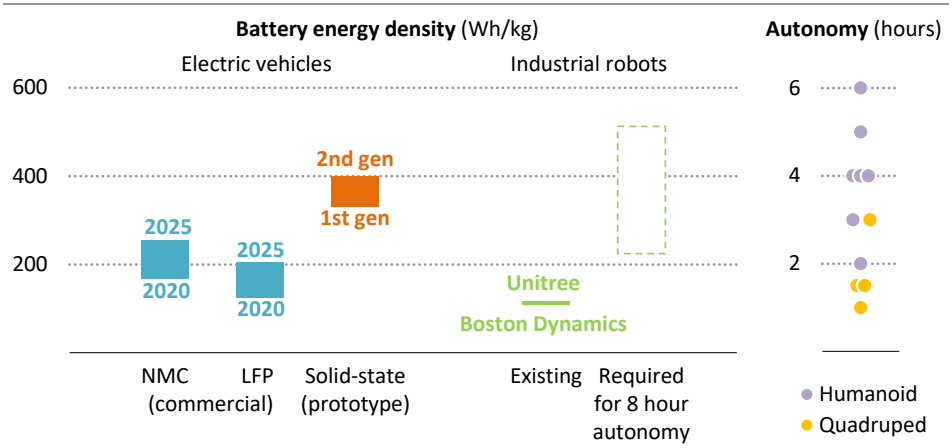
Sources: International Federation of Robotics (2025) and PATSTAT (2026).

The need for large upfront investments could act as a brake on the deployment of complex robots across the industrial and energy sectors, and beyond. Although prices are falling, multi-purpose industrial robots rely on expensive hardware and software, which means they are not yet competitive with human labour in many regions outside of very standardised processes. Existing humanoid robots for industrial use cost at least USD 100 000, although some producers have ambitions to reduce costs to around USD 20 000 to USD 30 000. Even humanoid robots, which can be integrated into spaces designed for humans, require software investment in domain models, which are complex to produce, particularly for small and medium-sized enterprises. “Cobots”, which work alongside humans, are emerging as one alternative approach. These robots can be deployed with less training since humans can help support their work in complex cases.

Battery technology may limit the durability of mobile industrial robots and the applications in which they can be used. While recent software developments in robotics have been remarkable, greater deployment will depend equally on hardware improvement and will benefit from battery R&D and supply chain performance. Some multi-purpose robots

operate for only one to two hours on a single charge, driving down utilisation factors and reducing their competitiveness. Beyond duration, many robots are intended to perform high-power tasks, such as lifting heavy objects, for which existing battery technologies are not well suited, and which pose additional challenges such as overheating. To achieve autonomy equivalent to that of a human shift worker, battery energy densities would need to increase by a factor of around two. At today's rate of battery improvement, this would take around 10 years. Battery swapping is a feasible alternative that could enable higher overall utilisation, but it is only practical in certain contexts, increases costs and may constrain battery technology selection because cells need to be arranged so that the pack can be removed easily from the robot's chassis. For these reasons, new battery technologies are likely to be required to enable the widespread uptake of industrial robotics. Leaders in battery development may therefore have an advantage in industrial robot design and deployment.

Figure 10.6 ▷ Battery energy density by chemistry and sector, and autonomy of commercial robots on a single battery charge today



Some existing industrial robots have short lifespans, and significant technical improvement is needed to enable 8-hour lifetimes

Note: NMC = lithium nickel cobalt manganese oxide; LFP = lithium iron phosphate. Battery energy density is reported at the pack level. "Required for 8-hour autonomy" refers to the energy density that would be required to enable existing robots to operate for 8 hours without other improvements to their energy efficiency, based on the average lifetime for the nominated battery type and assuming an existing density of 110 Wh/kg (around the average for industrial robots today).

Sources: IEA (2025j), Unitree (2026), and BostonDynamics (2026).

10.4 Key takeaways

Physical AI is advancing rapidly as falling hardware costs, more capable edge computing and new modelling approaches bring robotics, autonomous systems and domain-specific

industrial optimisation closer to large-scale deployment. Yet traditional AI applications that optimise production processes still offer substantial untapped potential, especially in energy-intensive industries such as steel and aluminium production, where existing tools can strengthen competitiveness by cutting energy costs by 3-10%. For advanced manufacturing processes where digitalisation is already prioritised and energy costs represent a smaller share of total production costs, emerging physical AI systems, such as increasingly adaptive robots, could provide productivity gains.

Robot deployment has largely stagnated in recent years, with global installations hovering at around 55 000 units in 2024, but the rise of physical AI and a step change in capabilities could alter this trend, with investor interest already rising and the number of robotics and AI patents surging. From an energy perspective, efficient robots can increase the competitiveness of clean energy technologies, and advances in battery technologies will play a key role in improving the efficiency of robots by extending their operating time.

11 If AI drives a boom in productivity and GDP, what would it mean for energy demand?

11.1 AI's impact on productivity: From micro-level gains to macro effects

Labour productivity growth has slowed markedly across advanced economies over the past two decades.[13] In many advanced economies, productivity growth has trended downward since the early 2000s, even in the United States, where performance has been comparatively stronger. Against this backdrop, recent rapid advances in AI have raised expectations that the slowdown in productivity growth could begin to reverse.

Micro-level evidence provides a consistent signal: AI tools can substantially increase worker efficiency across a wide range of tasks. Across controlled trials and field deployments, AI has repeatedly delivered sizeable efficiency improvements (Table 11.1), typically reducing completion times for writing, customer support, software development, legal drafting and translation tasks by 15-50%.

Table 11.1 ▷ Micro-level productivity gains (task and firm level) from AI

Task type	Productivity gain	Source
Professional writing	• Tasks completed 40% faster	Noy et al. (2023)
Software engineering	• Task completed 55.8% faster	Peng et al. (2023)
	• +26.1% completed tasks	Cui et al. (2025)
	• Task completed 19% faster	Becker et al. (2025)
Consulting deliverables	• +12.1% completed knowledge tasks, delivered 25.1% faster	Dell'Acqua et al. (2023)
Customer support	• +15% completed tasks	Brynjolfsson et al. (2025)

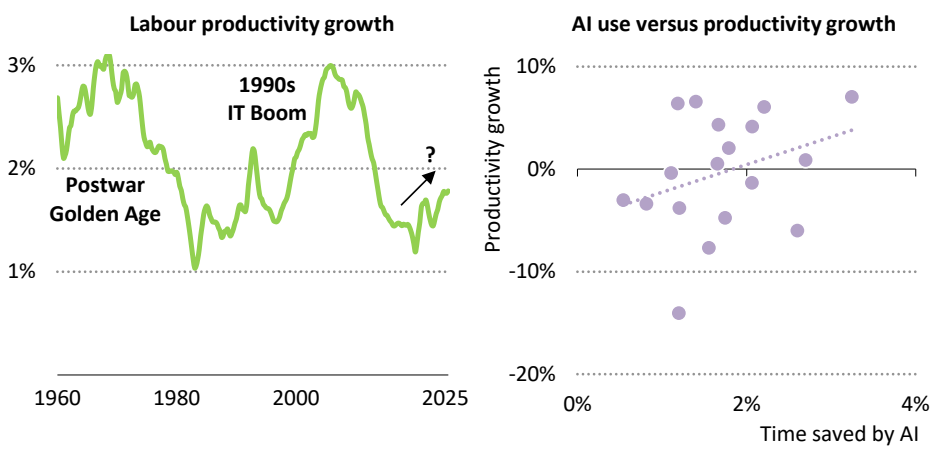
Labour productivity in the United States grew at an annualised rate of 4.9% in the third quarter of 2025, one of the strongest readings in two decades, while unit labour costs fell for a second consecutive quarter (Figure 11.1). While this recent acceleration has coincided with the diffusion of AI tools across firms, its underlying drivers remain uncertain. Productivity growth has historically been stronger in more digitally exposed sectors. The growing uptake of AI, with around 20% of US firms now using AI in their operations (U.S. Census Bureau, 2026), together with evidence of time savings from generative AI use, suggests that AI could be starting to contribute to measured productivity growth.

Beyond near-term productivity effects, estimates of AI's potential impact on GDP vary widely, reflecting differing assumptions about which tasks the technology will be used for, the speed of adoption and the scale of complementary investment. Even so, they typically project that AI will provide a boost to economic output (OECD, 2025). Some projections are extremely ambitious. Korinek and Suh (2024) explore stylised scenarios in which AI raises

[13] Labour productivity measures the output per hour worked.

GDP by 2% per year under "business-as usual" assumptions, with much higher growth rates (18% annually) arising if artificial general intelligence emerges. More modest, but still optimistic, assessments come from Acemoglu (2024), who estimates that AI could increase GDP by 0.92% to 1.6% in total over the next ten years, corresponding to roughly 0.55 percentage points of additional productivity growth. Goldman Sachs (2023) estimates that AI could raise global GDP by 7% over a ten-year adoption period, driven by an increase in annual productivity growth of around 1.5 percentage points. Aghion and Bunel (2024) place AI's impact between these estimates, suggesting that AI could boost aggregate productivity growth by around 0.8 to 1.3 percentage points per year. McKinsey Global Institute projects an even larger contribution, suggesting that AI and other automation technologies could lift global productivity growth by 0.5 to 3.4 percentage points per year over the period through to 2040 (McKinsey, 2023). The emerging consensus is that AI's impact will depend critically on the breadth and speed of adoption and on the availability of complementary assets such as data, compute capacity, organisational capabilities and human capital.

Figure 11.1 ▷ Labour productivity growth in the United States, 1960-2025, and correlation between post-Covid-19 labour productivity growth and time savings from AI use by sector in the United States



US labour productivity has strengthened recently, with early data suggesting that AI-related time savings may be contributing to stronger performance in some sectors

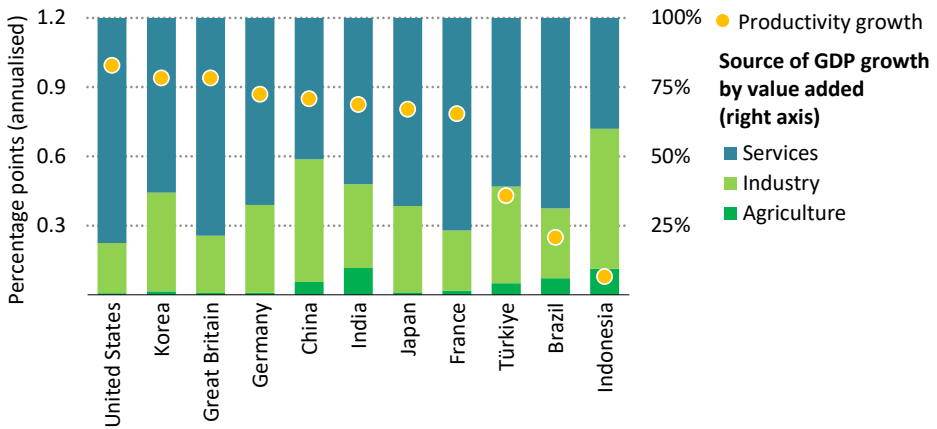
Sources: IEA analysis based on data from BLS (2026) and Bick et al. (2025).

While most previous studies have focused on the United States, OECD analysis (OECD, 2026) provides one of the first quantitative assessments of how AI-driven productivity improvements may translate into income gains across countries once sectoral and trade linkages are considered. This approach recognises that a boost in productivity in one sector

may translate into higher demand for the outputs of another, and that this higher demand may in turn be met through increased international trade. It starts from experimental evidence on how AI affects task performance, maps this to sector-level productivity improvements based on each country's economic structure and exposure to AI-relevant tasks and then evaluates how these effects propagate across sectors within national economies and across borders via trade.

In the central scenario, which assumes a medium pace of economic adoption of AI tools and improvements in AI capabilities consistent with current trends, AI is estimated to raise per capita real income growth by around 0.1 to 0.95 percentage points per year across major economies over the next decade. The largest gains occur in countries where knowledge-intensive service sectors, such as information and communications technology, finance and professional services, make up a significant share of the economy and where conditions for AI adoption, including digital infrastructure, skills and innovation capacity, are strongest (Figure 11.2).

Figure 11.2 ▷ AI-induced productivity and macroeconomic gains over the next 10 years in selected large economies



AI-driven gains are strongest where AI-intensive sectoral structures, high adoption rates and deep integration in global value chains align to support productivity impacts

Source: IEA analysis based on OECD (2026).

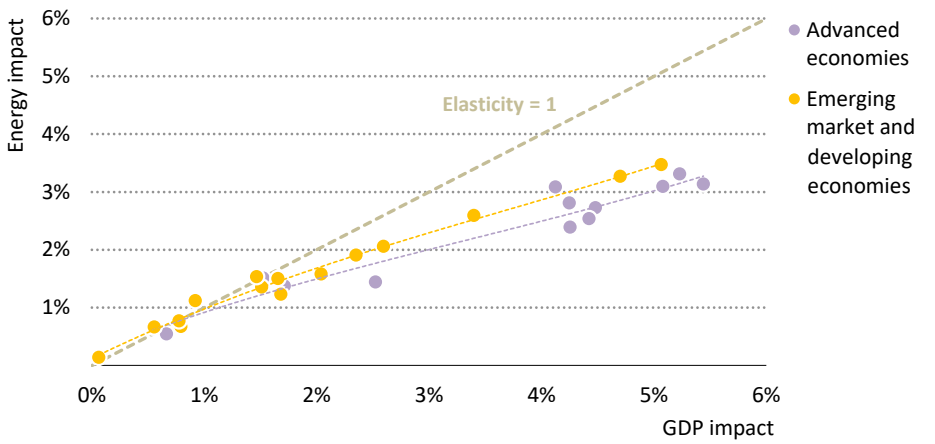
11.2 What could be the impact of higher productivity from AI on energy demand?

Macroeconomic trends are an important driver of energy demand. Since 2000, energy demand has increased by an average of 1.9% annually, compared with average GDP growth of 3.3% per year. As economic growth supports rising incomes, people tend to use existing

goods more, driving cars more often, spending more in the services sector and purchasing additional products.

However, the impact of faster economic growth on energy demand varies significantly across sectors and countries, reflecting differences in the responsiveness of energy service demand to rising incomes (Figure 11.3). In advanced economies, per capita energy demand peaked in the mid-2000s and has since fallen by 15%. Over the same period, GDP per capita increased by one-quarter. In these economies, structural shifts in the economy towards services and improved technologies have led to a saturation in energy demand. In contrast, in emerging markets and developing economies, rising incomes translate more directly into higher demand for energy-intensive goods and services, and energy demand has continued to increase. Similarly, there are wide differences between countries in the deployment of technologies and their efficiency. In analysing the potential impacts of an AI-driven GDP boost on energy demand, capturing these dynamics at a granular regional and sectoral level is essential.

Figure 11.3 ▷ Total demand elasticity by region in the high GDP case, 2035



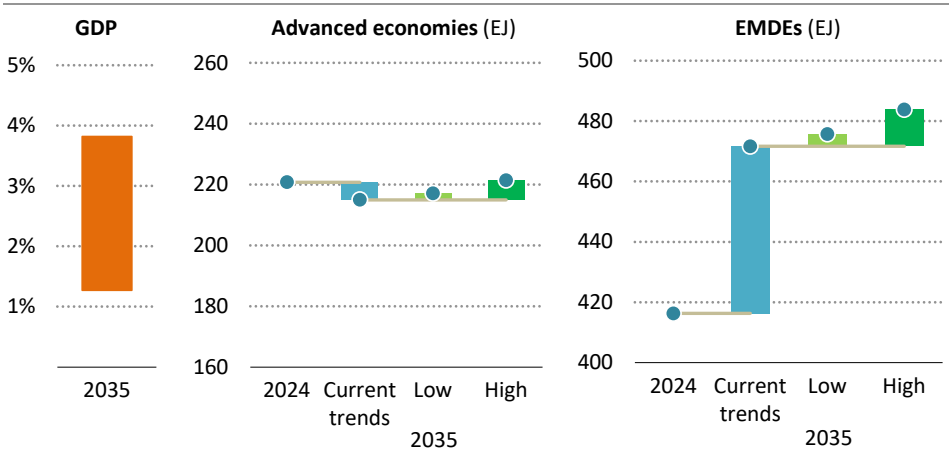
The impact on energy demand is lower than the increase in GDP in most regions, especially in regions with a higher GDP impact

Source: IEA analysis with GDP data from OECD (2026).

Based on the OECD economic modelling described above, AI adoption could lift global GDP by 1.3-3.8% above its baseline level in 2035 (OECD, 2026). For this report, we use IEA energy demand models for each major economy to estimate the effects of this additional GDP growth on energy demand. We estimate that global energy demand could rise by 0.9% to 2.6%, or by 6 EJ to 19 EJ, relative to current trends by 2035. This range of additional demand is comparable to the current annual demand of the United Kingdom and Japan, respectively.

In advanced economies, total energy demand is projected to continue its declining trend in a scenario aligned with today's policy settings, even as GDP continues to grow. In our cases with an additional GDP increase due to AI, energy demand in advanced economies receives a modest boost. At the upper end of our AI growth cases, this boost is sufficient to offset the decline in energy demand seen in our baseline scenarios, holding energy demand at its current level (Figure 11.4).

Figure 11.4 ▷ Impact of higher AI-related productivity on GDP and energy demand by economic grouping by 2035



An AI-related GDP increase could lead to 0.9% - 2.6% higher energy demand by 2035, offsetting declines in advanced economies and accelerating growth in EMDEs

Notes: EMDE = emerging markets and developing economies. Low and high refer to the upper and lower range of the GDP impact.

Source: IEA analysis with GDP data from OECD (2026).

In emerging markets and developing economies, the additional GDP in our AI scenarios is concentrated largely in China and India. This is due to the presence in both countries of knowledge-intensive services (which benefit directly from the AI productivity boost) and manufacturing (which benefits indirectly through higher global demand). Overall, energy demand is projected to grow strongly in emerging markets and developing economies in a scenario aligned with today's policy settings, increasing by around 13%. At the high end of our AI cases, this increase rises to 16%.

There are, however, several sources of uncertainty that could limit the scale of AI-related impacts on energy demand. In macroeconomic projections, the magnitude of additional GDP growth attributable to AI is difficult to isolate from the growth already embedded in baseline projections. Productivity gains could also have negative implications for labour markets, which could place downward pressure on macroeconomic outcomes. In the energy system, higher incomes do not automatically translate into proportionally higher energy use.

Households and firms may invest in more efficient technologies, replacing appliances, air conditioners or vehicles earlier than assumed in current models, or they may switch to electric mobility, all of which would temper the rise in energy demand. At the same time, widespread deployment of existing AI applications could generate substantial energy savings within the energy system itself. Optimisation of industrial production processes, advanced building energy management systems and improved transport routing could all reduce energy consumption (Box 11.1). These savings could be amplified further if AI enables disruptive innovations that reshape processes and technologies across multiple sectors.

Box 11.1 ▷ Can we estimate the net impact of AI on the global energy system?

In addition to quantifying the additional energy demand resulting from potentially higher macroeconomic productivity gains, the IEA has already quantified other aspects of the energy and AI nexus. These include energy demand from data centres, the energy savings potential from widespread AI deployment and rebound effects from lower energy prices. Nonetheless, aggregating these estimates to determine the overall impact of AI is complex, as the underlying measurements vary significantly in their scope and level of uncertainty.

The estimates with the highest degree of confidence and evidence are the projections on the future demand of data centres. These rely on a bottom-up approach that considers shipments of different types of IT equipment, complemented by detailed tracking of data centre projects. The outlook reflects rising demand for a broad range of server classes alongside anticipated improvements in server efficiency. Data centre electricity demand is projected to rise from around 500 TWh today to between 700 TWh and 1 700 TWh in 2035, depending on the case. Uncertainties remain around the pace of AI adoption and the rapid evolution of model and hardware efficiency. The robustness of the modelling framework and the growing systemic importance of data centres mean that these projections are fully integrated into the long-term scenarios of the *World Energy Outlook*.

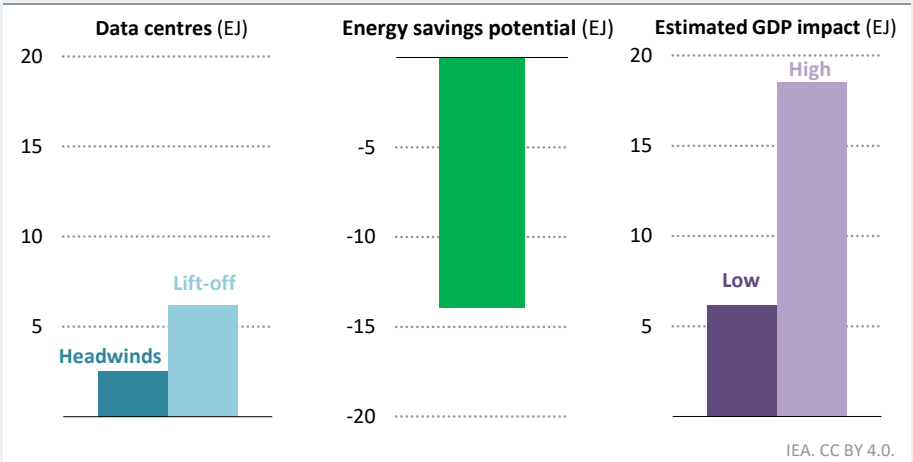
The quantification of potential energy savings from the adoption of known AI optimisations in end-use sectors is based on well-documented empirical case studies. However, extrapolating from case studies to estimates of sector-wide potential is uncertain. To reach the potential estimated savings of around 14 EJ, AI solutions would need to reach high deployment rates over the next ten years. Important barriers to reaching these rates include progress on regulation, digitalisation, trust and skills. There is also further uncertainty about a potential rebound effect following these savings if AI-driven savings lower the cost of energy services and therefore ultimately result in higher demand.

Estimating the impact of AI-related productivity gains on energy demand across all sectors is challenging and highly uncertain. First, as noted above, models of AI's impact on GDP produce a wide range of outcomes and depend on assumptions regarding AI capabilities and uptake, as well as macroeconomic uncertainties, such as the impact on

labour markets and incomes. Second, higher economic activity does not translate directly into higher energy demand, as the effects are influenced by technology choices, consumer behaviour, innovation trends and policies. These factors result in a wide range of potential outcomes for the boost to energy demand in 2035, ranging from 6 EJ to 19 EJ.

For these reasons, the effects of data centres on electricity demand are already integrated into the modelling framework and scenarios of the IEA's *World Energy Outlook*. The cases analysing the potential impact of an AI-driven economic boost on energy demand are exploratory in nature and are included in this report to provide stakeholders with a framework for considering the key determinants, uncertainties and possible orders of magnitude. Neither these estimates, nor the IEA's estimates of the energy-saving potential of AI applications, are included in the scenarios of the *World Energy Outlook*.

Figure 11.5 ▷ Impact of different AI-related effects on global energy demand, 2035



Notes: Low = low GDP additionality, high = high GDP additionality. The quantification of potential energy savings refers to the widespread scaling of existing AI applications across different end-use sectors as assessed in previous IEA analysis (IEA, 2025a).

11.3 Key takeaways

AI has the potential to lift productivity and accelerate economic growth, but the scale and timing of any macroeconomic boost depend on how quickly AI is adopted, how effectively firms reorganise to use these tools, and whether the required complementary investments in skills, data and digital infrastructure materialise. The implications for energy demand are similarly complex. Higher activity does not necessarily translate into proportionally higher

energy use, as technology choices, efficiency improvements, innovation and policy frameworks all shape the outcome. The analysis presented here indicates that, if AI adoption were to lift global GDP levels by around 1.3-3.8% above the baseline in 2035, as modelled by the OECD, global energy demand could increase by roughly 0.9-2.6% at that horizon compared with the baseline outlook.

Other AI-related trends monitored by the IEA will also affect outcomes. These include rising electricity needs from the expansion of data centres and potential efficiency improvements across end-use sectors as energy services become increasingly automated and responsive. As highlighted in this analysis, these elements cannot simply be combined to estimate a single net impact of AI on the energy system. The IEA will continue to track these changes and enhance the evidence base, helping to improve modelling methods and ensure that long-term scenarios incorporate AI-related shifts with appropriate recognition of the key uncertainties.

ANNEXES

Methodology and data tables

Methodology for data centre energy demand

The modelling of data centre electricity demand relies on a bottom-up approach developed by the Lawrence Berkeley National Laboratory over the past two decades. In this modelling approach, IT equipment shipments are the key driver of data centre electricity demand. We analyse three types of IT equipment: servers, storage systems, and network equipment. The last category refers to network equipment hosted within data centre facilities to connect servers and storage devices to the data network. It should not be confused with the data transmission network, which connects data centres and end-users (for example, 5G network towers). The latter falls outside the scope of the modelling of data centre electricity consumption in this study.

The central input to the model is the annual shipment of servers. These come from:

- IDC's (International Data Corporation), which provides shipment projections for the period 2019-2028 (IDC, 2024)
- These are triangulated with additional data inputs from Omdia (OMDIA, 2025), (OMDIA, 2026), SemiAnalysis (SemiAnalysis, 2025c), and Borderstep Institute (Hintemann, Hinterholtzer, and Konrat, 2024), and additional literature (Koomey, 2007), (Koomey, 2011), (Shehabi, et al., 2024), (Shehabi, et al., 2018), (Shehabi, et al., 2016), (Gartner, 2014a, 2014b, 2014c, 2015a, 2015b, 2015c, 2015d, 2016a, 2016b, 2017a, 2017b, 2017c, 2018a, 2018b, 2020), (Masanet, et al., 2020), (Malmodin, et al., 2024).

The stock of storage drives is derived from hard-disk drive shipment data from Forbes (Forbes, 2021) and the split between HDDs and solid-state drives from (SSDs) (Shehabi, et al., 2024). The stock of network equipment is estimated based on server port density.

We distinguish three types of data centres, which serve as archetypes in this model: enterprise data centres, colocation and service provider data centres, and hyperscale data centres.

The technical characteristics of the server stock, such as lifetime and power consumption, as well as operational characteristics like idle power ratio and utilisation rates, are based on estimates from the United States (Shehabi, et al., 2024). Similarly, for storage drives, the split between storage technologies and average utilisation rates is also based on US estimates. The characteristics of storage drives are assumed to be constant across all data centre types. The network port distribution is also assumed to be constant, with one exception: specific InfinityBand-like network equipment, whose stock depends solely on accelerated servers.

Based on these datasets and input assumptions, we estimate the installed capacity for each type of IT equipment. It is important to note that these values differ from the maximum designed capacity, as they consider only the installed units of each IT equipment type and do not reflect total rack capacity.

The regional allocation of global installed IT capacities relies on several factors. The primary driver is the regional breakdown provided by our third-party data provider (IDC), which is based on market dynamics in each region. To achieve finer regional granularity, we also consider the level of digitalisation of economies based on the digital adoption index (World Bank, 2016) and the development of the local data centre market, using publicly available data on data centre market revenues (Statista, 2024), (Turner&Townsend, 2024), and project databases (OMDIA, 2025), (BNEF, 2025).

Network equipment is assumed to have a 100% utilisation rate. Storage systems utilisation rates are considered constant. Idle power assumptions are based on trends observed in the SERT database for conventional servers and estimates from the literature for accelerated servers (SPEC, 2024), (Shehabi, et al., 2024).

Aggregation of utilisation rates is conducted by data centre type. The equation for the electricity consumption of servers is as follows:

$$E = (P_{max} - P_{idle}) * u + P_{idle}$$

Where:

- P_{max} is the maximum power draw of an operating server (distinct from the maximum rated power, especially for accelerated servers).
- P_{idle} is the power drawn by a server when not processing useful tasks.
- u is the server utilisation rate.

For each region and data centre type, IT electricity demand is multiplied by the corresponding Power Usage Effectiveness (PUE) to obtain the total electricity demand of the infrastructure and hosted IT equipment.

PUE primarily accounts for cooling equipment, power supply equipment, and lighting. Power supply equipment and lighting are collectively referred to as “auxiliary equipment”. Data centre type influences PUE due to variations in infrastructure efficiency, climate also affects PUE by directly impacting cooling requirements. PUE estimates are based on regional climate and data centre type (enterprise, colocation and service provider, and hyperscale) (Lei and Masanet, 2022). We assume that regional differences within the same data centre category arise from variations in cooling needs. The relative evolution of PUE over time is informed by improvements reported in company-level data (Google, 2025c).

The simplified equation for data centre electricity demand in each region is as follows:

$$E_{data\ centre} = \sum_{i\ =\ data\ centre\ type} (E_{server,i} + E_{storage,i} + E_{network,i}) * PUE_i$$

Data tables

General note to the tables

This annex includes the following datasets:

- **Table A.1 - World Data centres by case:** Includes global historical and projected data by case and data centre type (hyperscale, colocation and service provider and enterprise) for the following metrics:
  - Total and IT installed capacity (GW)
  - Power usage effectiveness
  - Capacity factor (%)
  - Total and IT electricity consumption (TWh)
- **Table A.2 - Data centres installed capacity by region:** Includes regional historical and projected total and IT installed capacity (GW) for the Base Case
- **Table A.3: Data centres power usage effectiveness and load factor by region**
- **Table A.4: Data centres electricity consumption by region**

Tables A.2 A.3 and A.4 include data for these regions: world, North America, United States, Central and South America, Europe, Africa, Middle East, Asia Pacific and China. The definitions for regions are in Annex B.

Both in the text of this report and in these annex tables, rounding may lead to minor differences between totals and the sum of their individual components.

Annex A licencing

Table A.1: World data centres by case

				Base		Lift-Off		High Efficiency		Headwinds	
	2023	2024	2025	2030	2035*	2030	2035*	2030	2035*	2030	2035*
Installed capacity (GW)											
Total	**83**	**97**	**114**	**226**	**277**	**241**	**383**	**205**	**231**	**200**	**218**
Hyperscale	31	36	44	85	103	91	148	88	105	61	64
Colocation and service provider	27	35	41	86	116	92	160	90	117	74	88
Enterprise	25	27	29	54	58	57	75	27	9	65	66
IT	**57**	**68**	**82**	**174**	**228**	**185**	**317**	**164**	**204**	**148**	**174**
Hyperscale	27	31	39	77	94	83	136	79	97	56	58
Colocation and service provider	17	23	28	65	96	69	132	68	101	55	73
Enterprise	13	14	16	32	38	33	49	16	6	38	43
Power usage effectiveness											
Total	**1.43**	**1.41**	**1.38**	**1.29**	**1.21**	**1.29**	**1.21**	**1.24**	**1.13**	**1.33**	**1.25**
Hyperscale	1.15	1.14	1.13	1.10	1.09	1.10	1.09	1.10	1.09	1.10	1.09
Colocation and service provider	1.56	1.53	1.49	1.33	1.21	1.33	1.21	1.32	1.15	1.33	1.21
Enterprise	1.95	1.92	1.89	1.71	1.54	1.71	1.53	1.67	1.46	1.72	1.54
Capacity factor (%)											
Total	**49**	**49**	**48**	**48**	**49**	**48**	**49**	**48**	**50**	**48**	**49**
Hyperscale	54	53	52	51	52	50	50	51	52	51	52
Colocation and service provider	48	48	47	47	48	47	48	47	48	47	49
Enterprise	46	45	45	45	46	45	47	45	49	45	47
Electricity consumption (TWh)											
Total	**360**	**416**	**485**	**945**	**1 193**	**1 008**	**1 637**	**868**	**1 013**	**833**	**942**
Hyperscale	148	166	200	378	466	404	652	390	480	275	292
Colocation and service provider	112	144	170	355	492	380	679	371	495	303	378
Enterprise	100	106	115	213	234	225	306	107	37	255	272
IT	**252**	**295**	**352**	**732**	**985**	**782**	**1 356**	**698**	**895**	**625**	**755**
Hyperscale	129	146	177	342	427	366	597	354	440	249	268
Colocation and service provider	72	94	114	266	406	285	559	280	430	228	311
Enterprise	51	55	61	124	152	131	199	64	25	148	177

*2035 numbers serve as exploratory scenarios given the high level of uncertainty around data centre demand growth.

Table A.2: Data centres installed capacity by region

	2023	2024	2025	Base Case 2030
Total installed capacity (GW)				
World	83	97	114	226
North America	35	43	53	102
United States	35	42	52	100
Central and South America	0.3	0.3	0.4	0.7
Europe	15	16	17	27
Africa	0.3	0.4	0.4	0.7
Middle East	0.3	0.4	0.4	0.7
Asia Pacific	30	36	41	92
China	20	24	28	67
IT installed capacity (GW)				
World	57	68	82	174
North America	26	32	40	82
United States	26	31	39	81
Central and South America	0.2	0.2	0.2	0.4
Europe	10	11	12	21
Africa	0.2	0.2	0.2	0.5
Middle East	0.2	0.2	0.2	0.4
Asia Pacific	19	24	28	67
China	13	16	19	49

Table A.3: Data centres power usage effectiveness and capacity factor by region

	2023	2024	2025	Base Case 2030
Power usage effectiveness				
World	1.43	1.41	1.38	1.29
North America	1.32	1.32	1.30	1.24
United States	1.31	1.32	1.30	1.23
Central and South America	1.73	1.70	1.67	1.50
Europe	1.47	1.45	1.42	1.29
Africa	1.85	1.81	1.77	1.59
Middle East	1.96	1.92	1.88	1.70
Asia Pacific	1.55	1.50	1.46	1.35
China	1.56	1.50	1.47	1.35
Capacity factor (%)				
World	49	49	48	48
North America	51	50	49	48
United States	51	50	50	49
Central and South America	48	47	47	46
Europe	49	48	48	48
Africa	46	46	45	45
Middle East	46	46	45	45
Asia Pacific	48	48	48	47
China	48	48	48	47

Table A.4: Data centres electricity consumption by region

	2023	2024	2025	Base Case 2030
Total electricity consumption (TWh)				
World	360	416	485	945
North America	158	187	229	434
United States	154	183	224	426
Central and South America	1.2	1.4	1.5	2.7
Europe	66	68	72	114
Africa	1.3	1.4	1.6	2.9
Middle East	1.3	1.5	1.7	3.0
Asia Pacific	128	150	173	378
China	84	102	117	277
IT electricity consumption (TWh)				
World	252	295	352	732
North America	120	142	176	351
United States	117	139	173	345
Central and South America	0.7	0.8	0.9	1.8
Europe	45	47	51	88
Africa	0.7	0.8	0.9	1.8
Middle East	0.7	0.8	0.9	1.7
Asia Pacific	82	100	118	281
China	54	68	80	205

Definitions

This annex provides general information on terminology used throughout this report including: units and general conversion factors; definitions of fuels, processes and sectors; regional and country groupings; and abbreviations and acronyms.

Units

Area	km^2	square kilometres
	Mha	million hectares
Distance	km	kilometre
Emissions	ppm	parts per million (by volume)
	t CO_2	tonnes of carbon dioxide
	Gt CO_2-eq	gigatonnes of carbon-dioxide equivalent (using 100-year global warming potentials for different greenhouse gases)
	kg CO_2-eq	kilogrammes of carbon-dioxide equivalent
	g CO_2/km	grammes of carbon dioxide per kilometre
	g CO_2/kWh	grammes of carbon dioxide per kilowatt-hour
	kg CO_2/kWh	kilogrammes of carbon dioxide per kilowatt-hour
Energy	MJ	megajoule (1 joule x 10^6)
	GJ	gigajoule (1 joule x 10^9)
	TJ	terajoule (1 joule x 10^{12})
	PJ	petajoule (1 joule x 10^{15})
	EJ	exajoule (1 joule x 10^{18})
	W	watt (1 joule per second)
	kW	kilowatt (1 watt x 10^3)
	MW	megawatt (1 watt x 10^6)
	GW	gigawatt (1 watt x 10^9)
	TW	terawatt (1 watt x 10^{12})
	kWh	kilowatt-hour
	MWh	megawatt-hour
	GWh	gigawatt-hour
	TWh	terawatt-hour
	MBtu	million British thermal units
Energy density	Wh/kg	watt hours per kilogramme
Energy equivalence	boe	barrel of oil equivalent
	mboe/d	million barrels of oil equivalent per day
	toe	tonne of oil equivalent
	ktoe	thousand tonnes of oil equivalent
	Mtoe	million tonnes of oil equivalent
	Lge	litre of gasoline equivalent
	bcme	billion cubic metres of natural gas equivalent
	Mtce	million tonnes of coal equivalent (equals 0.7 Mtoe)
	btce	billion tonnes of coal equivalent

Mass	kg	kilogramme
	t	tonne (1 tonne = 1 000 kg)
	kt	kilotonne (1 tonne x 10^3)
	Mt	million tonnes (1 tonne x 10^6)
	Gt	gigatonne (1 tonne x 10^9)
Monetary	USD million	1 US dollar x 10^6
	USD billion	1 US dollar x 10^9
	USD trillion	1 US dollar x 10^{12}
	USD/t CO_2	US dollars per tonne of carbon dioxide
Volumetric	bcm	billion cubic metres
	tcm	trillion cubic metres
	barrel	one barrel of crude oil
	kb/d	thousand barrels per day
	mb/d	million barrels per day

Definitions

Accelerated server: A specialised server equipped with hardware accelerators such as graphics processing units (GPUs) or tensor processing units (TPUs), to significantly boost computational performance for parallelisable and compute-intensive workloads. These servers are particularly critical for applications such as AI model training, inference, and high-performance computing.

Aeroderivative gas turbine: A class of gas turbine derived from aviation jet engine designs, characterised by a compact footprint, rapid start-up capability (typically under ten minutes to full load) and high part-load efficiency.

Agentic AI: A class of AI systems capable of autonomously decomposing high-level objectives into sequences of sub-tasks, selecting and invoking external tools, and iterating on intermediate results with limited or no human intervention.

Back-up generation capacity: Households and businesses connected to a main power grid may also have a source of back-up power generation capacity that, in the event of disruption, can provide electricity. Back-up generators are typically fuelled with diesel or gasoline. Capacity can be as little as a few hundred watts. Such capacity is distinct from mini-grid and off-grid systems that are not connected to a main power grid.

Battery storage: Energy storage technology that uses reversible chemical reactions to absorb, store and release electricity on demand.

Buildings: The buildings sector includes energy used in residential and services buildings. Services buildings include commercial and institutional buildings (e.g. schools, hospitals, public offices.) and other non-specified buildings. Building energy use includes space heating and cooling, water heating, lighting, appliances and cooking equipment.

Capacity market: A market-based mechanism designed to ensure resource adequacy by remunerating generators and other resources for being available to supply electricity or reduce demand when needed. In many capacity markets, capacity is procured ahead of the delivery period through auctions, with payments reflecting the value of maintaining sufficient reliable capacity.

Capacity utilisation rate: The ratio of the average load to the nameplate capacity of an asset over a specified period, expressed as a percentage. The capacity utilisation rate measures how intensively the installed capacity is used. Higher utilisation indicates more effective use of the available capacity, while lower utilisation suggests underuse of the installed assets.

Carbon dioxide (CO_2)**:** A gas consisting of one part carbon and two parts oxygen. It is an important greenhouse (heat-trapping) gas.

Cloud computing: Cloud computing is the provision of computing services via the internet ("the cloud"). It enables users to access scalable and flexible services on demand, without the need to manage physical infrastructure directly.

Coal: Consists of both primary coal, i.e. lignite, coking and steam coal, and derived fuels, e.g. patent fuel, brown-coal briquettes, coke-oven coke, gas coke, gas works gas, coke-oven gas, blast furnace gas and oxygen steel furnace gas. Peat is also included.

Colocation and service provider data centres: These facilities either lease space to customers to house their own computing and storage equipment (colocation) or provide both the space and computing equipment (service providers).

Concentrating solar power (CSP)**:** Thermal power generation technology that collects and concentrates sunlight to produce high temperature heat to generate electricity.

Connection queue: A register of projects, whether generation or demand, that have applied to a system operator for a new or expanded connection to the electricity grid and are awaiting technical assessment, approval or construction of the required infrastructure.

Conventional server: A conventional server relies solely on central processing units (CPUs) for processing, without the use accelerator chips. It handles general computing tasks using standard memory, storage, and networking components.

Critical minerals: A wide range of minerals and metals that are essential in clean energy technologies and other modern technologies and have supply chains that are vulnerable to disruption. Although the exact definition and criteria differ among countries, critical minerals for clean energy technologies typically include chromium, cobalt, copper, graphite, lithium, manganese, molybdenum, nickel, platinum group metals, zinc, rare earth elements and other commodities.

Data centre cluster: A geographic concentration of data centre facilities within a defined area that collectively represents a significant share of local or regional electricity demand.

Demand-side integration (DSI)**:** Consists of two types of measures: actions that influence load shape such as energy efficiency and electrification; and actions that manage load such as demand-side response measures.

Demand-side response (DSR)**:** Describes actions which can influence the load profile such as shifting the load curve in time without affecting total electricity demand, or load shedding such as interrupting demand for a short duration or adjusting the intensity of demand for a certain amount of time.

Dispatchable generation: Electricity from technologies whose power output can be readily controlled up to the nameplate capacity, i.e. increased to maximum rated capacity or decreased to zero, in order to help match supply with demand.

Domain-specific model: An AI model developed and trained for a narrowly defined task or operating environment, using data and constraints particular to that domain. Domain-specific models are typically much smaller and less computationally intensive than general-purpose models.

Edge computing: A distributed computing architecture in which data processing occurs on or near the device generating the data, rather than in a centralised data centre. Edge computing reduces network latency and bandwidth requirements, making it suitable for real-time applications such as autonomous driving, industrial robots and on-device AI inference.

Electric vehicles (EVs)**:** Electric vehicles comprise of battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs).

Electricity demand: Defined as total gross electricity generation less own use generation, plus net trade (imports less exports), less transmission and distribution losses.

Electricity generation: Defined as the total amount of electricity generated by power only or combined heat and power plants including generation required for own use. This is also referred to as gross generation.

End-use sectors: Include industry, transport, buildings, agriculture and other non-energy use.

Energy demand: See total energy supply.

Energy-intensive industries: Includes production and manufacturing in the branches of iron and steel, chemicals, non-metallic minerals (including cement), non-ferrous metals (including aluminium), and paper, pulp and printing.

Energy-related and industrial process CO_2 emissions: Carbon dioxide emissions from fuel combustion, industrial processes, and fugitive and flaring CO_2 from fossil fuel extraction. Unless otherwise stated, CO_2 emissions in the *World Energy Outlook* refer to energy-related and industrial process CO_2 emissions.

Energy sector greenhouse gas (GHG) **emissions:** Energy-related and industrial process CO_2 emissions plus fugitive and vented methane (CH_4) and nitrous dioxide (N_2O) emissions from the energy and industry sectors.

Energy services: A personal or societal gain from the use of energy. Include, *inter alia,* heating, cooling, lighting, entertainment, mobility, nourishment, hygiene and education. Also see useful energy.

Enterprise data centres: These facilities are run by businesses or institutions for their own use. They are typically smaller and less efficient than other types of data centres.

Floating-point operation (FLOP)**:** A floating-point operation is an arithmetic calculation involving floating-point numbers, such as addition, subtraction, multiplication, or division. It is commonly used as a unit for measuring computational workload. Floating-point operations per second (FLOPS) is a common metric for evaluating the performance of accelerated servers.

Fossil fuels: Consist of coal, oil and natural gas. Total fossil fuel use is equal to unabated fossil fuels plus fossil fuels with CCUS plus non-energy use of fossil fuels.

Frontier model: A large-scale AI model, typically based on transformer architectures, trained on vast datasets and representing the most advanced general-purpose capabilities available at a given time.

Gallium nitride (GaN)**:** A wide-bandgap semiconductor material used in high-frequency, high-efficiency power conversion devices. Gallium nitride transistors can switch faster and at higher voltages than conventional silicon devices.

Geothermal: Heat derived from the sub-surface of the earth, usually using a working fluid such as water and/or steam to bring the energy to the surface. Depending on its characteristics, geothermal energy can be used for heating and cooling purposes or be harnessed to generate clean electricity if the temperature is adequate.

Graphics processing unit (GPU)**:** Graphics processing units (GPUs) and other accelerators, such as tensor processing units (TPUs), are optimised for parallel computations, enabling faster processing of certain tasks. These types of processors are pivotal for AI model training, inference, and high-performance computing.

Grid connection: The physical link and contractual arrangement through which a facility is connected to the public electricity transmission or distribution network. Obtaining a grid connection typically requires a technical assessment by the relevant system operator, which evaluates whether the facility can be accommodated within existing network capacity or whether upgrades are needed.

Heat (end-use)**:** Can be obtained from the combustion of fossil or renewable fuels, direct geothermal or solar heat systems, exothermic chemical processes and electricity (through resistance heating or heat pumps which can extract it from ambient air and liquids). This category refers to the wide range of end-uses, including space and water heating, and cooking in buildings, desalination and process applications in industry. It does not include cooling applications.

Heat (supply)**:** Obtained from the combustion of fuels, nuclear reactors, large-scale heat pumps, geothermal or solar resources. It may be used for heating or cooling, or

converted into mechanical energy for transport or electricity generation. Commercial heat sold is reported under total final consumption with the fuel inputs allocated under power generation.

Heavy industries: Iron and steel, chemicals and cement.

High-bandwidth memory (HBM)**:** A class of memory technology that vertically stacks multiple memory dies and connects them via wide internal interfaces, achieving significantly higher data transfer rates than conventional memory modules.

Hourly matching: A procurement approach in which an electricity consumer seeks to match its demand with generation on an hour-by-hour basis, rather than on a net annual basis.

Hydropower: Refers to the electricity produced in hydropower projects, with the assumption of 100% efficiency. It excludes output from pumped storage and marine (tide and wave) plants.

Hyperscale data centres: These are massive facilities operated by major technology companies, such as Amazon Web Services, Google, Meta, and Microsoft. They use scalable, highly efficient infrastructure to support cloud services, web hosting and, increasingly, AI services.

Idle power: Idle power refers to the amount of electricity a device consumes to perform essential background operations when it is not actively processing workloads. The idle power ratio is the same metric, expressed as a percentage of the device's maximum rated power. Lower levels of idle power indicate higher operational efficiency.

Industry: The sector includes fuel used within the manufacturing and construction industries. Key industry branches include iron and steel, chemicals and petrochemicals, cement, aluminium, and paper, pulp and printing. Use by industries for the transformation of energy into another form or for the production of fuels is excluded and reported separately under other energy sector. There is an exception for fuel transformation in blast furnaces and coke ovens, which are reported within iron and steel. Consumption of fuels for the transport of goods is reported as part of the transport sector, while consumption of fuels by off-road vehicles is reported under the specific sector. For instance, fuels consumed by bulldozers as a part of industrial operations is reported in industry.

Inference: Inference is the process of deploying a trained model to analyse new or real-time data in order to generate outputs such as predictions, classifications, decisions, or responses. Unlike training, which involves learning from data, inference focuses on using learned patterns to perform tasks in production environments.

Installed IT capacity: In a data centre, installed IT capacity refers to the total rated capacity of servers, storage, and networking devices and is measured in megawatts (MW).

Investment: Investment is the capital expenditure in energy supply, infrastructure, end-use and efficiency. Fuel supply investment includes the production, transformation and transport of oil, gas, coal and low-emissions fuels. *Power sector* investment includes new construction and refurbishment of generation, electricity grids (transmission, distribution and public

electric vehicle chargers), and battery storage. *Energy efficiency* investment includes efficiency improvements in buildings, industry and transport. *Other end-use* investment includes the purchase of equipment for the direct use of renewables, electric vehicles, electrification in buildings, industry and international marine transport, equipment for the use of low-emissions fuels, and CCUS in industry and direct air capture. Data and projections reflect spending over the lifetime of projects and are presented in real terms in year-2024 US dollars converted at market exchange rates unless otherwise stated. Total investment reported for a year reflects the amount spent in that year.

Latency: Network latency is a measure of the time that data takes to be communicated across the network. Networks with a longer delay or lag have high latency, while those with fast response times have low latency.

Large language model (LLM)**:** A type of AI model based on the transformer architecture, trained on large corpora of text data and designed to generate, interpret and reason about natural language.

Levelised cost of electricity (LCOE)**:** An indicator of the expected average production cost for each unit of electricity generated by a technology over its economic lifetime. The LCOE combines into a single metric all the cost elements directly associated with a given power technology, including construction, financing, fuel, maintenance and costs associated with a carbon price. It does not include network integration or other indirect costs.

Light industries: Include non-energy-intensive industries: food and tobacco; machinery; mining and quarrying; transportation equipment; textiles; wood harvesting and processing and construction.

Liquid cooling: A thermal management approach in which a liquid medium, typically water or a dielectric fluid, is used to transfer heat away from IT equipment. Liquid cooling systems can absorb substantially more heat per unit volume than air-based systems and are increasingly required in data centres with high-density AI hardware.

Load factor: The ratio of average electricity demand to peak electricity demand over a specified period, expressed as a percentage. A higher load factor indicates greater sustained consumption over the period.

Low-emissions electricity: Includes output from renewable energy technologies, nuclear power, fossil fuels fitted with CCUS, hydrogen and ammonia.

Maximum designed capacity: In a data centre, this refers to the maximum theoretical capacity the facility can support when fully populated with IT equipment and operating at its design limits. This includes constraints such as power delivery, cooling infrastructure and rack space. In practice, the total installed capacity is often lower due to redundancy requirements, operational safety margins, or partial buildouts.

Natural gas: A gaseous fossil fuel, consisting mostly of methane. Occurs in deposits, whether liquefied or gaseous. In IEA analysis and statistics, it includes both non-associated gas originating from fields producing hydrocarbons only in gaseous form, and associated gas

produced in association with crude oil production, as well as methane recovered from coal mines (colliery gas). Natural gas liquids, manufactured gas (produced from municipal or industrial waste, or sewage) and quantities vented or flared are not included. Natural gas has a specific energy content of 44.09 MJ/kg on a higher heating value basis. Natural gas data in cubic metres are expressed on a gross calorific value basis and are measured at 15 °C and at 760 mm Hg (Standard Conditions). Natural gas data expressed in tonnes of oil equivalent, mainly to allow comparison with other fuels, are on a net calorific basis. The difference between the net and the gross calorific value is the latent heat of vapourisation of the water vapour produced during combustion of the fuel.

Neo-cloud: A category of newer, typically smaller cloud computing and data centre companies that have emerged or expanded rapidly in response to surging demand for AI compute capacity.

Non-energy-intensive industries: See other industry.

Nuclear power: Refers to the electricity produced by a nuclear reactor, assuming an average conversion efficiency of 33%.

Oil: A liquid fuel. Usually refers to fossil fuel mineral oil. Includes oil from both conventional and unconventional oil production. Petroleum products include refinery gas, ethane, liquid petroleum gas, aviation gasoline, motor gasoline, jet fuel, kerosene, gas/diesel oil, heavy fuel oil, naphtha, white spirits, lubricants, bitumen, paraffin, waxes and petroleum coke.

Onsite generation: Electricity generation capacity installed at or immediately adjacent to the point of consumption, operating independently of or in parallel with the main electricity grid. In the data centre context, onsite generation is increasingly considered as a means of securing power supply more rapidly than grid connection timelines would allow. Onsite facilities may serve as the primary source of electricity, as temporary bridge power pending a grid connection, or as backup.

Physical AI: A broad category of AI systems designed to perceive, reason about and act within physical environments, typically embodied in hardware such as robotic arms, mobile robots, drones or autonomous vehicles. Physical AI integrates advances in machine perception, motor control and real-world simulation, drawing on techniques developed in generative AI and reinforcement learning.

Power density: The amount of electrical power consumed per unit of physical space.

Power electronics: Semiconductor-based devices and circuits that convert, control and condition electrical power. Applications include rectification (AC to DC conversion), voltage regulation, frequency conversion and power factor correction.

Power generation: Refers to electricity generation and heat production from all sources of electricity, including electricity-only power plants, heat plants, and co-generation (i.e. combined heat and power) plants. Both main activity producer plants and small plants that produce fuel for their own use (auto-producers) are included.

Power usage effectiveness (PUE)**:** The power usage effectiveness is the ratio of total facility electricity consumption to the electricity consumption of the IT equipment (PUE = total consumption/IT consumption). It is commonly used as a key indicator of how efficiently a data centre uses energy. It focuses on the amount of energy used by computing equipment, rather than electricity consumption by other facility infrastructure (such as cooling and lighting). A low level of PUE indicates a high level of energy efficiency.

Rack: A standardised enclosure used to mount and organise servers, networking equipment and other hardware in a data centre. The rack is the basic physical unit around which data centre infrastructure is designed.

Rare earth elements (REEs)**:** A group of seventeen chemical elements in the periodic table, specifically the fifteen lanthanides plus scandium and yttrium. REEs are key components in some clean energy technologies, including wind turbines, electric vehicle motors and electrolysers.

Reasoning model: A category of AI model designed to engage in multi-step logical inference, exploring and evaluating multiple solution paths before producing a final output.

Rectifier: An electrical device that converts alternating current (AC) to direct current (DC).

Renewables: Include modern bioenergy, geothermal, hydropower, solar photovoltaics, concentrating solar power, wind, marine (tide and wave) energy, and renewable waste.

Residential: Energy used by households including space heating and cooling, water heating, lighting, appliances, electronic devices and cooking.

Services: A component of the buildings sector. It represents energy used in commercial facilities, e.g. offices, shops, hotels, restaurants and in institutional buildings, e.g. schools, hospitals, public offices. Energy use in services includes space heating and cooling, water heating, lighting, appliances, cooking and desalination.

Silicon carbide (SiC)**:** A wide-bandgap semiconductor material used in power electronics for high-voltage, high-temperature and high-efficiency power conversion applications. Silicon carbide devices offer lower switching losses than conventional silicon components, making them suitable for rectifiers and voltage converters in high-power-density environments.

Solar: Includes solar photovoltaics (PV), concentrating solar power (CSP), and solar heating and cooling.

Solar photovoltaics (PV)**:** Electricity produced from solar photovoltaic cells including utility-scale and small-scale installations.

Solid-state transformer: A power conversion device that uses semiconductor-based switching circuits to transform voltage levels. Solid-state transformers offer a substantially smaller material footprint and greater precision in power control than conventional transformers.

Token: A discrete unit of text processed by a language model, typically representing a word, sub-word fragment or punctuation mark, depending on the model's tokenisation scheme.

Total energy supply (TES)**:** Represents domestic demand only and is equivalent to electricity and heat generation plus the other energy sector, excluding electricity, heat and hydrogen, plus total final consumption, excluding electricity, heat and hydrogen. TES does not include ambient heat from heat pumps or electricity trade.

Total final consumption (TFC)**:** Is the sum of consumption by the various end-use sectors. TFC is broken down into energy demand in the following sectors: industry (including manufacturing, mining, chemicals production, blast furnaces and coke ovens); transport; buildings (including residential and services); and other (including agriculture and other non-energy use). It excludes international marine and aviation bunkers, except at world level where it is included in the transport sector.

Total installed capacity: In a data centre, total installed capacity refers to both IT capacity and the power capacity of auxiliary equipment. In practice, this is often lower than the maximum designed capacity due to redundancy requirements, operational safety margins, or partial buildouts.

Training: The computational process by which an AI model is developed, involving the iterative adjustment of model parameters in response to large volumes of input data so that the model learns to perform a specified task or set of tasks.

Transformer: A passive electrical device that transfers energy between circuits through electromagnetic induction, used to step voltage up or down for transmission, distribution and end-use applications.

Transport: Includes fuels and electricity used in the transport of goods or people within the national territory irrespective of the economic sector within which the activity occurs. This includes: fuel and electricity delivered to vehicles using public roads or for use in rail vehicles; fuel delivered to vessels for domestic navigation; fuel delivered to aircraft for domestic aviation; and energy consumed in the delivery of fuels through pipelines. Energy consumption from marine and aviation bunkers is presented only at the world level and is excluded from the transport sector at a domestic level.

Variable renewable energy (VRE): Sources of renewable energy (usually electricity) where the maximum output of an installation at a given time depends on the availability of fluctuating environmental inputs. VRE includes a broad array of technologies such as wind power, solar PV, run-of-river hydro, concentrating solar power (where no thermal storage is included) and marine (tidal and wave).

Uninterruptible power supply (UPS)**:** An uninterruptible power supply is equipment used to maintain power to a data centre during outages. UPS systems – most commonly batteries – are crucial to ensuring the extremely high levels of reliability that data centres must meet.

Utilisation rate: The utilisation rate of IT equipment measures the proportion of the available computing resources actively used over a given period.

Regional and country groupings

Advanced economies: Organisation for Economic Co-operation and Development (OECD) grouping and Bulgaria, Croatia, Cyprus[1,2], Malta and Romania.

Africa: North Africa and sub-Saharan Africa regional groupings.

Asia Pacific: Southeast Asia regional grouping and Australia, Bangladesh, Democratic People's Republic of Korea (North Korea), India, Japan, Korea, Mongolia, Nepal, New Zealand, Pakistan, The People's Republic of China (China), Sri Lanka, Chinese Taipei, and other Asia Pacific countries and territories.[3]

Caspian: Armenia, Azerbaijan, Georgia, Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan and Uzbekistan.

Central and South America: Argentina, Plurinational State of Bolivia (Bolivia), Bolivarian Republic of Venezuela (Venezuela), Brazil, Chile, Colombia, Costa Rica, Cuba, Curaçao, Dominican Republic, Ecuador, El Salvador, Guatemala, Guyana, Haiti, Honduras, Jamaica, Nicaragua, Panama, Paraguay, Peru, Suriname, Trinidad and Tobago, Uruguay and other Central and South American countries and territories.[4]

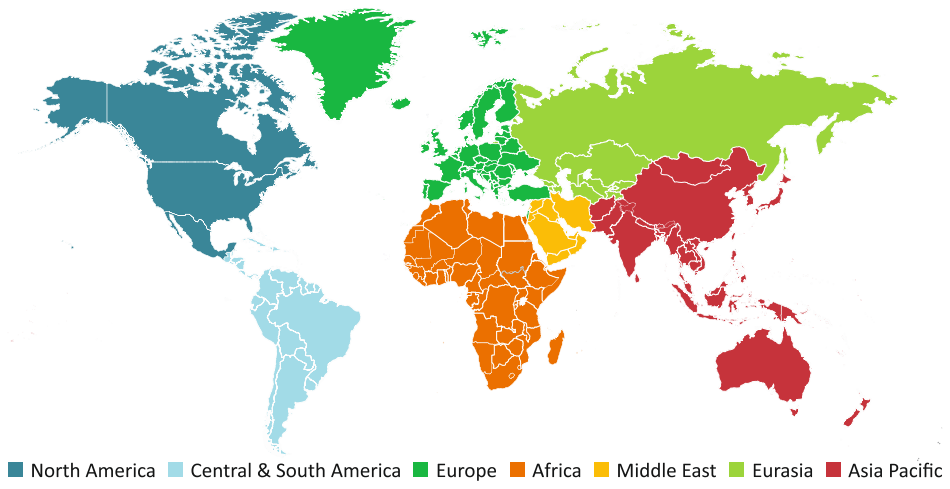
China: Includes (The People's Republic of) China and Hong Kong, China.

Developing Asia: Asia Pacific regional grouping excluding Australia, Japan, Korea and New Zealand.

Emerging market and developing economies: All other countries not included in the advanced economies regional grouping.

Eurasia: Caspian regional grouping and the Russian Federation (Russia).

Figure B.1 ▷ Main country groupings



Note: This map is without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Europe: European Union regional grouping and Albania, Belarus, Bosnia and Herzegovina, Gibraltar, Iceland, Israel[5], Kosovo[6], Montenegro, North Macedonia, Norway, Republic of Moldova, Serbia, Switzerland, Türkiye, Ukraine and United Kingdom.

European Union: Austria, Belgium, Bulgaria, Croatia, Cyprus[1,2], Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovak Republic, Slovenia, Spain and Sweden.

IEA (International Energy Agency)**:** Australia, Austria, Belgium, Canada, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Japan, Korea, Latvia, Lithuania, Luxembourg, Mexico, New Zealand, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, Switzerland, The Netherlands, Türkiye, United Kingdom and United States.

Latin America and the Caribbean (LAC)**:** Central and South America regional grouping and Mexico.

Middle East: Bahrain, Islamic Republic of Iran (Iran), Iraq, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Syrian Arab Republic (Syria), United Arab Emirates and Yemen.

Non-OECD: All other countries not included in the OECD regional grouping.

Non-OPEC: All other countries not included in the OPEC regional grouping.

North Africa: Algeria, Egypt, Libya, Morocco and Tunisia.

North America: Canada, Mexico and United States.

OECD (Organisation for Economic Co-operation and Development)**:** IEA grouping plus Chile, Colombia, Costa Rica, Iceland, Israel, Latvia and Slovenia.

OPEC (Organization of the Petroleum Exporting Countries)**:** Algeria, Angola, Bolivarian Republic of Venezuela (Venezuela), Equatorial Guinea, Gabon, Iraq, Islamic Republic of Iran (Iran), Kuwait, Libya, Nigeria, Republic of the Congo (Congo), Saudi Arabia and United Arab Emirates.

OPEC+: OPEC grouping plus Azerbaijan, Bahrain, Brunei Darussalam, Kazakhstan, Malaysia, Mexico, Oman, Russian Federation, South Sudan and Sudan.

Southeast Asia: Brunei Darussalam, Cambodia, Indonesia, Lao People's Democratic Republic (Lao PDR), Malaysia, Myanmar, Philippines, Singapore, Thailand and Viet Nam. These countries are all members of the Association of Southeast Asian Nations (ASEAN). Timor-Leste joined ASEAN on 26 October 2025 and is excluded from this *WEO* grouping for this publication, but is included in aggregate within the overarching Asia Pacific group.

Sub-Saharan Africa: Angola, Benin, Botswana, Cameroon, Côte d'Ivoire, Democratic Republic of the Congo (DRC), Equatorial Guinea, Eritrea, Ethiopia, Gabon, Ghana, Kenya, Kingdom of Eswatini, Madagascar, Mauritius, Mozambique, Namibia, Niger, Nigeria, Republic of the Congo (Congo), Rwanda, Senegal, South Africa, South Sudan, Sudan, United Republic of Tanzania (Tanzania), Togo, Uganda, Zambia, Zimbabwe and other African countries and territories.[7]

Country notes

[1] Note by Republic of Türkiye: The information in this document with reference to "Cyprus" relates to the southern part of the island. There is no single authority representing both Turkish and Greek Cypriot people on the island. Türkiye recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Türkiye shall preserve its position concerning the "Cyprus issue".

[2] Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Türkiye. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

[3] Individual data are not available and are estimated in aggregate for: Afghanistan, Bhutan, Cook Islands, Fiji, French Polynesia, Kiribati, Macau (China), Maldives, New Caledonia, Palau, Papua New Guinea, Samoa, Solomon Islands, Timor-Leste, Tonga and Vanuatu.

[4] Individual data are not available and are estimated in aggregate for: Anguilla, Antigua and Barbuda, Aruba, Bahamas, Barbados, Belize, Bermuda, Bonaire, Sint Eustatius and Saba, British Virgin Islands, Cayman Islands, Dominica, Falkland Islands (Malvinas), Grenada, Montserrat, Saint Kitts and Nevis, Saint Lucia, Saint Pierre and Miquelon, Saint Vincent and Grenadines, Saint Maarten (Dutch part), Turks and Caicos Islands.

[5] The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD and/or the IEA is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

[6] This designation is without prejudice to positions on status, and is in line with United Nations Security Council Resolution 1244/99 and the Advisory Opinion of the International Court of Justice on Kosovo's declaration of independence.

[7] Individual data are not available and are estimated in aggregate for: Burkina Faso, Burundi, Cabo Verde, Central African Republic, Chad, Comoros, Djibouti, Gambia, Guinea, Guinea-Bissau, Lesotho, Liberia, Malawi, Mali, Mauritania, Sao Tome and Principe, Seychelles, Sierra Leone and Somalia.

Abbreviations and acronyms

AC	alternating current
AESO	Alberta Electric System Operator
AI	artificial intelligence
ASIC	application-specific integrated circuit
ASR	automatic speech recognition
AV	autonomous vehicle
AWS	Amazon Web Services
BLS	Bureau of Labor Statistics
BNEF	Bloomberg New Energy Finance
BNetzA	Bundesnetzagentur
CAISO	California Independent System Operator
CAPEX	capital expenditure
CBOE	Chicago Board Options Exchange
CPU	central processing unit
CRU	Commission for Regulation of Utilities
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DC	direct current
DFT	density functional theory

DSCR	debt service coverage ratio
EDB	Singapore Economic Development Board
EIA	Energy Information Administration
EMDE	emerging markets and developing economies
ENTSO-E	European Network of Transmission System Operators for Electricity
ERCOT	Electric Reliability Council of Texas
FLAP-D	Frankfurt, London, Amsterdam, Paris and Dublin
FLOPS	floating-point operations per second
GDP	gross domestic product
GEM	Global Energy Monitor
GOES	grain-oriented electrical steel
GPU	graphics processing unit
HVAC	heating, ventilation and air conditioning
ICT	information and communications technology
IEA	International Energy Agency
IMDA	Infocomm Media Development Authority (Singapore)
IoT	internet of things
IP	Internet Protocol
IPF	international patent family
IRFM	Institute for Research on Fusion by Magnetic Confinement
IT	information technology
LCOP	levelised cost of production
LFP	lithium iron phosphate
Li-ion	lithium-ion
LLM	large language model
LM	language model
MER	market exchange rates
MoE	mixture of experts
MOU	memorandum of understanding
NCSL	National Conference of State Legislatures
NMC	lithium nickel cobalt manganese oxide
NPU	neural processing unit
O&M	operations and maintenance
OECD	Organisation for Economic Co-operation and Development
OPEX	operational expenditure
PJM	PJM Interconnection
PPA	power purchase agreement
PPP	purchasing power parity

References

Acemoglu, D. (2024), The simple macroeconomics of AI. Economic Policy, https://economics.mit.edu/sites/default/files/2024-04/The%20Simple%20Macroeconomics%20of%20AI.pdf

AESO (Alberta Electric System Operator) (2025), AESO Announces Interim Approach to Large Load Connections, https://www.aeso.ca/aeso/newsroom/aeso-announces-interim-approach-to-large-load-connections/

Aghion, P., & Bunel, S. (2024), AI and Growth: Where Do We Stand, https://www.frbsf.org/wp-content/uploads/AI-and-Growth-Aghion-Bunel.pdf

Altman (2025), The Gentle Singularity, https://blog.samaltman.com/the-gentle-singularity

Amazon (2025), 2024 Amazon Sustainability report AWS Summary, https://sustainability.aboutamazon.com/2024-amazon-sustainability-report-aws-summary.pdf

AP (Associated Press) (2026), Iran war halts Qatar helium output, threatening global tech supply chains, https://apnews.com/article/iran-chips-semiconductor-helium-exports-war-fe934332f7c83bb722ca87db22cd57d0

Baker McKenzie (2025), Malaysia: (i) Implementation of Data Centre Framework and (ii) Initiative for sustainable data centre development, https://insightplus.bakermckenzie.com/bm/real-estate_1/malaysia-i-implementation-of-data-centre-framework-and-ii-initiative-for-sustainable-data-centre-development

Becker, et al. (2025). Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity, https://arxiv.org/abs/2507.09089

Benchmark (2026), Q1 2026 Battery Energy Stationary Storage Forecast, https://www.benchmarkminerals.com/battery-energy-stationary-storage

Bick et al. (2025), The State of Generative AI Adoption in 2025, https://www.stlouisfed.org/on-the-economy/2025/nov/state-generative-ai-adoption-2025

Bloomberg (2026), DDR5 and DDR4 Spot prices, accessed through Bloomberg Terminal.

Bloomberg (2025), The $3 Trillion AI Data Center Build-Out Becomes All-Consuming For Debt Markets, https://www.bloomberg.com/news/articles/2026-02-02/the-3-trillion-ai-data-center-build-out-spurs-a-debt-market-boom

BLS (Bureau of Labor Statistics) (2026), Productivity, https://www.bls.gov/productivity/

BNEF (Bloomberg New Energy Finance) (2026), Power Purchase Agreements (PPA) database, https://www.bnef.com/interactive-datasets/2e68fe885a1313d9

BNEF (2025), Global Data Center Live IT Database, https://www.bnef.com/interactive-datasets/2e68fe885a1313d9

BNetzA (Bundesnetzagentur / Federal Network Agency) (2025), Rahmenfestlegung der Allgemeinen Netzentgeltsystematik Strom (AgNes) [Framework for the General Network Charges System for Electricity], https://www.bundesnetzagentur.de/DE/Beschlusskammern/1_GZ/GBK-GZ/2025/GBK-25-01-1x3_AgNes/Downloads/Diskussionspapier_AgNes.pdf?__blob=publicationFile&v=6

BostonDynamics (2026), Spot Specifications, https://support.bostondynamics.com/s/article/Spot-Specifications-49916

Brynjolfsson, E., Li, D., & Raymond, L. (2025), Generative AI at work, Quarterly Journal of Economics, https://academic.oup.com/qje/article/140/2/889/7990658

C3.ai (n.d.), Predictive Maintenance for Electric Grid, https://c3.ai/customers/predictive-maintenance-for-electric-grid/

CAISO (2026), Large Load Considerations - Issue Paper, https://www.caiso.com/documents/issue-paper-large-load-consideration-jan-20-2026.pdf

Chan, K. (2026), Does China care about AGI?, https://www.high-capacity.com/p/does-china-care-about-agi

China Daily (2026), Chinese AI chips gaining market traction, https://www.chinadaily.com.cn/a/202601/30/WS697cb910a310d6866eb36b0a.html

Counterpoint Research (2025), Global Smart Meter Installations Set to Surpass 3 Billion by 2030 Amid Accelerating Adoption, https://counterpointresearch.com/en/insights/global-smart-meter-installations-set-to-surpass-3-billion-by-2030-amid-accelerating-adoption

Credit Sights (2026), Utility Credit: Ten Themes for 2026, https://know.creditsights.com/utility-credit-ten-themes-for-2026-pdf

CRU (Commission for Regulation of Utilities) (2025), Large Energy Users Connection Policy, https://cruie-live-96ca64acab2247eca8a850a7e54b-5b34f62.divio-media.com/documents/CRU2025236_Large_Energy_User_connection_policy_decision_paper.pdf

Crunchbase (2026), (database) accessed January 2026, https://www.crunchbase.com/

Cui, K., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2024), The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4945566

Data Center Watch (2025), Q2 2025 Update, https://www.datacenterwatch.org/q22025

Data Centre Dynamics (2025), Chinese government plans data center capacity reseller network amid overbuild concerns, https://www.datacenterdynamics.com/en/news/chinese-government-plans-data-center-capacity-reseller-network-amid-overbuild-concerns/

Dell'Acqua, F., Lakhani, K. R., Lifshitz Assaf, H., Tushman, M. L., & Kellogg, K. C. (2023), Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality, Harvard Business School Working Paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4573321

EDB/IMDA (Singapore Economic Development Board/ Infocomm Media Development Authority) (2025) Launch of second data centre – Call for application, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2025/launch-of-second-data-centre

EIA (U.S. Energy Information Administration) (2025), Sales (consumption), revenue, prices & customers (EIA-861), https://www.eia.gov/electricity/data.php#sales

Elevate Energy Consulting (2025), Practical Guidance and Considerations for Large Load Interconnections, https://www.icc.illinois.gov/docket/P20250679/documents/371138/files/650731.pdf

Elsworth C., Huang K., Patterson D., Schneider I., Sedivy R., Goodman S., Townsend B., Ranganathan P., Dean J., Vahdat A., Gomes B., and Manyika J. (2025), Measuring the environmental impact of delivering AI at Google Scale, https://arxiv.org/pdf/2508.15734

Ember (2025), From AI to emissions: Aligning ASEAN's digital growth with energy transition goals, https://ember-energy.org/latest-insights/from-ai-to-emissions-aligning-asean-digital-growth-with-energy-transition/

Epoch AI (2026a), GPU Clusters, https://epoch.ai/data/gpu-clusters#use-this-work

Epoch AI (2026b), Data on Machine Learning Hardware, https://epoch.ai/data/machine-learning-hardware

EPRI (Electric Power Research Institute) (2026), Powering Intelligence 2026: Updated Scenarios of U.S. Data Center Electricity Use and Power Strategies, https://www.epri.com/research/products/000000003002034696

European Commission (2025a) AI Continent Action Plan, https://commission.europa.eu/topics/competitiveness/ai-continent_en

European Commission (2025b), European Grids Package, https://energy.ec.europa.eu/topics/infrastructure/european-grids_en

Eurostat (2025), Harmonised index of consumer prices (HICP), https://ec.europa.eu/eurostat/databrowser/view/prc_hicp_midx__custom_19656255/default/table

Federal Reserve Bank of St. Louis (2025), Consumer Price Indexes, https://fred.stlouisfed.org/categories/9

Felipe Oviedo (2025), Energy Use of AI Inference: Efficiency Pathways and Test-Time Compute, https://arxiv.org/pdf/2509.20241

FERC (Federal Energy Regulatory Commission) (2025), Federal Energy Regulatory Commission directs nation's largest grid operator to create new rules to embrace innovation and ensure reliability, https://www.ferc.gov/news-events/news/fact-sheet-ferc-directs-nations-largest-grid-operator-create-new-rules-embrace

Fervo Energy (2023), Fervo Energy Breaks Ground on the World's Largest Next-gen Geothermal Project, https://fervoenergy.com/fervo-energy-breaks-ground-on-the-worlds-largest-next-gen-geothermal-project/

Financial Times (2025), China gains dexterous upper hand in humanoid robot tussle with US, https://www.ft.com/content/4ebac441-d5a8-4c6a-950c-a160274d389b

Forbes (2021), In 2020 HDD Companies Shipped Over 1ZB of Storage Capacity, https://www.forbes.com/sites/tomcoughlin/2021/02/07/in-2020-hdd-companies-shipped-over-1zb-of-storage-capacity/

Gartner (2020), Gartner Says Worldwide Server Revenue Grew 5.1% in the Fourth Quarter of 2019, While Shipments Increased 11.7%, https://www.gartner.com/en/newsroom/press-releases/2020-03-19-gartner-says-worldwide-server-revenue-grew-5-percent-in-the-fourth-quarter-of-2019-while-shipments-increased-11-percent

Gartner (2018a), Gartner Says Worldwide Server Revenue Grew 25.7 Percent in the Fourth Quarter of 2017, While Shipments Increased 8.8 Percent, https://www.gartner.com/en/newsroom/press-releases/2018-03-08-gartner-says-worldwide-server-revenue-grew-in-the-fourth-quarter-of-2017

Gartner (2018b), Gartner Says Worldwide Server Revenue Grew 33.4 Percent in the First Quarter of 2018, While Shipments Increased 17.3 Percent, https://www.gartner.com/en/newsroom/press-releases/2018-06-11-gartner-says-worldwide-server-revenue-grew-33-percent-in-the-first-quarter-of-2018-while-shipments-increased-17-percent

Gartner (2017a), Gartner Says Worldwide Server Revenue Grew 16 Percent in the Third Quarter of 2017; Shipments Grew 5.1 Percent, https://www.gartner.com/en/newsroom/press-releases/2017-12-11-gartner-says-worldwide-server-revenue-grew-16-percent-in-third-quarter-of-2017

Gartner (2017b), Gartner Says Worldwide Server Shipments Declined 4.2 Percent in the First Quarter of 2017; Revenue Declined 4.5 Percent, https://www.gartner.com/en/newsroom/press-releases/2017-06-07-gartner-says-worldwide-server-shipments-declined-4-percent-in-the-first-quarter-of-2017-revenue-declined-4-percent

Gartner (2017c), Gartner Says Worldwide Server Shipments Grew 2.4 Percent in the Second Quarter of 2017; Revenue Grew 2.8 Percent, https://www.gartner.com/en/newsroom/press-releases/2017-09-12-gartner-says-worldwide-server-shipments-grew-2-percent-in-the-second-quarter-of-2017-revenue-grew-2-percent

Gartner (2016a), Gartner Says Worldwide Server Revenue Declined 0.8 Percent in the Second Quarter of 2016, While Shipments Increased 2 Percent, https://www.gartner.com/en/newsroom/press-releases/2016-09-14-gartner-says-worldwide-server-revenue-declined-1-percent-in-the-second-quarter-of-2016-while-shipments-increased-2-percent

Gartner (2016b), Gartner Says Worldwide Server Revenue Grew 8.2 Percent in the Fourth Quarter of 2015, While Shipments Increased 9.2 Percent, https://www.gartner.com/en/newsroom/press-releases/2016-03-09-gartner-says-worldwide-server-revenue-grew-8-percent-in-the-fourth-quarter-of-2015-while-shipments-increased-9-percent

Gartner (2015a), Gartner Says Worldwide Server Market Grew 4.8 Percent in Shipments, While Revenue Increased 2.2 Percent in Fourth Quarter of 2014, https://www.gartner.com/en/newsroom/press-releases/2015-03-03-gartner-says-worldwide-server-market-grew-4-percent-in-shipments-while-revenue-increased-2-percent-in-fourth-quarter-of-2014

Gartner, (2015b), Gartner Says Worldwide Server Revenue Grew 7.5 Percent in the Third Quarter of 2015, While Shipments Increased 9.2 Percent, https://www.gartner.com/en/newsroom/press-releases/2015-12-02-gartner-says-worldwide-server-revenue-grew-7-percent-in-the-third-quarter-of-2015-while-shipments-increased-9-percent

Gartner (2015c), Gartner Says Worldwide Server Shipment Market Grew 8 Percent in the Second Quarter of 2015, While Revenue Increased 7.2 Percent, https://www.gartner.com/en/newsroom/press-releases/2015-08-26-gartner-says-worldwide-server-shipment-market-grew-8-percent-in-the-second-quarter-of-2015-while-revenue-increased-7-percent

Gartner (2015d), Gartner Says Worldwide Server Shipments Grew 13 Percent in the First Quarter of 2015, While Revenue Increased 17.9 Percent, https://www.gartner.com/en/newsroom/press-releases/2015-05-28-gartner-says-worldwide-server-shipments-grew-13-percent-in-the-first-quarter-of-2015-while-revenue-increased-18-percent

Gartner (2014a), Gartner Says Worldwide Server Shipments Grew 1 Percent in the Third Quarter of 2014 While Revenue Increased 1.7 Percent, https://www.gartner.com/en/newsroom/press-releases/2014-12-03-gartner-says-worldwide-server-shipments-grew-1-percent-in-the-third-quarter-of-2014-while-revenue-increased-1-percent

Gartner (2014b), Gartner Says Worldwide Server Shipments Market Grew 1.3 Percent in the Second Quarter of 2014 While Revenue Increased 2.8 Percent, https://www.gartner.com/en/newsroom/press-releases/2014-08-27-gartner-says-worldwide-server-shipments-market-grew-1-percent-in-the-second-quarter-of-2014-while-revenue-increased-2-percent

Gartner, (2014c), Gartner Says Worldwide Server Shipments Market Grew 1.4 Percent in the First Quarter of 2014, While Revenue Declined 4.1 Percent, https://www.gartner.com/en/newsroom/press-releases/2014-05-28-gartner-says-worldwide-server-shipments-market-grew-1-percent-in-the-first-quarter-of-2014-while-revenue-declined-4-percent

Gas World (2026), Korea chipmakers have helium stockpiles 'for six months', https://www.gasworld.com/story/korea-chipmakers-have-helium-stockpiles-for-six-months/2174283.article/

GE Vernova (2025), 2025 Investor Update, https://www.gevernova.com/investors/events/2025-investor-update

GEM (Global Energy Monitor) (2026), Global Oil and Gas Plant Tracker, https://globalenergymonitor.org/projects/global-oil-gas-plant-tracker/

Georgia General Assembly (2026), SB 34; HB 1063, https://www.legis.ga.gov/legislation/69551 and https://www.legis.ga.gov/legislation/72514

Global Energy Monitor (2025), Global Oil and Gas Plant Tracker, https://globalenergymonitor.org/projects/global-oil-gas-plant-tracker/

Goldman Sachs (2026), The Macroeconomic Spillovers From AI Electricity Demand, https://www.gspublishing.com/content/research/en/reports/2026/02/11/cefdda9f-3b7c-4a53-9950-dfcfa3c6bbc2.html

Google (2025a), Google Environmental Report 2025, https://www.gstatic.com/gumdrop/sustainability/google-2025-environmental-report.pdf

Google (2025b), How much energy does Google's AI use? We did the math, https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference

Google (2025c), Google DataCenters: Efficiency, https://datacenters.google/efficiency/

HGBR (Hampton Global Business Review) (2025), Localizing The Global SiC Supply Chain: Industrial Policy, Capacity Investment, and Competitive Disruption, https://hgbr.org/research_articles/localizing-the-global-sic-supply-chain-industrial-policy-capacity-investment-and-competitive-disruption/

Hintemann, R., Hinterholtzer, S. and Konrat, F. (2024), Server Stock Data — A Basis for Determining the Energy and Resource Requirements of Data Centres, 2024 Electronics Goes Green 2024+ (EGG), pp. 1-5, https://ieeexplore.ieee.org/abstract/document/10631194

Hoffmann, M., Boysel, S., Nagle, F., Peng, S., & Xu, K. (2024), Generative AI and the Nature of Work. CESifo Working Paper/ Harvard Business School Working Paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5007084

Ho, A. et al. (2024), Algorithmic progress in language models, https://arxiv.org/abs/2403.05812

Hugging Face (2025), AI Energy Score Leaderboard, https://huggingface.co/spaces/AIEnergyScore/Leaderboard

Hydro-Québec (2026), A new rate for data centres and a rate adjustment for blockchains to reflect the value of renewable electricity, https://news.hydroquebec.com/news/press-releases/all-quebec/hydro-quebec-proposing-regie-energie-new-rate-large-data-centres-adjustment-rate-cryptographic-use-applied-blockchains.html

IDC (2025), Global Memory Shortage Crisis: Market Analysis and the Potential Impact on the Smartphone and PC Markets in 2026, https://www.idc.com/resource-center/blog/global-memory-shortage-crisis-market-analysis-and-the-potential-impact-on-the-smartphone-and-pc-markets-in-2026/

IDC (2024), Datacenter Trends: Sustainable Builds and Carbon Emissions, https://www.idc.com/getdoc.jsp?containerId=IDC_P33186

IEA (International Energy Agency) (2026), Sodium-ion battery momentum grows, but challenges remain, https://www.iea.org/commentaries/sodium-ion-battery-momentum-grows-but-challenges-remain

IEA (2026b), Real-World Impact of AI in Energy, http://d19ob9sqegt2wc.cloudfront.net/stage/uploads/Casebook_on_AI_A5_202447a94e.pdf

IEA (2025a), Energy and AI, https://www.iea.org/reports/energy-and-ai

IEA (2025b), Overcoming energy constraints is key to delivering on Europe's data centre goals, https://www.iea.org/commentaries/overcoming-energy-constraints-is-key-to-delivering-on-europe-s-data-centre-goals

IEA (2025c), Global Critical Minerals Outlook 2025, https://www.iea.org/reports/global-critical-minerals-outlook-2025

IEA (2025d) Energy Prices, https://www.iea.org/data-and-statistics/data-product/energy-prices#energy-prices-and-taxes

IEA (2025e), Real-Time Electricity Tracker, https://www.iea.org/data-and-statistics/data-tools/real-time-electricity-tracker?from=2026-3-3&to=2026-4-2&category=demand

IEA (2025f), “Demand Response 4.0” enabled by machine learning, https://www.iea.org/data-and-statistics/data-tools/energy-and-ai-observatory?tab=AI+for+energy&casestudy=%E2%80%9CDemand+Response+4.0%E2%80%9D%C2%A0enabled+by+machine+learning

IEA (2025g), AI-powered wind speed forecasting, https://www.iea.org/data-and-statistics/data-tools/energy-and-ai-observatory?tab=AI+for+energy&casestudy=AI-powered+wind+speed+forecasting

IEA (2025h), AI-powered materials discovery for battery innovation, https://www.iea.org/data-and-statistics/data-tools/energy-and-ai-observatory?tab=AI+for+energy&casestudy=AI-powered+materials+discovery+for++battery+innovation

IEA (2025i), World Energy Employment 2025, https://www.iea.org/reports/world-energy-employment-2025

IEA (2025j), What Next for the Global Car Industry, https://www.iea.org/reports/what-next-for-the-global-car-industry

IMDA (Infocomm Media Development Authority) (2025), Launch of second data centre – Call for application, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2025/launch-of-second-data-centre

International Federation of Robotics (2025), International Federation of Robotics, https://ifr.org/

IoT Analytics (2024), Smart electricity meter market 2024: Global adoption landscape, https://iot-analytics.com/smart-meter-adoption/

IRFM (Institute for Research into Fusion by Magnetic Confinement) (2025), AI to monitor the walls of the Tokamak WEST, https://irfm.cea.fr/en/2025/12/ai-to-monitor-the-walls-of-the-tokamak-west/

Koomey, J. (2011), Growth in data center electricity use 2005 to 2010. A report by Analytical Press, completed at the request of The New York Times, https://alejandrobarros.com/wp-content/uploads/old/Growth_in_Data_Center_Electricity_use_2005_to_2010.pdf

Koomey, J. G. (2007), Estimating total power consumption by servers in the U.S. and the world, http://uploadi.www.ris.org/editor/1203418697svrpwrusecompletefinal.pdf

Korean Statistical Information Service (2025), Consumer Price Survey, https://mods.go.kr/board.es?mid=a20109020000&bid=11751&eng_board_type=01

Korinek, A., & Suh, D. (2024), Scenarios for the Transition to AGI, https://www.nber.org/system/files/working_papers/w32255/w32255.pdf

Lei, N. and Masanet, E. (2022), Climate- and technology-specific PUE and WUE estimations for U.S. data centers using a hybrid statistical and thermodynamics-based approach, https://www.sciencedirect.com/science/article/pii/S0921344922001719

McCoy (2026), GT Turbine Order Data, https://www.mccoypower.net/

McKinsey (2023), The economic potential of generative AI: The next productivity frontier, https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

Malmodin, J. et al. (2024), ICT sector electricity consumption and greenhouse gas emissions – 2020 outcome, https://www.sciencedirect.com/science/article/pii/S0308596123002124

Masanet, E. et al. (2020), Recalibrating global data center energy-use estimates. Science, vol. 367(6481), pp. 984-986, https://www.science.org/doi/abs/10.1126/science.aba3758

Meta (2025), 2025 Sustainability Report - Environmental Data Index, https://sustainability.atmeta.com/wp-content/uploads/2025/10/Meta_2025-Environmental-Data-Index.pdf

Micron (2024), Financial results F Q3 2024, https://investors.micron.com/static-files/a531c7f0-fca2-48f3-8f24-79c945aaa2d2

Microsoft (2026), Building Community-First AI Infrastructure, https://blogs.microsoft.com/on-the-issues/2026/01/13/community-first-ai-infrastructure/

Microsoft (2025), 2025 Enviornmental Sustainability Report - Accelerating progress to 2030, https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report/

Monitoring Analytics (2025), Analysis of the 2026/2027 RPM Base Residual Auction, https://www.monitoringanalytics.com/reports/Reports/2025/IMM_Analysis_of_the_20262027_RPM_Base_Residual_Auction_Part_A_20251001.pdf

NCSL (National Conference of State Legislatures) (2025), Policy Snapshot: Data Center Incentives, https://www.ncsl.org/fiscal/policy-snapshot-data-center-incentives

Noy, S., & Zhang, W. (2023), Experimental evidence on the productivity effects of generative Artificial Intelligence, Science, https://www.science.org/doi/10.1126/science.adh2586

Nvidia (2025), 800 VDC Architecture for Next-Generation AI Infrastructure, https://nvdam.nvidia.com/assets/share/asset/zlg5snufeo

OECD (Organization for Economic Cooperation and Development) (2026), AI meets trade: Global linkages and the cross-country distribution of the gains from AI, https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/03/ai-meets-trade_6001acf4/13081644-en.pdf

OECD (2025), Macroeconomic productivity gains from Artificial Intelligence in G7 economies, https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/macroeconomic-productivity-gains-from-artificial-intelligence-in-g7-economies_dcf91c3e/a5319ab5-en.pdf

OMDIA (2026), Long Range Server Forecast – 1H26, https://omdia.tech.informa.com/om143626/long-range-server-forecast--1h26

OMDIA (2025), Data Center Building Tracker – 2H25, https://omdia.tech.informa.com/om138977/data-center-building-tracker--2h25

Patterson D., Gonzalez J., Hölzle U., Le Q., Liang C., Munguia L-M., Rothchild D., So D., Texier M., and Dean J. (2022), The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink, https://arxiv.org/abs/2204.05149

PATSTAT (2026), (database) accessed January 2026, https://www.epo.org/en/searching-for-patents/business/patstat

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023), The impact of AI on developer productivity: Evidence from GitHub Copilot, https://arxiv.org/abs/2302.06590

PJM (2026), Load Forecast Report, https://www.pjm.com/-/media/DotCom/library/reports-notices/load-forecast/2026-load-report.ashx

PJM (2025a), Long-Term Load Forecast Report, https://www.pjm.com/-/media/DotCom/library/reports-notices/load-forecast/2025-load-report.ashx

PJM (2025b), 2027/2028 Base Residual Auction Report, https://www.pjm.com/-/media/DotCom/markets-ops/rpm/rpm-auction-info/2027-2028/2027-2028-bra-report.pdf

PJM (2024), Load Forecast Report January 2024, https://www.pjm.com/-/media/DotCom/library/reports-notices/load-forecast/2024-load-report.ashx

PJM (2023), Load Forecast Report January 2023, https://www.pjm.com/-/media/DotCom/library/reports-notices/load-forecast/2023-load-report.ashx

PJM (2022), Load Forecast Report January 2022, https://www.pjm.com/-/media/DotCom/library/reports-notices/load-forecast/2022-load-report.ashx

PJM (2021), Load Forecast Report January 2021, https://www.pjm.com/-/media/DotCom/library/reports-notices/load-forecast/2021-load-report.ashx

PJM (2020), Load Forecast Report January 2020, https://www.pjm.com/-/media/DotCom/library/reports-notices/load-forecast/2020-load-report.pdf

Praas, R., Sánchez, R., & Balland, P. (2025), From the bottom to the top: Robotics datasets lead on Hugging Face, https://aiworld.eu/story/from-the-bottom-to-the-top-robotics-datasets-lead-on-hugging-face

Qian, Z. (2026), China's AI Landscape: a free-for-all, not a central plan: what 6000+ filings with regulators reveal, https://www.chinatalk.media/p/chinas-ai-landscape-a-free-for-all

RSR Wireless (2024), Smart meters and massive IoT – what the next 10 years will look like, https://www.rcrwireless.com/20240816/fundamentals/smart-meters-and-massive-iot-what-the-next-10-years-will-look-like

Refinitiv (2025), Edited Transcript - 2330.TW - Q1 2025 Taiwan Semiconductor Manufacturing Co Ltd Earnings Call, https://investor.tsmc.com/english/encrypt/files/encrypt_file/reports/2025-04/7630274eecc1197a4e3ea6a415f44a47204fe10a/TSMC%201Q25%20Transcript.pdf

Reuters (2025), The AI frenzy is driving a memory chip supply crisis, https://www.reuters.com/world/china/ai-frenzy-is-driving-new-global-supply-chain-crisis-2025-12-03/

S&P Global (2026), Capital IQ (financial data platform) (database), accessed January 2026, https://www.spglobal.com/marketintelligence/en/solutions/sp-capital-iq-pro

S&P Global (2026), Corporate Renewables Contracts (database) accessed January 2026, https://core.spglobal.com/#platts/powerbi?dashboardName=CorporateRnwContracts&tab=Corporate%20renewables%20contracts%20(PPAs)

SemiAnalysis (2025a), NVIDIA GTC 2025 - Built For Reasoning, Vera Rubin, Kyber, CPO, Dynamo Inference, Jensen Math, Feynman, https://newsletter.semianalysis.com/p/nvidia-gtc-2025-built-for-reasoning-vera-rubin-kyber-cpo-dynamo-inference-jensen-math-feynman

SemiAnalysis (2025b), Huawei AI CloudMatrix 384 – China's Answer to Nvidia GB200 NVL72, https://newsletter.semianalysis.com/p/huawei-ai-cloudmatrix-384-chinas-answer-to-nvidia-gb200-nvl72

SemiAnalysis (2025c), Custom Market Intelligence, https://semianalysis.com/datacenter-industry-model/

Shehabi, A., et al. (2024), 2024 United States Data Center Energy Usage Report, https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf

Shehabi, A., et al. (2018), Data center growth in the United States: decoupling the demand for services from electricity use, https://iopscience.iop.org/article/10.1088/1748-9326/aaec9c

Shehabi, A., et al. (2016), United States Data Center Energy Usage Report, https://eta.lbl.gov/publications/united-states-data-center-energy

Siemens Energy (2026), Siemens Energy is investing $1 billion and creating highly skilled jobs in the United States, https://www.siemens-energy.com/global/en/home/press-releases/siemens-energy-is-investing--1-billion-and-creating-highly-skill.html

Siemens Energy (2025a), Capital Markets Day 2025 Gas Services, https://assets.siemens-energy.com/dam/5fc093ca-4fd8-482c-8390-b39b004f66f3/251120_CMD-2025_Gas_Services_FINAL-pdf_Original%20file.pdf

Siemens Energy (2025b), Capital Markets Day 2025 Grid Technologies, https://assets.siemens-energy.com/dam/d8d5d849-a7a9-4b19-a969-b39b004f6c26/251120_CMD-2025_Grid_Technologies_FINAL-pdf_Original%20file.pdf

SPEC (2024), SPECpower_ssj 2008, https://www.spec.org/power_ssj2008/results/

Stanford (2025), Artificial Intelligence Index Report 2025, https://hai.stanford.edu/ai-index/2025-ai-index-report

Statista (2024), Data Center – Worldwide, https://www.statista.com/outlook/tmo/data-center/worldwide

Statistics of Japan (2025), Base Consumer Price Index, https://www.e-stat.go.jp/en/stat-search/files?page=1&query=electricity&layout=dataset&toukei=00200573&tstat=000001150147&cycle=0&tclass1=000001150151&tclass2=000001150152&tclass3=000001150153&tclass4=000001150156&tclass5val=0&metadata=1&data=1

TechCrunch (2025), ChatGPT users send 2.5 billion prompts a day, https://techcrunch.com/2025/07/21/chatgpt-users-send-2-5-billion-prompts-a-day/

Turner & Townsend (2024), Data centre cost index 2024, https://reports.turnerandtownsend.com/dcci-2024/

UK Department for Energy Security and Net Zero (2025), Domestic energy price indices, https://www.gov.uk/government/statistical-data-sets/monthly-domestic-energy-price-statistics

UK Power Networks (2026), Data Centre Demand Profiles, https://ukpowernetworks.opendatasoft.com/explore/dataset/ukpn-data-centre-demand-profiles/information/

Unitree (2026), A1 Battery, https://www.unitree.com/a1/battery

U.S. Census Bureau (2026), Business Trends and Outlook Survey, https://www.census.gov/hfp/btos/data

U.S. DOE (United States Department of Energy) (2024), AI for Energy, https://www.energy.gov/cet/articles/ai-energy

USGS (US Geological Survey) (2026), Helium Statistics and Information, https://www.usgs.gov/centers/national-minerals-information-center/helium-statistics-and-information

Utility Dive (2026), ERCOT's large load queue jumped almost 300% last year. Retrieved from https://www.utilitydive.com/news/ercots-large-load-queue-jumped-almost-300-last-year-official/808820/

Virginia State Corporation Commission (2025), Final Order in Dominion Energy Virginia Biennial Review (PUR 2025 00058), https://www.scc.virginia.gov/docketsearch/DOCS/89g601!.PDF

White House (2026), Ratepayer Protection Pledge, https://www.whitehouse.gov/articles/2026/03/ratepayer-protection-pledge/

Wiser et al. (2025), Factors influencing recent trends in retail electricity prices in the United States, The Electricity Journal, vol. 34, no. 4, https://www.sciencedirect.com/science/article/pii/S1040619025000612#sec0015

Wood Mackenzie (2025), AI, power and the new geopolitics of energy, https://www.woodmac.com/news/opinion/ai-power-and-the-new-geopolitics-of-energy2/

World Bank (2016), World Development Report 2016: Digital Dividends, https://www.worldbank.org/en/publication/wdr2016

IEA Publications
International Energy Agency
Website: www.iea.org
Contact information: www.iea.org/contact

Typeset in France by IEA - April 2026
Cover design: IEA
Photo credits: © Shutterstock

Key Questions on Energy and AI

World Energy Outlook Special Report

The development and uptake of artificial intelligence (AI) has accelerated in recent years – elevating the question of what widespread deployment of the technology will mean for the energy sector. There is no AI without energy – specifically electricity for data centres. At the same time, AI could transform how the energy industry operates if it is adopted at scale. However, until now, policy makers and other stakeholders have often lacked the tools to analyse both sides of this issue due to a lack of comprehensive data.

This report from the International Energy Agency (IEA) aims to fill this gap based on new global and regional modelling and datasets, as well as extensive consultation with governments and regulators, the tech sector, the energy industry and international experts. It includes projections for how much electricity AI could consume over the next decade, as well as which energy sources are set to help meet it. It also analyses what the uptake of AI could mean for energy security, emissions, innovation and affordability.