axonpartnersgroup.com

# The Transformative Impact of AI on Data Centres

January 2025


**Authors**

Víctor Martínez
SENIOR MANAGER

Borja González
SENIOR ASSOCIATE


AXON CONSULTING

# Contents


| | | |
|---|---|---|
| **1.** | **Introduction** | **3** |
| **2.** | **How does AI affect computing power?** | **4** |
| 2.1. | Why – and how – AI is changing | 5 |
| 2.2. | Computing power for training | 8 |
| 2.3. | Computing power for inference | 10 |
| **3.** | **How is AI transforming data centres?** | **12** |
| 3.1. | Storage | 12 |
| 3.2. | AI processing hardware | 14 |
| 3.3. | Energy | 16 |
| **4.** | **How should AI data centres be enabled?** | **18** |
| 4.1. | Strategic location | 20 |
| 4.2. | Energy and resource supply | 21 |
| 4.3. | Regulatory frameworks and incentives | 22 |
| 4.4. | Public-private collaboration and investment | 23 |
| **5.** | **Conclusion** | **24** |

# 1. Introduction

Artificial Intelligence (AI) is redefining the role of data centres (DCs). Groundbreaking advances in AI models are setting new benchmarks in performance and multimodal capabilities. As a result, computational demand has skyrocketed, putting pressure on storage and processing in data centres.

Companies are now competing to develop better, faster, and more powerful AI models. The emergence in early 2025 of Chinese company DeepSeek, with a model claimed to be at least as efficient and powerful as established rivals – not to mention cheaper – is one of the most newsworthy efforts, but it won't be the last. The continuous evolution and release of new AI models will add to the already enormous pressure on computational infrastructure.

**The continuous evolution and release of new AI models will add to the already enormous pressure on computational infrastructure.**

This increased workload is directly linked to two fundamental processes: the training and inference of AI models.

Training advanced models requires managing large volumes of data and performing complex calculations, often involving weeks or even months of intensive operations using thousands of graphics processing units (GPUs) or tensor processing units (TPUs).

Inference, the use of trained models to generate real-time results or predictions, has grown in importance due to the massive adoption of AI, particularly in personalised and high-frequency applications. Indeed, computing requirements for inference now exceed those for training models.

Data centres have had to adapt – and they are adapting. They are currently embracing specialised hardware and next-generation storage solutions, redefining their operational priorities to manage this new workload and ensure efficiency. The scalability – capacity to change in size or scale – of data centres has also become a key focus.

Governments can aid this process. Strategic policies, fiscal incentives, clear regulations, and public-private collaborations will attract investments and promote innovation. These measures can not only facilitate economic growth but also create a favorable environment for research and technological development.

How is AI affecting computing power? What does this mean for data centres? How can countries ensure they have access to the computing supply that will power global AI development? Where should AI data centres be deployed? These are the questions this paper aims to answer.

# 2. How does AI affect computing power?

AI has transformed innovation, efficiency, and decision-making across multiple industries. 2024 saw such breakthroughs as the introduction of multimodal capabilities, expanded context windows, and more complex reasoning from AI models like GPT-4 and LLaMA 3.3.

These developments have significantly increased computational demand and, in turn, led to the evolution of data centres that can handle the training of increasingly sophisticated models. But computing power faces another challenge. The widespread adoption of these technologies – millions of users now interact daily with AI tools – has highlighted the computing required to run models while in use, a process known as inference.

This constantly evolving technological landscape needs sustainable and scalable infrastructures capable of meeting its growing demands.

## 2.1.
## Why – and how – AI is changing

In recent years, AI has made extraordinary advances, from language models capable of generating text to computer vision systems that analyse and process images. In turn computational demand in data centres has increased.

AI models can be classified into several main categories, each with specific applications and a direct impact on computational workload. They include:

- **Large language models (LLMs)**
  Designed to process and generate text in natural language, such as ChatGPT, these models are trained on vast amounts of data and are used for tasks like writing, translation, and text comprehension.

- **Computer vision models**
  These models analyse images and videos for tasks like classification, object detection, and segmentation. Tesla Autopilot, for example, uses object detection and segmentation to enable autonomous driving.

- **Recommendation models**
  These models suggest products, content, or services based on user preferences. For instance, Netflix recommends movies and TV shows tailored to users' viewing history and preferences.

- **Simulation models**
  These models mimic physical, chemical, or biological behaviours to make complex predictions. An example is pharmaceutical research, where AI models simulate molecular interactions to aid in drug development.

- **Other models**
  Other models include data processing, voice and audio analysis, and time series forecasting, tackling tasks like pattern analysis, sound signal generation, and event prediction. Examples include voice assistants like Siri and Alexa, among others.

Given that large language models are the most representative model type and that their evolution follows a general trend shared with other types of models, this document will primarily focus on the analysis of LLMs.

As we have noted, 2024 was a significant year for the development of AI systems and particularly next-generation language models, with major launches from leading AI research groups such as OpenAI, Meta, Google, and Anthropic. These organisations have driven transformative innovations, including multimodal processing (text, images and audio), enhanced reasoning capabilities for complex tasks, and expanded context windows that enable the handling of larger datasets and extended conversations. Their most noteworthy achievements include:

- **OpenAI**
  Since the introduction of GPT-1 in 2018, OpenAI has made significant strides with successive models like GPT-2, GPT-3 and GPT-3.5. GPT-4o, released in December 2024, is a model integrating advanced reasoning, multimodal processing, and optimisation for complex tasks such as scientific analysis, programming, and understanding deep contexts.

- **Meta**
  With its LLaMA 3 model family, launched in April 2024, Meta transformed the open-source AI domain. This model introduced multimodal capabilities for processing text, images, and videos, along with notable improvements in contextual reasoning and content generation. LLaMA 3.1, 3.2, and 3.3 have followed.

- **Google**
  The introduction of Gemini 2.0 in December 2024 revolutionised the market. This model combines speed, efficiency, and advanced multimodal capabilities, enabling precise analysis and generation of text, images, and audio. Variants such as Gemini 2.0 Pro and Gemini Ultra offer optimised solutions for critical and complex tasks.

- **Anthropic**
  Claude 3.5, released in June 2024, is a model that excels in programming and technical tasks. It includes a "computer use" functionality, allowing the model to interact with operating systems and execute commands, significantly expanding its practical applications. Variants such as Claude 3.5 Sonnet and Claude 3.5 Haiku have enhanced this innovation.

The progress of these models can be measured through various metrics that not only illustrate remarkable leaps in performance and capacity but also highlight the engineering innovations driving these transformations. From the exponential growth in the number of parameters and training tokens to significant extensions in context length and vocabulary size, these metrics demonstrate how leading companies in the sector are pushing the boundaries of what AI can achieve. For example:

AXON CONSULTING

- **Parameters**
  Parameters represent a model's adjustable internal values, essential for processing input data and learning complex patterns. The number of parameters has skyrocketed since the earliest releases, demonstrating the increase in the models' capacity. For example, the evolution from GPT-2 to GPT-4o saw a leap from 1.5 billion to 200 billion parameters. Gemini 2.5 grew to 20 billion parameters compared to about 1.6 billion for Gemini 1.0.

- **Training tokens**
  These are the smallest units of text used to train language models. The number of tokens processed during training is a key indicator of the size and diversity of the datasets employed. Token usage during training has again seen vast growth[2]. For example, LLaMA 1 used 1.4 trillion training tokens, compared to 15 trillion tokens used by LLaMA 3.3, while Claude 1.0 processed 120 billion tokens and Claude 3.5 scaled up to 1.5 trillion tokens.

- **Context length**
  This term defines the maximum number of tokens a model can simultaneously process as input to generate a response or predict the next token. A longer context window allows the model to capture more complex relationships within the text, including those spanning multiple sentences or paragraphs. Context-related parameters have also increased[3], as seen in models like GPT (from 1,024 tokens in GPT-2 to 128,000 tokens in GPT-4) or LLaMA (from 2,048 tokens in LLaMA 1 to 128,000 tokens in LLaMA 3.3).

- **Vocabulary size**
  This is the total number of unique tokens a model can recognise or use as part of its vocabulary during training and usage. Growth is smaller here, but still noticeable[4], with expansion from 40,000 tokens in GPT-1 or LLaMA 1 to 50,000 in GPT-4o or LLaMA 3.3.

While these changes are essential to expand the functionality and versatility of AI models, data centres are having to adapt to the increase computing power they require in two specific areas: training and inference, as the next two sections explain.

**From the exponential growth in the number of parameters and training tokens to significant extensions in context length and vocabulary size, these metrics demonstrate how leading companies in the sector are pushing the boundaries of what AI can achieve.**

2 https://lifearchitect.ai/models-table/
3 https://mattrickard.com/the-context-length-observation
4 https://blog.getbind.co/2024/12/13/llama-3-3-70b-vs-gpt-4o-which-is-better-for-coding

## 2.2.
## Computing power for training

The training process enables AI models to perform specific tasks with efficiency and precision. Models acquire knowledge from datasets, adjusting their internal parameters through advanced mathematical algorithms designed to minimise prediction errors. This learning mechanism not only allows them to identify patterns but also equips them to generalise that knowledge and apply it to diverse contexts and scenarios.

**Models acquire knowledge from datasets, adjusting their internal parameters through advanced mathematical algorithms designed to minimise prediction errors.**

But there are two problems: the exponential growth in data collection and the significant rise in computational demand for training. Yes, data is essential for optimising model performance. However, it also leads to a substantial increase in time and operational costs. Increasingly complex training processes mean that models have evolved from handling millions of parameters to dealing with trillions. The expansion of context windows and the introduction of multimodal capabilities add to these challenges

Figure 1 illustrates how improvements in model performance, driven by greater data availability, are directly linked to an exponential increase in computational demand, typically calculated in Floating Point Operations Per Second (FLOPs), a way of measuring a computer's processing speed.


Figure 1: Comparison of the precision of AI models against the computational demands their training requires [Source: Axon Partners Group][5]

5 https://heim.xyz/documents/Training-Compute-Thresholds.pdf and https://artificialanalysis.ai/leaderboards/models

Model accuracy, typically assessed via tests such as ChatBot Arena, Massive Multitask Language Understanding (MMLU), General Purpose Question Answering (GPQA), Math-500 or Human Evaluation, has generally doubled in less than five years – but in turn computing demand has vastly increased. Indeed, the upper right corner of Figure  hints at the existence of a practical limit where each marginal improvement in accuracy requires a disproportionately greater computational effort to train.

Computational costs are projected to grow exponentially as they approach these limits. These diminishing returns pose a challenge for the future development of these models. According to a study published by Epoch AI[6], a team of scientists investigating the future of AI, in June 2024, technology companies could face a shortage of publicly available training data between 2026 and 2032.

This deficit threatens to hinder the development and scalability of language models. More data can mean more accuracy – but it also means greater demands on computational capacity. Data centre infrastructures and training strategies will have to adjust to deal with this issue. And it's not the only one they'll have to cope with. The inference process, which we examine next, offers further challenges.

6 https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

## 2.3.
## Computing power for inference

Language models have become more sophisticated and accurate – and accessible. AI has evolved from being an experimental tool to becoming an indispensable part of everyday life. According to technology publication The Verge, ChatGPT had 200 million users worldwide per week by August 2024[7].

In 2020 only 116 million people worldwide regularly used AI-based applications. By 2024, this number had passed 314 million – an increase of over 170% in just four years[8]. The time spent using these tools has increased too, rising from an average of 10 minutes per user to one hour and 10 minutes daily.

AI model training has long been seen as the main computational challenge AI poses due to its complexity and resource intensity. However, with the mass adoption of AI-based tools, inference – the process of applying the knowledge acquired by a trained model to make predictions, generate responses, or solve problems – has overtaken training in terms of load and computational demands.

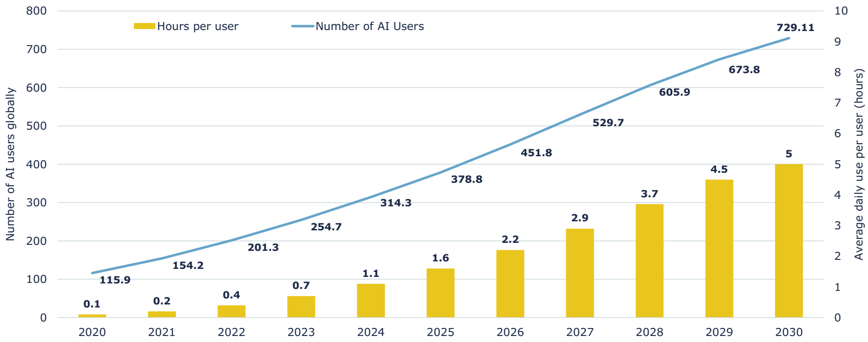



Figure 2: Evolution in AI users and daily consumption of AI applications [Source: Axon Partners Group based on Statista][9]

7 https://www.theverge.com/2024/8/29/24231685/openai-chatgpt-200-million-weekly-users

8 A regular user is someone who uses the platform habitually, has knowledge of the platform, engages in frequent and extended interactions, optimizes their queries, and returns regularly for tasks such as work, learning, or problem-solving.

9 https://www.statista.com/forecasts/1449844/ai-tool-users-worldwide

This shift is driven by three key factors: the increase in the number of users, the growing complexity of models, and the diversification of their applications. As we have seen, advanced models like GPT-4 or LLaMA 3.3 incorporate billions of additional parameters compared to earlier versions, significantly increasing the computational requirements per query. The transition to multimodal models, which process text, images, audio, and video in an integrated manner, adds to processing demands. Nor should we forget that these applications, particularly those requiring personalised or contextual responses, often demand memory and computational capacity in real time.

Austrian artificial intelligence company Ikangai[10] suggests that a single query to a model like GPT-4 can require between 3 and 60 teraflops. If 10 million queries are processed daily, the computational demand for inference could match the initial training demand in less than a month. Over an entire year, the computational load of inference could exceed the resources needed to train the original model by up to 100 times.

Figure 3 compares inference costs for leading companies in the sector with the costs associated with training[11].

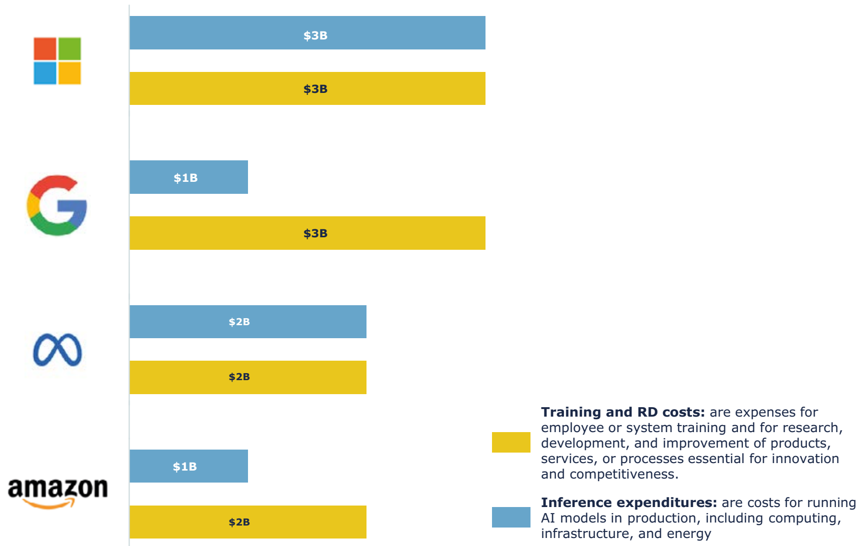



Figure 3: Costs incurred by companies in AI data centres for model inference and training from January to August 2024 [Source: Axon based on JP Morgan[12]]

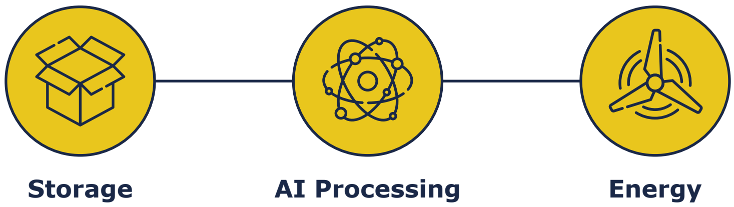
10 https://www.ikangai.com/the-secrets-of-gpt-4-leaked

11 Leaving aside investment in hardware (as shown in Figure 5) that is shared for both training and inference.

12 https://am.jpmorgan.com/content/dam/jpm-am-aem/global/en/insights/eye-on-the-market/a-severe-case-of-covidia-amv.pdf

# 3.
# How is AI transforming data centres?

AI models are becoming more complex and computationally demanding. In this section we examine three key features of data centres that are essential to supporting the development and operation of AI: storage, AI processing hardware, and energy.




## 3.1.
## Storage

Demand for data is already high. AI will make demand increase even further. This could lead to capacity shortages in data centres. Strategic planning is needed to meet future infrastructure needs.

Future demand will depend on factors such as the adoption rate of advanced AI use cases, the balance between cloud computing and edge computing for managing AI workloads, and the computing, storage, and networking requirements that these workloads entail.

As illustrated in Figure 4, AI models process vast – and growing – amounts of data, putting pressure on storage capacity.

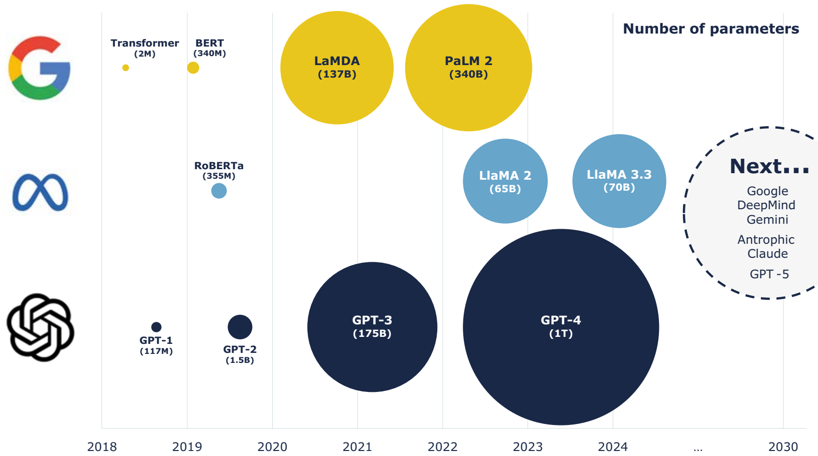



Figure 4: Different language model sizes of AI models over time [Source: Axon Partners Group based on Scalian][13]

For example, GPT-4 requires almost 10,000 times the storage capacity of GPT-1, while PaLM 2 requires 170,000 times the storage capacity of the original Transformer model. And growth isn't going to ease up. According to a recent McKinsey report[14], the need for data centre storage is expected to drive an annual increase of between 19% and 27% over the 2024-2030 period. According to the same report, by 2030 approximately 70% of the total demand for data centre capacity is projected to be for facilities equipped to handle advanced AI workloads.

**By 2030 approximately 70% of the total demand for data centre capacity is projected to be for facilities equipped to handle advanced AI workloads.**

13 https://www.scalian-spain.es/la-revolucion-de-la-inteligencia-artificial-el-poder-transformador-de-los-modelos-de-lenguaje-a-gran-escala-llm/

14 https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-centre-capacity-to-meet-growing-demand

## 3.2.
## AI processing hardware

Increasingly complex models require more parameters and greater computational capabilities. As we saw earlier, training advanced models demands thousands of GPUs or TPUs operating in parallel for weeks or even months. This drives up costs and places enormous pressure on data centres. These facilities must adapt – not only to handle greater computational loads but also to ensure speed and efficiency, particularly for applications requiring real-time responses.

Meanwhile major technology companies are allocating 40%-50% of their investments to the development and deployment of specialised hardware for training and operating advanced AI models, as illustrated in Figure 5.


Figure 5: Investments in hardware and other assets made by tech companies for AI Data Centres from January to August 2024 [Source: Axon Partners Group based on JP Morgan[15]]

15 https://am.jpmorgan.com/content/dam/jpm-am-aem/global/en/insights/eye-on-the-market/a-severe-case-of-covidia-amv.pdf

And this investment seems to be showing results. Chips used today are up to 30 times faster than those from 2017. NVIDIA's Blackwell chips are said to be between 20 and 30 times faster than their predecessors, the H100 GPUs[16].

**Data centres have managed to maintain relatively stable energy usage even as their workloads have grown. However, energy demand from data centres is projected to grow by 160% by 2030.**

Another example comes from the scientific journal Optica[17], which announced late last year that researchers from the Chinese University of Hong Kong have developed an artificial neuron that emulates the functions of a biological neuron but with processing speeds a billion times faster. Using laser technology, this neuron can reach speeds of 10 GBaud, processing data equivalent to 100 million heartbeats or 34.7 million handwritten digital images in just one second.

In parallel, IBM[18] announced in December 2024 a breakthrough in optical technology that could drastically improve how data centres train and run generative AI models. IBM researchers have developed a new process for co-packaged optics (CPO), the next generation of optical technology, enabling connectivity within data centres at the speed of light, complementing existing short-range electrical cables. This innovation could significantly increase data centre communication bandwidth, minimise GPU downtime, and accelerate AI processing.

16 https://www.reuters.com/technology/artificial-intelligence/nvidias-design-flaw-with-blackwell-ai-chips-now-fixed-ceo-says-2024-10-23

17 https://www.optica.org/about/newsroom/news_releases/2024/december/laser-based_artificial_neuron_mimics_nerve_cell_functions_at_lightning_speed

18 https://newsroom.ibm.com/2024-12-09-ibm-brings-the-speed-of-light-to-the-generative-ai-era-with-optics-breakthrough

## 3.3.
## Energy

Data centres have managed to maintain relatively stable energy usage even as their workloads have grown. However, energy demand from data centres is projected to grow by 160% by 2030.

According to figures from the International Energy Agency (IEA), data centres today consume between 1% and 2% of global energy. This share is expected to rise to 3%-4% by the end of the decade. AI models specifically are estimated to account for approximately 0.7%-1.4% of global energy consumption today, potentially increasing to 2.1%-2.8% by 2030. This growth in global electricity demand is shown in Figure 6.

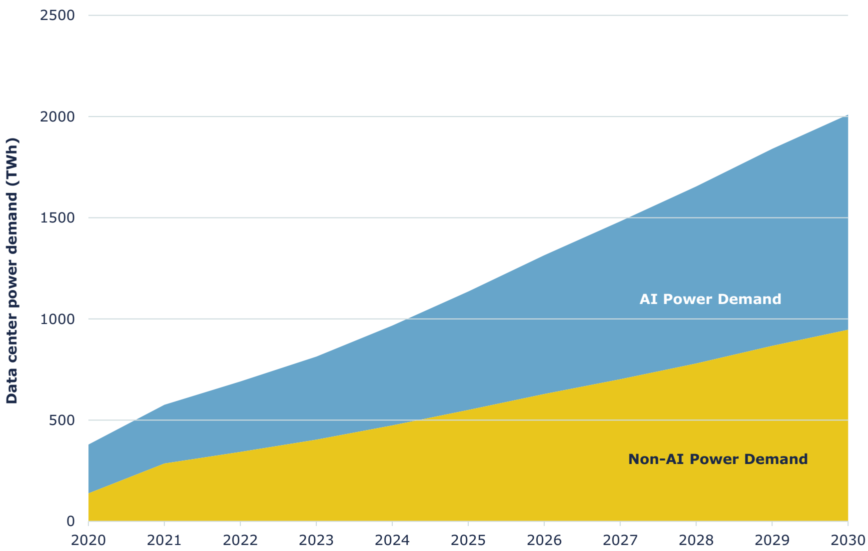



Figure 6: Estimated data centre power demand 2020-2030 [Source: Axon Partners Group based on Goldman Sachs[19]]

19 https://www.goldmansachs.com/images/migrated/insights/pages/gs-research/gs-sustain-generational-growth-ai-data-centers-global-power-surge-and-the-sustainability-impact/sustain-data-center-redaction.pdf

The increase in energy consumption is related to several specific factors, including the growing computational demands of large-scale AI models, the need for high-performance GPUs and the continuous operation of cooling systems. For example, as highlighted by BestBrokers on its website[20] and based on data from the Electric Power Research Institute, processing a ChatGPT query requires, on average, nearly 10 times more electricity than performing a Google search.

This increase in computing intensity translates directly into higher energy consumption. While traditional data centres typically operate at 3–6 kW (kVA) per rack, high-performance computing (HPC) data centres exceed 10 kW per rack. AI-driven data centres push these limits even further, with power demands reaching 60–100 kW per rack, driven by the extreme computational requirements and the need for advanced cooling solutions.

As for future demand, the scientific journal Joule[21] reports that a potential scenario of ultra-large-scale AI adoption suggests that Google’s AI alone could consume as much electricity as a country like Ireland (29.3 TWh per year).

A large portion of this energy comes from fossil fuel-based sources; the transition to a more sustainable energy model remains uneven and slow in many regions. As highlighted in a report by the International Air Transport Association[22], data centres are responsible for between 2% and 3% of global greenhouse gas emissions, more than the total emissions of the aviation industry. If current trends continue, carbon dioxide emissions from these facilities could double between 2024 and 2030. This would have not only ecological consequences but also economic ones, with what is described as a "worldwide social cost" estimated at between $125 and $140 billion, according to recent analyses published by Goldman Sachs[23].

While advances in energy efficiency per server have been made through innovations in chip design, liquid cooling technologies, and improved workload distribution, as the demand for processing and training AI models grows exponentially, data centres’ electricity consumption will grow. So can the expansion of AI be managed in a way that does not compromise global sustainability goals? That is one of the questions we will address in chapter 4.

20 https://www.bestbrokers.com/forex-brokers/ais-power-demand-calculating-chatgpts-electricity-consumption-for-handling-over-78-billion-user-queries-every-year/

21 https://www.cell.com/action/showPdf?pii=S2542-4351%2823%2900365-3

22 https://www.iata.org/en/iata-repository/publications/economic-reports/greenhouse-gases-in-review-

23 https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

# 4.

# How should AI data centres be enabled?

Data centres are fundamental components for economic and technological development in general and AI computing needs in particular. It's no surprise then that many nations are planning to build or establish data centres or trying to attract data centre developers.

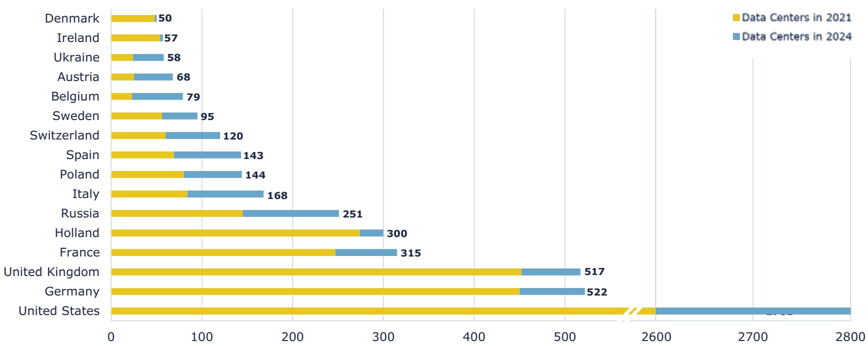



Figure 7: Number of data centres in different European countries and the US as of February 2024 [Source: Axon Partners Group based on CBRE Research[24]]

24 https://www.expansion.com/inmobiliario/mercado/2024/07/13/669172b5468aeb626e8b4588.html

And data centres developers are responding. Major hyperscalers have recently announced ambitious infrastructure projects for AI-specialised data centres. One such is the "Stargate" alliance, led by OpenAI, Oracle, and SoftBank, which plans to invest up to $500 billion in the United States over the next four years, with an initial phase of $100 billion allocated for the construction of data centres in Texas (as recently announced by Donald Trump[25]). Meta, meanwhile, has selected the ACS Group to develop one of the world's largest data centre campuses in Richland Parish, Louisiana, with an investment of €9.5 billion (about $9.77 billion)[26].

In Europe, Blackstone[27] has announced an ambitious plan to build a large data centre campus in Northumberland, United Kingdom, with an estimated investment of £10 billion (about $12.4 billion), which is in line with the general global trend to double investment in digital infrastructure over the next five years. In July 2024 it predicted that there would be $2 trillion worth of capital expenditures worldwide over the next five years to build and facilitate new data centres. In Spain, AWS[28] is positioning itself as a key player with a projected investment of €15.7 billion ($16.2 billion) in AI-optimised infrastructure through 2033, while Microsoft has committed €1.9 billion ($1.96 billion) through 2025 for its cloud infrastructure in Madrid, with a projected capacity of over 150 MW in the country by 2030[29].

Countries and policymakers are implementing measures and strategies to encourage the construction of AI data centres, but location, resources, regulation, and public-private collaboration are needed to ensure a favorable environment for expansion – issues we discuss in the next part of this chapter.

25 https://apnews.com/article/trump-ai-openai-oracle-softbank-son-altman-ellison-be261f8a8ee07a0623d4170397348c41

26 https://cincodias.elpais.com/companias/2024-12-04/meta-elige-a-acs-para-levantar-su-mayor-campus-de-centros-de-datos-con-una-inversion-de-9500-millones.html

27 https://www.bbc.com/news/articles/c3e957k9d1yo

28 https://www.aboutamazon.eu/news/job-creation-and-investment/aws-plans-to-invest-15-7-billion-in-spain-supporting-the-creation-of-17-500-jobs-annually-in-local-businesses

29 https://news.microsoft.com/es-es/2024/02/20/microsoft-quadruples-its-investment-in-ai-and-cloud-infrastructure-in-spain-to-promote-the-deployment-of-responsible-and-secure-artificial-intelligence-in-companies-and-public-administration/

4.1.
## Strategic location

Strategic locations are essential for data centre development. Factors that should be considered include:

- **Submarine cable connectivity**
  Locations near submarine cable landing stations will ensure high-speed global communication.
- **Proximity to primary energy sources**
  AI data centres should be located near power generation plants to reduce transmission losses and ensure a stable supply.
- **Local climate conditions**
  Cooler climates reduce cooling costs.

Dedicated technology parks with pre-approved land and essential services enable the development of such projects in strategically chosen locations. Examples[30] include the Maharashtra Data Centre Park in Mumbai, the development of the Condor Galaxy AI Data Centre in UAE (which has driven significant investments in renewable energy and advanced energy infrastructure), and Meta's AI Data Centre in Louisiana, which is located near submarine cables.

**Dedicated technology parks with pre-approved land and essential services enable the development of such projects in strategically chosen locations.**

30 https://indianexpress.com/article/cities/mumbai/maharashtra-cabinet-gives-approval-for-green-integrated-data-centre-park-9604187/ and https://www.opportunitylouisiana.gov/metadatacenter

## 4.2.
## Energy and resource supply

AI-intensive data centres demand enormous energy capacity and efficient cooling systems. Developers must therefore consider:

- **Low-cost energy**
  Energy expenses can account for up to 50% of an AI data centre's total cost. AI data centres' energy demands can reach hundreds or even thousands of GWh annually. Cheap energy is therefore vital.

- **Energy redundancy**
  To guarantee operational continuity in case of disruptions to the primary supply, backup systems should be established along with connections to multiple substations from different providers.

- **Availability of renewable energy**
  To reduce long-term costs and meet sustainability standards. It is useful to identify regions with renewable energy sources such as solar, wind, or hydroelectric power.

- **Access to water**
  For water-based cooling systems it is essential to ensure sustainable and reliable access to water sources.

Examples[31] of efforts to address the energy demands of AI include Argentina's Nuclear Plan, which includes the construction of a new nuclear power plant, and the Green Mountain Data Centre in Norway, which leverages Norway's abundant renewable hydroelectric energy to reduce operational costs while maintaining sustainability. Meta's Data Centre in Luleå, Sweden meanwhile, combines advanced energy storage systems and backup generators to ensure uninterrupted operations, while its Arctic Circle location naturally reduces the need for energy-intensive cooling.

31 https://www.world-nuclear-news.org/articles/argentina-aims-to-pioneer-new-nuclear-president-milei-says, https://greenmountain.no/green-mountain-data-centers-host-hpe-ai-and-hpc-infrastructure/ and https://datacenters.atmeta.com/

## 4.3.
## Regulatory frameworks and incentives

Creating an environment conducive to data centre development requires regulatory support and fiscal incentives such as:

- **Tax incentives**
  Financial measures – like tax exemptions, reduced corporate rates, reduced prices of energy and accelerated depreciation – facilitate the establishment and operation of data centres, making investments more attractive.

- **Incentives for technological innovation**
  It is crucial to promote collaboration with technology companies to integrate innovative solutions.

- **Simplified administrative processes**
  Streamlined procedures, such as fast-track permits and special technology zones (for example free zones), reduce bureaucratic delays and facilitate efficient data centre setup.

- **Clear and stable regulations**
  Transparent legal frameworks ensure investment protection, long-term stability, and compliance with environmental and operational standards, boosting investor confidence.

An example of regulatory support comes from Dubai, which is developing a free zone focused on artificial intelligence and data centres as part of the "Dubai AI Strategy"[32], offering incentives like tax exemptions to attract global technology companies.

32 https://www.wam.ae/en/article/b547v40-dmcc-launches-centre-drive-global-aisolutions

AXON CONSULTING

## 4.4.
## Public-private collaboration and investment

Public-private cooperation leverages the capabilities and capital of the private sector to accelerate data centre development. In practice this could mean:

- **Creation of joint investment funds**
  Mixed public-private investment funds can help to finance key infrastructure for AI data centres, such as high-capacity power grids, advanced cooling systems, and telecommunications connectivity.

- **Government guarantees**
  Providing legal and financial guarantees attracts investors and minimises risks in large-scale AI data centre projects.

- **Investment in talent and technological ecosystems**
  Developing educational programs, creating innovation hubs, and attracting talent can ensure the long-term sustainability of data centre infrastructures.

There are already a number of examples[33] of such initiatives. The European Union offers attractive conditions for foreign investors interested in this sector. Among individual countries, Germany has implemented a national strategy that includes visa facilitation, funding for research projects, and collaboration with academic institutions to attract top AI talent. Similarly, the Chinese government and private sector have committed over $10 billion to developing AI-focused data centres across the country. The Stargate project, mentioned in chapter 4, is a public-private collaboration that plans to invest up to $500 billion over the next four years.

These are just some of the ways in which governments can attract investment and promote economic development but also position themselves as technological leaders in an increasingly AI-driven world.

**Public-private cooperation leverages the capabilities and capital of the private sector to accelerate data centre development.**

33 https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence and https://www.reuters.com/technology/bytedance-plans-20-billion-capex-2025-mostly-ai-sources-say-2025-01-23

# 5.
# Conclusion

Computational demand is growing at a staggering pace – and AI is driving that growth. As AI models become more complex and their adoption expands, the pressure on data centres will increase significantly. Can data centres handle this load? And are governments able and willing to lead this transformation?

AI-specialised data centres present both a unique opportunity and significant challenges. These challenges include ensuring access to sustainable energy, selecting strategic locations with robust infrastructure, and attracting investment. Government support is also needed to deal with these challenges and secure the social and economic benefits AI promises.

But adapting to this new reality is not merely a challenge – it is a strategic imperative. Countries that prioritise appropriate initiatives will not only support the growth of AI but also secure a competitive edge in the global digital economy.

By laying the groundwork for advanced and sustainable infrastructure, these countries can attract international investment, create high-value jobs, and foster innovation – innovation that benefits sectors far beyond technology.

**Countries that prioritise appropriate initiatives will not only support the growth of AI but also secure a competitive edge in the global digital economy.**

AXON CONSULTING

# About
# **Axon Consulting**

Axon is an international firm founded in 2006 that provides investment and advisory services to a broad client base in the ICT and digital space in more than 70 countries across the world.

Axon's emerging technologies practice is at the core of our work and a growing priority for clients. We focus on strategy, policy, and market intelligence around a broad spectrum of next-generation technologies, engaging both private and governmental stakeholders. By partnering closely with decision-makers, Axon helps them anticipate trends, address critical challenges, and develop actionable roadmaps to harness the full potential of emerging technologies.

**Tel:** +34 913 102 894
**Email:** marketing@axonpartnersgroup.com