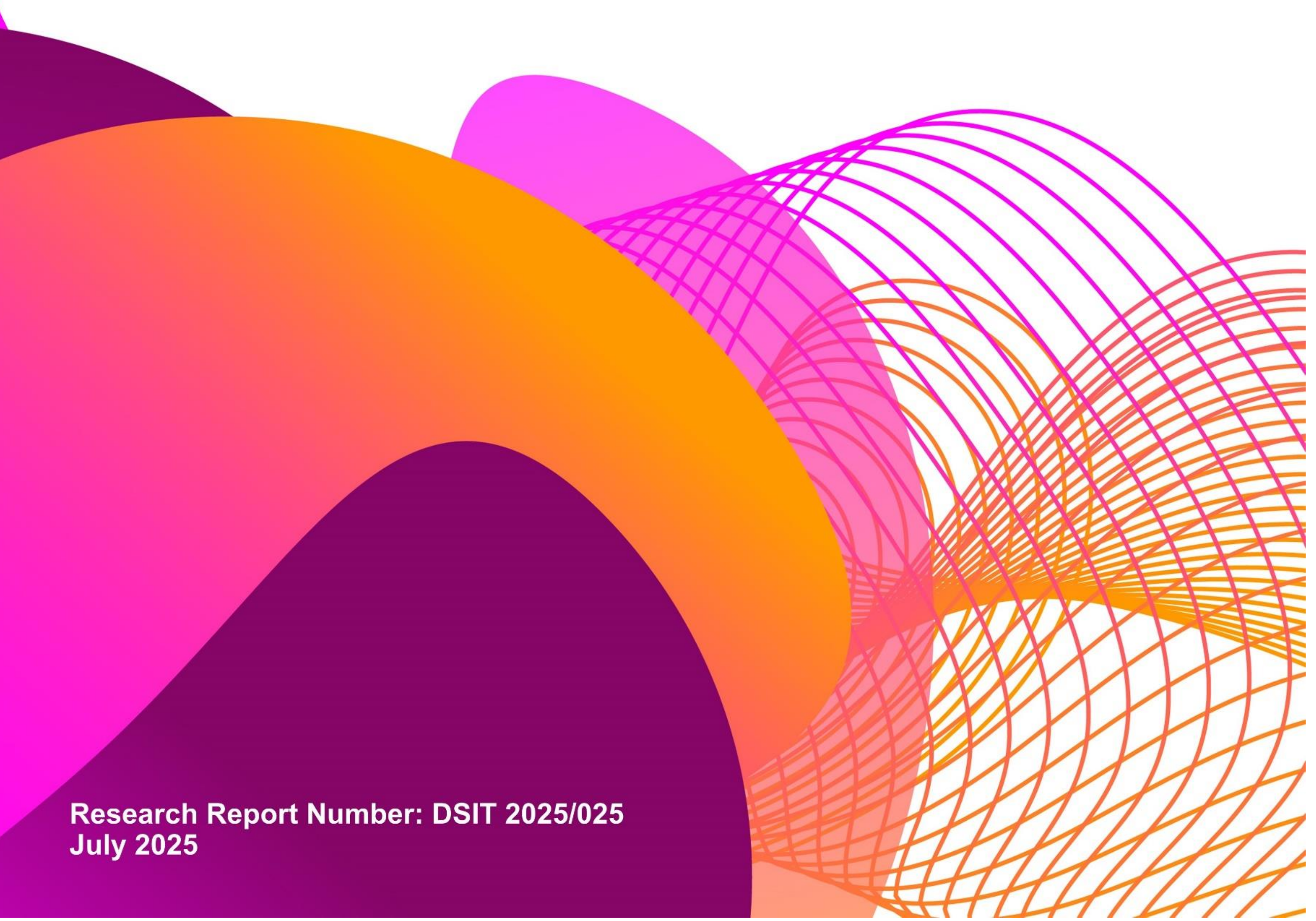
Department for
Science, Innovation
& Technology

Compute Evidence Annex



Research Report Number: DSIT 2025/025
July 2025

This publication is available at www.gov.uk/official-documents.

Any enquiries regarding this publication should be sent to us at compute@dsit.gov.uk

Printed on paper containing 40% recycled fibre content minimum

Printed in the UK by HH Associates Ltd. on behalf of the Controller of His Majesty’s Stationery Office

Contents

Executive Summary

1. Artificial Intelligence (AI) capabilities are advancing rapidly, with significant developments in foundation models like OpenAI's ChatGPT. In 2024 alone, 165 new large-scale models of above an 100,000 exaFLOPs (Floating Point Operation) training compute threshold were released, up from just 30 in 2022.

2. UK AI compute demand is projected by McKinsey to grow, with the IT load growing from 1.8 gigawatts (GW) in 2024, to as much as 13.6GW by 2035 in a high-demand scenario - a 7.5x increase.

3. Inference is expected to overtake training as the dominant driver in UK AI compute demand, with some scenarios suggesting that up to 62% of future demand could come from Gen-AI inference by 2035. This shift will be driven by widespread adoption in sectors like retail, banking, and public services, where real-time, low-latency AI is also becoming more important.

4. Even under optimistic conditions, current McKinsey projections suggest the UK's compute supply could fall short. Without intervention, the UK could face a 5GW compute gap by 2030. This gap risks undermining national security, economic growth, and the UK's ambition to lead in AI.

5. Public sector compute demand could reach 0.2GW by 2030, 20 times more than the combined capacity of the UK's current public flagship supercomputers, Isambard-AI and Dawn. This growth is largely driven by growing AI healthcare applications and complex modelling.

6. Rising demand is partly balanced by more efficient models and hardware. Energy use per ChatGPT prompt has fallen by a factor of 10 since 2023.[1] *[The original content of this section*

referring to DSIT's environmental impacts model has been temporarily removed. We keep analysis under routine review, and are updating this modelling to ensure it reflects the most up to date assumptions and analysis. This section will be updated by summer 2026.]

Note, future projections are based on modelling and analysis conducted by McKinsey & Company. It is intended to inform ongoing policy development and does not represent a final cross-cutting view of government. Scenario projections are subject to uncertainty and may evolve as further evidence and perspectives are incorporated.

Introduction

AI has the potential to transform the economy and society. The IMF estimates that the widespread adoption of AI could add up to 1.5% to annual productivity growth.[3] Compute will be needed to underpin this transformation. From developing the latest AI models that increasingly need more compute to analyse the vast amounts of data these models rely on, to the growing inference requirements used to respond to AI prompts as adoption rises.

Compute is essentially the computer chips used to process data, whether processing the training data to build AI models or processing user prompts. Compute can be broadly categorised into several types. Traditional workloads or general-purpose compute is typically powered by central processing units (CPUs) and support standard workloads like databases and web services. Graphics processing units (GPUs) were originally designed to render images and video but are now more widely used for AI and scientific computing. High-performance compute (HPC) clusters combine CPUs and GPUs to tackle complex scientific and AI problems, often housed in supercomputing facilities.

AI capabilities are developing at a fast pace. Most developments in the last decade were enabled by machine learning – an approach to AI where models are trained by exposure to enormous amounts of data. More recently, there has been rapid AI developments following breakthroughs in foundation models. These are large AI systems capable of completing a range of different tasks, such as OpenAI's ChatGPT model which can respond to and create text and images.

Alongside this has been a rapid rise in the adoption of AI. Currently, 21% of UK businesses are using AI in July 2025, up from 13% a year ago.[4] The most popular types of AI are text generation using large language models (9.6% of firms), followed by visual content creation (7.1%) and data processing using machine learning (5.3%).[5] Consumer adoption is reportedly higher, with four-in-ten adults using AI chatbots for personal use at least monthly. [6]

These are key drivers of the rapid growth of the UK AI sector. The sector contributed £11.8bn in gross value added in 2024 and grew at a rate faster than the UK average.[7] There were 650 new AI company registrations in 2024, meaning the total number of AI firms was over 5,800 in 2024. Just under one fifth (17%) of UK AI revenue is generated by AI infrastructure firms.[8]

This annex sets out the latest evidence on AI compute to inform the development of the Compute Roadmap. From developments in the computing chips that make compute, to estimates of the current and future demand and supply of compute. Building on this foundational understanding, the following section explores the latest trends in compute technology that are driving these advancements.

Trends in Compute

As AI demand grows, AI-specific compute has appeared as a distinct category. This includes AI training compute, which requires massive GPU resources, or equivalents like Google's tensor processing units (TPUs). AI-specific compute is also used to support inference, which involves running trained models to serve specific applications. Edge computing enables real-time processing by bringing computation closer to the data source. This is especially critical for applications that require ultra-low latency, such as autonomous vehicles. Cloud compute offers scalable access to these resources, while edge compute enables real-time AI processing closer to data sources.

AI compute demand is driven by three converging forces: the rapid expansion of model size, ongoing growth in the scale and variety of datasets, and the accelerating adoption of AI across sectors. In 2024, 165 new large-scale models (above an 100,000 exaFLOPs compute threshold) were released, up from just 30 in 2022.[9] The growing number and complexity of models has been accompanied by significant growth in the size of datasets used to train models, which has grown by 3.7x annually since 2010.[10] Together, the computational power used in state-of-the-art AI training runs has been rising seven times faster than Moore's Law.[11] Machine learning chip performance has improved by 31% annually[12], doubling roughly every 2.2 years.[13] These gains can reduce the cost per FLOP. For example, the cost per token for OpenAI's GPT-4o mini model has dropped by 99% compared to earlier models.[14] While GPUs are becoming more efficient by delivering more performance per unit of energy or cost, the total cost of acquiring and operating them remains high, and the rapid growth in AI model size and usage is outpacing these efficiency gains. For example, the overall cost of training AI models has still increased by 2.6x per year since 2016.[15] More complex models have led to better AI applications. Building on the recent increases outlined above, AI adoption is also expected to boom in the coming years, with developers continuously refining and deploying new models to meet evolving needs across healthcare, finance, retail, and public services.

The growing demand for AI is shifting from model training to inference.[16] This is the stage where a trained model is used to generate outputs, such as answering questions or making predictions. While each inference task is far less computationally intensive than training, it happens much more frequently and at scale. This makes both low latency and cost efficiency critical. For example, generating 10 tokens per second per user (equivalent to around 450 words per minute) may not sound large in isolation, but when multiplied across millions of users, it results in a substantial and continuous compute load.[17] This throughput is also necessary to maintain a smooth, real-time user experience. In contrast, training involves teaching models using massive datasets and typically requires thousands of GPUs or TPUs running in parallel for extended periods. Training a model like GPT-4, for instance, is estimated to consume between 10MW and 100MW of power,[18] highlighting its significantly higher computational demands compared to inference.

Inference is becoming a dominant driver of compute demand, in particular Gen-AI applications, has implications for compute architecture. Compute architecture needs distinct levels of data granularity, also known as floating points (FP). For training, accuracy matters most, so FP16 or FP32 is preferred. In inference, speed, lower memory use, and reduced precision are priorities, making FP8 or FP4 common choices. More recent GPUs, like Nvidia's Blackwell, have been prioritising FP8 inference performance as a result, as well as some compute systems taking a mixed approach overall.[19]

Inference can account for the majority of total compute costs. While training models is a one-off investment, ongoing inference demand, especially as adoption of AI increases, will require a greater share of total compute. McKinsey research for DSIT estimates that inference's share of total UK compute demand could rise from 16% in 2023 to 57-71% in 2035, depending on the forecast scenario. Already, it was estimated that OpenAI's ChatGPT costs over $700,000 per day in cloud infrastructure in 2023, which is around 0.36 cents per query.[20] The compute per token could increase by up to 1.3 times a year, although this estimate carries some uncertainty due to evolving model architectures and deployment strategies.[21] This growth is largely driven by the increasing complexity and capability of AI models, such as their ability to process longer and more context-rich prompts from Gen-AI applications. This is only partially counteracted due to increases in algorithmic efficiency in how AI models work.[22] For example, OpenAI found that by 2019, it took 44x less compute to reach the same ImageNet model performance as in 2012 due to better algorithms.[23]

Low latency and sovereign capabilities requirements are also growing. Gen-AI applications often involve fast response requirements for text and voice interactions and are latency and sovereignty sensitive, so they need UK-based data centres. For example, in autonomous driving, low-latency AI models are essential for processing sensor data, such as LiDAR and camera feeds, in milliseconds to make split-second decisions like obstacle avoidance or speed adjustments.[24] Another example is Singapore's GovTech developed PAIR, an LLM-powered assistant used across 100 government agencies to draft emails and conduct research. This is a Gen-AI deployment that handles sensitive government data and thus runs in sovereign cloud infrastructure, illustrating the rising need for local inference capacity.[25] These examples reflect a broader trend: as Gen-AI moves from experimentation to deployment, UK-based data centres are becoming critical to meet performance expectations and regulatory requirements.

There is demand for compute to develop narrow models within the UK. AI training tasks are generally not latency sensitive, with over 100ms round-trip latency being acceptable, allowing for the use of overseas compute resources.[26] However, sectors such as healthcare, energy, defence, and public services require complete sovereignty due to data sensitivity.[27] For example, Congenica served as the exclusive clinical decision support partner for the NHS Genomic Medicine Service, performing all genomic data analysis within the UK.[28] Similarly, AION, developed by Nostos Genomics, was clinically validated through the UK's 100,000 Genomes Project and supports small labs in diagnosing rare diseases using sovereign genomic data.[30] MILTON, a machine learning tool trained on data from nearly 500,000 UK Biobank participants, can predict over 1,000 diseases before diagnosis.[31] These use cases highlight the need for domestic infrastructure to ensure compliance with data protection regulations and to maintain public trust in the use of sensitive health data. Future compute demand projections are shaped by these use cases.

The Case for Intervention

To ensure the UK remains competitive and resilient in the age of AI, government intervention in compute infrastructure is essential. Government intervention in this case is providing the infrastructure and resources needed for compute to ensure that critical national needs are met, particularly where the market falls short. This includes investing in public infrastructure, subsidising access for researchers and SMEs, prioritising high-impact public projects, enforcing secure data standards, and fostering cross-sector collaboration to meet national strategic needs. This section sets out the analytical and economic rationale for intervention, structured around the four strategic objectives of the UK Compute Roadmap.

Building a modern public compute ecosystem providing the capacity and certainty UK researchers and innovators need:

The UK's compute demand is projected to grow, driven by the widespread adoption of AI across science, the economy and the public services. However, access to compute remains a significant barrier, particularly for researchers, startups, and SMEs. High setup and operating costs, such as GPUs costing up to £36,000 for a Nvidia H100 GPU, can prevent researchers from entering the AI market. [33] As a result, the UK's innovation potential is constrained, with much of its AI research capacity currently reliant on private actors. The market failure lies in the under-provision of public compute resources, which are essential for AI research and innovation. The UK must invest in a robust and scalable public compute infrastructure to meet the growing demands of AI research and development. In the UK, demand for such resources far exceeds supply. UKRI's AI GPU cluster was oversubscribed by 350% last year as per UKRI reporting and commercial cloud costs are prohibitively high.

Government intervention in AI compute directly addresses an existing market failure of the under-provision of public goods. Government programmes such as the AI Research Resource (AIRR) are critical to expanding access for researchers, SMEs, and mission-driven projects. By addressing the oversubscription of existing resources and high commercial cloud costs, the UK can democratise access to compute and unlock innovation potential across academia and industry.

Putting Compute to Use, Powering Innovation Across the Public and Private Sector

It is essential to ensure that compute is directed toward high-impact, mission-driven projects. By prioritising access for foundational science, national AI programmes, and SMEs, the government can ensure that compute becomes a catalyst for transformation across public services and national priorities.

AI is already delivering real-world impact in areas aligned with the government's Plan for Change. From virtual biopsies to AI-assisted IVF and predictive analytics in cardiology, UK researchers are pioneering applications that improve lives and reduce costs. Yet, these breakthroughs depend on access to compute infrastructure that is secure, scalable, and mission aligned. For example, RETFound is the first AI foundation model in ophthalmology, developed by researchers at UCL and Moorfield's Eye Hospital to diagnose eye diseases based on historic eye scan images. The model was trained with 8-16 Nvidia A100 GPUs over one month using 1.6 million retinal images.[34] Assuming it operated continuously for one month, this equates to approximately 2,300 to 4,600 kWh.[35] However, compromises to the sophistication of the training model were made due to limited compute resources and the need to store data securely. For example, the resolution of eye scan images was reduced from 1000 x 1000 pixels to 128 x 128 pixels. Additionally, the batch size was reduced to fit into the available resources, which increased the time needed to train the model.[36]

This surge in AI-driven R&D directly supports the UK Government's missions to improve public services through advanced technologies, strengthen scientific leadership, and drive economic growth through innovation. The UK has had notable early success in leveraging AI for science to improve lives and deliver commercial impact, for example:

- Clinicians at Imperial College Healthcare NHS Trust and researchers at Imperial College London, and University of St Andrews have used advanced AI to make IVF treatments more effective. This AI had to analyse thousands of high-resolution microscope images, a process requiring considerable GPU time which required training a complex neural network on a large dataset of medical images.[37]
- Glasgow University spinout Chemify has raised £36m of investment to digitise chemistry and automate chemistry laboratories by combining AI and robotics. Such systems rely on both cloud compute for training and on-premises edge compute for real-time experiment control.[38]
- AIRE is an AI model designed by researchers at Imperial College to support clinical decision-making by analysing patient history and imaging scans (electrocardiogram (ECG)), improving diagnostic outcomes through predictive analytics.[39]
- Researchers at NIHR IBTC have shown combining medical imaging with AI can be used to provide a 'virtual biopsy' for cancer patients, by extracting information about the chemical makeup of lung tumours from medical scans. Processing image data in real time to be clinically useful demands reliable local compute resources.[40]

Building AI Infrastructure to Keep the UK at the Cutting Edge of AI Development

To ensure the UK can lead at the frontier of AI, government intervention must prioritise the development of cutting-edge AI infrastructure that supports both the training and deployment of advanced models. This includes investment in high-performance computing, AI-optimised data centres, and specialised hardware such as GPUs and TPUs. As AI adoption accelerates across sectors, from healthcare and finance to manufacturing and public services, the demand for scalable, low-latency, and energy-efficient compute infrastructure is growing exponentially. Without intervention, the UK risks falling behind global competitors who are rapidly expanding their AI capabilities through sovereign infrastructure and public-private partnerships.

Through initiatives like AI Growth Zones, the UK is laying the groundwork for a new generation of AI infrastructure capable of delivering at least 6GW of AI-ready capacity by 2030. These zones will not only anchor national-scale training and inference but also serve as platforms for innovation, investment, and regional growth. By embedding compute infrastructure within a broader ecosystem of energy, talent, and research, the UK can accelerate AI adoption across sectors, from healthcare and defence to finance and manufacturing, while ensuring that the benefits of AI are realised across the country. This forward-looking approach positions the UK to lead in the global AI race, not just as a consumer of technology, but as a builder of the infrastructure that powers it.

Creating Sovereign, Secure and Sustainable Capability

Compute infrastructure is no longer just a technical asset; it is a strategic enabler. The UK must act decisively to avoid over-reliance on foreign infrastructure, which poses risks to data sovereignty, latency-sensitive applications, and national resilience. Sovereign compute capacity is particularly critical for sectors like defence, healthcare, and public services. Moreover, the UK has a strong foundation in chip design and systems engineering. By investing in domestic infrastructure and supporting UK-based innovation, the government can build strategic advantage across the AI and compute value chain.

Without intervention, the UK could face the potential risk of lagging in its efforts to become a strategic leader in AI. Other countries are also boosting their AI compute capacity. France is investing €109 billion to build competitive and independent infrastructure[41]. Meanwhile, India is focusing on democratising compute access through public-private partnerships, supported by a $1.3 billion commitment to expand infrastructure and subsidise AI resources for vital sectors.[42] These approaches exemplify how nations leverage unique strengths to overcome compute challenges, offering potential pathways for the UK to bolster its AI aspirations.

Compute Demand

Compute demand is rising across all parts of the economy. The rapid expansion of Gen-AI inferencing, particularly within consumer-facing applications is expected to dominate demand due to increasing interaction frequency and model complexity. McKinsey’s 2025 global analysis reported that 71% of companies now use Gen-AI in at least one function, with marketing and sales leading the way.[43] Adoption is highest in sectors that are expected to benefit from AI the most, such as the Information and communications sector (37% adoption rate, well above the all-sector average of 18%) and Professional, scientific and technical activities (32%).[44] A closer look at the distinct use cases for inference and training helps understand the nuances of compute demand.

Total UK compute demand projected to be between 4-7 times higher than today’s demand by 2035.[45] DSIT commissioned McKinsey to model both demand and supply of compute over the next 10 years. This generated three scenario-based demand projections, shaped by varying assumptions around AI adoption and technical progress. The range in projections reflect the inherent uncertainty surrounding the evolution of AI and that they largely depend on existing trends continuing. This is preliminary analysis and does not represent an agreed government forecast of demand and supply of compute in the UK: we will continue to refine this analysis to develop a refined picture on the future balance between supply and demand.

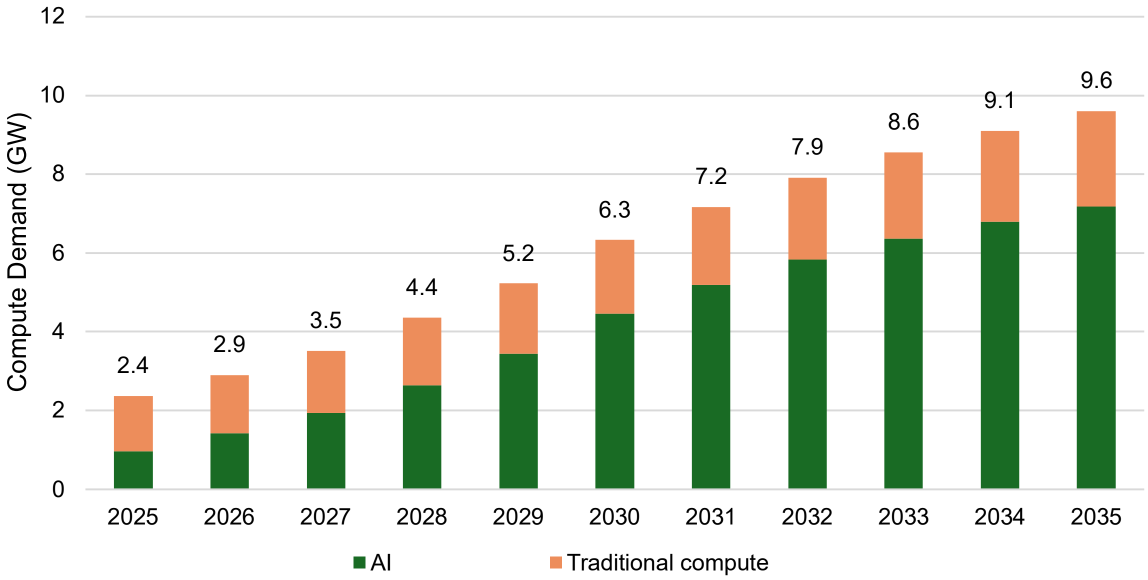
The main scenario is for total compute demand to grow from 1.8GW in 2024 to 9.6GW by 2035 (Table 1), driven by steadily increasing AI integration across sectors. However, this could vary between 7.4GW and 13.6GW. UK AI compute demand is measured here in **gigawatts (GW) of IT load** – a proxy for power needed to run the required hardware. 1GW of compute capacity can correspond to operating about 1 million high-end GPUs.[46]

Table 1: UK compute demand scenario-based projections by 2035 (GW, IT load capacity)

	High		Medium		Low	
	GW	% of total	GW	% of total	GW	% of total
Total Compute Demand	**13.6**		**9.6**		**7.4**	
AI	11.2	82%	7.2	75%	5.0	68%
Private Sector	10.9	97%	6.9	96%	4.8	96%
Public Sector	0.3	3%	0.3	4%	0.2	4%
Traditional	2.4	18%	2.4	25%	2.4	32%

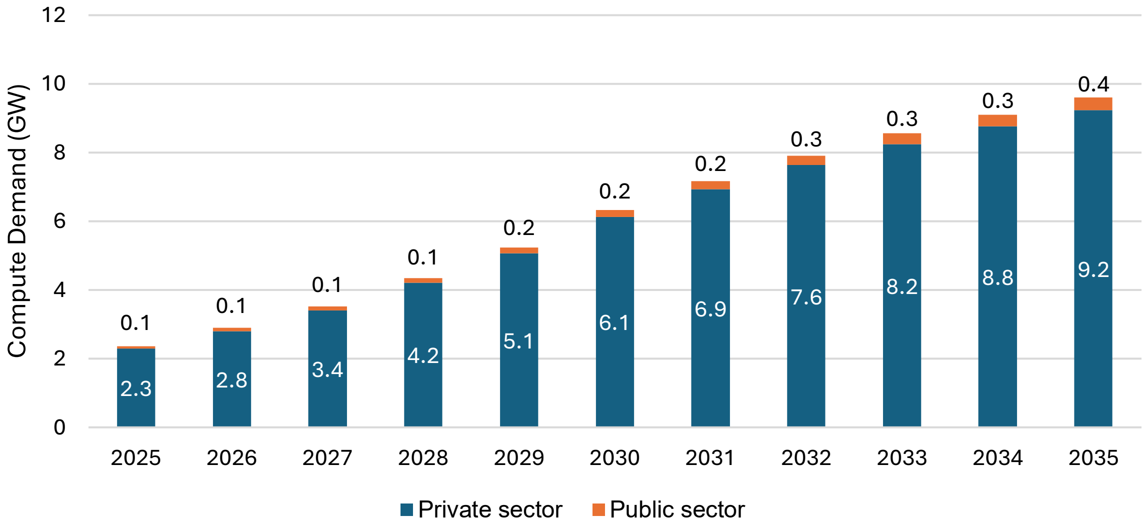
Specific AI compute demand is expected to be 13 times higher by 2035.[47] Traditional compute demand is expected to grow 6% annually on average between 2024 and 2035, which is five times slower than for AI compute demand (27%).[48] Consequently, while AI was only a third of total demand in 2024, it's share will rise dramatically to a three-quarters by 2035 (Figure 1).

Figure 1: UK compute demand by use case type in the Medium Scenario, 2025-2035 (GW)



It is expected that a high proportion of this compute growth will be from the private sector. In the medium scenario, private sector demand steadily increases to 6.1GW by 2030 or 12,000 exaFLOPs, while public sector demand remains relatively modest, rising from 0.1GW to 0.2GW over the same period (Figure 2).[49] This trend reflects the increased use of AI applications in Gen-AI and inferencing.

Figure 2: UK compute demand by user type in the Medium Scenario, 2025-2035 (GW)



Public sector demand alone could need up to 0.2GW of compute capacity by 2030, primarily driven by inference in healthcare uses.[50] This is the equivalent of up to an additional 420 exaFLOPs.[51] To put that in perspective, it means requiring 20 times more compute than University of Bristol's 'Isambard-AI' supercomputer and University of Cambridge's 'Dawn' supercomputer combined. Isambard-AI and Dawn currently provide less than 0.01GW of sustained compute.[52] This is because the public sector is more likely to use large, complex datasets for use cases such as modelling and simulations.

Compute Supply

AI data centres have evolved to meet the unique demands of AI workloads. Unlike traditional data centres that support general-purpose computing, AI data centres are purpose-built for high-intensity tasks. They rely heavily on GPUs and other accelerators rather than CPUs. They require advanced cooling systems due to higher energy demands and use high-performance storage to manage vast volumes of unstructured data. These centres are designed for rapid scaling and are driving a surge in global infrastructure investment. The UK has limited supercomputing facilities. Most of the world's 500 fastest supercomputers are in the US and China.[53]

While AI workloads can technically be hosted anywhere in the world, this introduces challenges around data control, latency, and regulatory compliance. Investing in sovereign compute enables the growth of AI data centres while ensuring long-term resilience. Sovereign compute refers to computing infrastructure and data that is physically located in the UK, which could be either owned or directly controlled by the UK Government. This includes publicly owned supercomputers like ARCHER2 and Isambard-AI, as well as government-rented capacity within UK-based data centres.

Specialised hardware

Over the past decade, GPUs have become a cornerstone of AI development, particularly for training deep learning models.[54] Their architecture makes them fundamental to AI and machine learning algorithms.[55] For example, a Nvidia Tesla V100 GPU has 5,120 cores, while a high-end Intel Xeon CPU has 28 cores, enabling greater parallel processing and more tasks to be done simultaneously. It means GPUs can be up to 100 times faster than CPUs.[56] Companies like Nvidia and AMD have been at the forefront of developing high-performance GPUs that significantly accelerate the training process.[57]

AI has also led to further advancements in hardware.[58] New chips like Tensor Processing Units (TPUs), Neural Processing Units (NPUs) and application-specific integrated circuits (ASICs), and field-programmable gate arrays (FPGAs) have been developed to meet performance and efficiency demands.[59] TPUs and NPUs are custom designed to run AI tasks faster and more energy efficient. For example, Google's first-gen TPU delivered between 15-30 times speed-up on inference with up to 80 times less energy compared to a contemporary CPU.[60] In 2024, the trend accelerated with the integration of AI chipsets into consumer devices like PCs, enabling on-device intelligence and hybrid AI processing.[61] Looking ahead, the AI hardware landscape is expected to diversify further, with innovations in AI companion devices, wearables, and edge computing solutions.[62] These developments will be driven by demand for real-time, low-latency AI experiences and the commercialisation of emotionally intelligent, IP-integrated AI products.[63]

Edge computing is leading to hardware changes. Hardware for on-device AI is improving to run advanced models locally. This is gaining traction for applications that demand low latency and real-time processing, such as autonomous vehicles, industrial IoT, and smart cities. IDC estimates that global spending on edge computing is set to grow by 13.8%, reaching $380bn by 2028.[64] This is driving investment in tinyML and energy-efficient chips that can perform inference on smartphones, sensors and vehicles.

This evolution in hardware is also closely tied to the varying compute needs across sectors. In healthcare, AI supports tasks like real-time image analysis, patient monitoring, and genomic studies, requiring robust computational power. For example, Queen's University Belfast develops AI diagnostic tools that process medical images with remarkable speed and accuracy. [65] This relies on specialised accelerators and FPGAs for efficient, low-latency performance. [66] In Convolutional Neural Net (CNN) inference tasks, FPGAs have proven 5 times faster latency (30 milliseconds vs 150 milliseconds) and over 6 times greater energy efficiency compared to GPUs.[68]

In scientific research, compute demands are often more intensive.[69] Applications such as climate modelling, astrophysics, and materials science require high levels of floating-point precision and massive parallel processing.[70] Such need large-scale data analysis and high-performance computing tasks. For example, the Multimodal Universe is the largest dataset designed for AI models, having been trained on 100TB of data.[71]

However, to fully use this specialised hardware, organisations must also invest in skills and talent. AI professionals who directly use AI compute also need the related skills and training to use it effectively.[72] The complexity of modern compute environments, which include specialised hardware such as GPUs, TPUs, and NPUs, requires a deep understanding of their architecture and functionality.

Data centres

Increased demand for AI is expected to drive the UK's data centre design.[73] The design of data centres to support AI compute is evolving, reflecting the changing demands of AI workloads and the increasing focus on factors like latency, data sovereignty, and resilience. There is a growing need for faster digital connectivity between racks, which is especially critical for latency-sensitive AI tasks that often require response times in the range of 15–200 milliseconds.[75] The ability to build infrastructure that is tailored to optimise interconnect speeds and reduce bottlenecks. The broader trend is toward a hybrid[76] and distributed model.[77] Large, centralised sites are complemented by smaller, geographically dispersed facilities. This includes edge deployments and decentralised architectures to support the diverse needs of AI applications.[78]

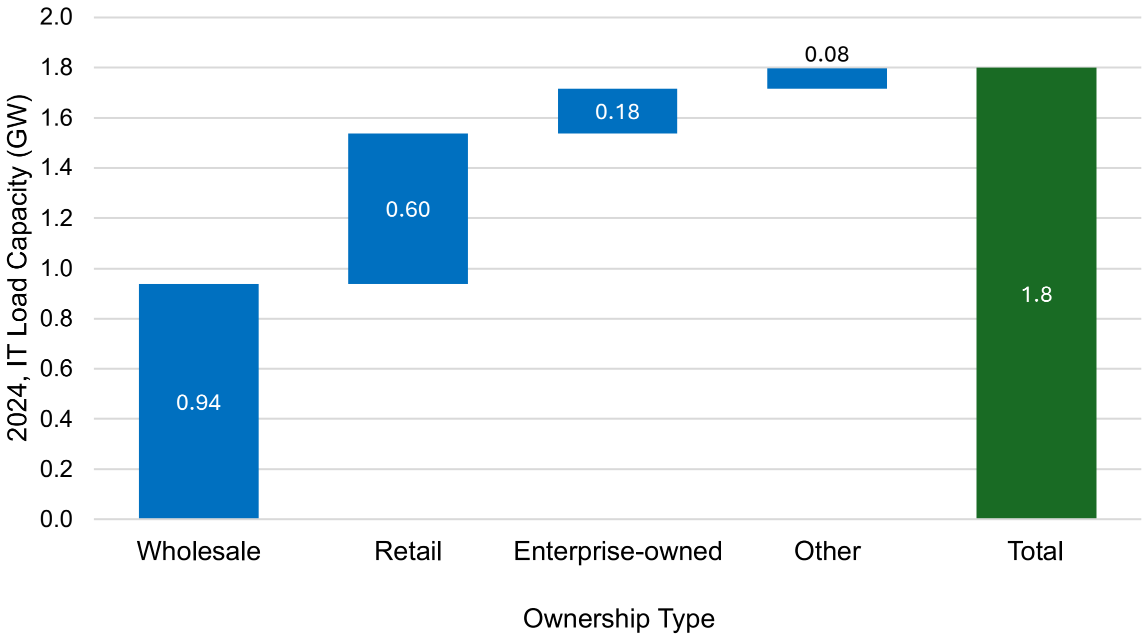
Simultaneously, there is also an increasing interest in decentralised data centre strategies. The growth in micro data centres and edge computing means that data and compute resources are distributed across multiple locations rather than relying on a single central point.[79] This approach offers potential benefits in terms of enhanced security by reducing the risk of a single point of failure, improved data privacy by allowing users more control over their information, and increased reliability through geographic redundancy.[80]

The growth of global over local hyperscalers of the UK's data centre supply carries significant implications for the UK. A hyperscaler is a large-scale data centre that provides cloud computing infrastructure and services at massive scale. Global hyperscalers offer benefits such as the rapid development of large-scale facilities like data centre campuses, insourced AI hardware that reduces dependency on other hardware developers, and support for large-scale international enterprises requiring substantial capacity.[81] However, this approach also comes with risks, including a potential loss of economic benefits from local construction and operation,[82] as global hyperscalers often rely on imported equipment and talent.[83] Additionally, there may be concerns regarding data sovereignty,[84] as locally constructed data centres could still allow foreign access to sensitive information.[85] Furthermore, global hyperscalers typically charge higher margins,[87] leading to increased costs for consumers, and their dominance may overshadow local hyperscalers or co-locators, limiting local capacity development.

Projected Compute Supply

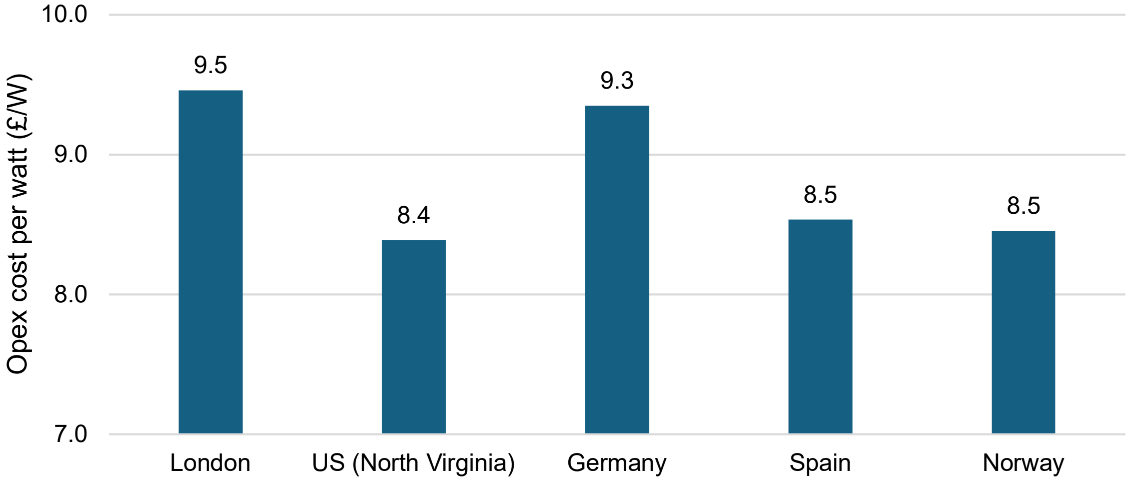
Around 1.8GW of UK's IT load (power capacity) is currently used to power around 330 data centres around the country.[89] This includes both workloads with a critical dependency on UK data centres (either given latency and/or data sovereignty requirements) and those that can be satisfied abroad.[90] There is a dominant hyperscaler presence through leasing wholesale co-locators, and this accounts for 58% of the DCs in the UK (Figure 3).[91] Historically, UK's estimated share of global hyperscaler DC capital expenditure has been around 9% and is expected to fall to 5% by 2030.[92] This could be driven by power costs, total lead times,[93] construction costs and availability of land.

Figure 3: UK data centre capacity, 2024 IT load capacity (GW)



High implied training costs for a model could deter private investment developers from locating in UK to abroad. The UK total lead time is over two years compared to the US, driven by planning delays.[94] The UK's construction costs are high versus Europe and US, with London ranked as most expensive city for construction by the ICC index 2024.[95] The UK's power costs are around four times higher than the US in 2023.[96] It's estimated that the implied training cost, if ChatGPT-4 was trained in the UK, is about £9.5bn.[97] This is around a £1bn cost difference between the US, largely driven by the operating cost for a large-scale data centre required to support compute (e.g., 1GW for ChatGPT-4).[98] The UK is a less attractive market for training from an operating cost only perspective.[99] Spain may be a more attractive market for LLM developers to locate their European hubs given more favourable conditions.[100]

Figure 4: Implied training and data centre operating cost for ChatGPT-4 (2023) by country (per watt)[101]



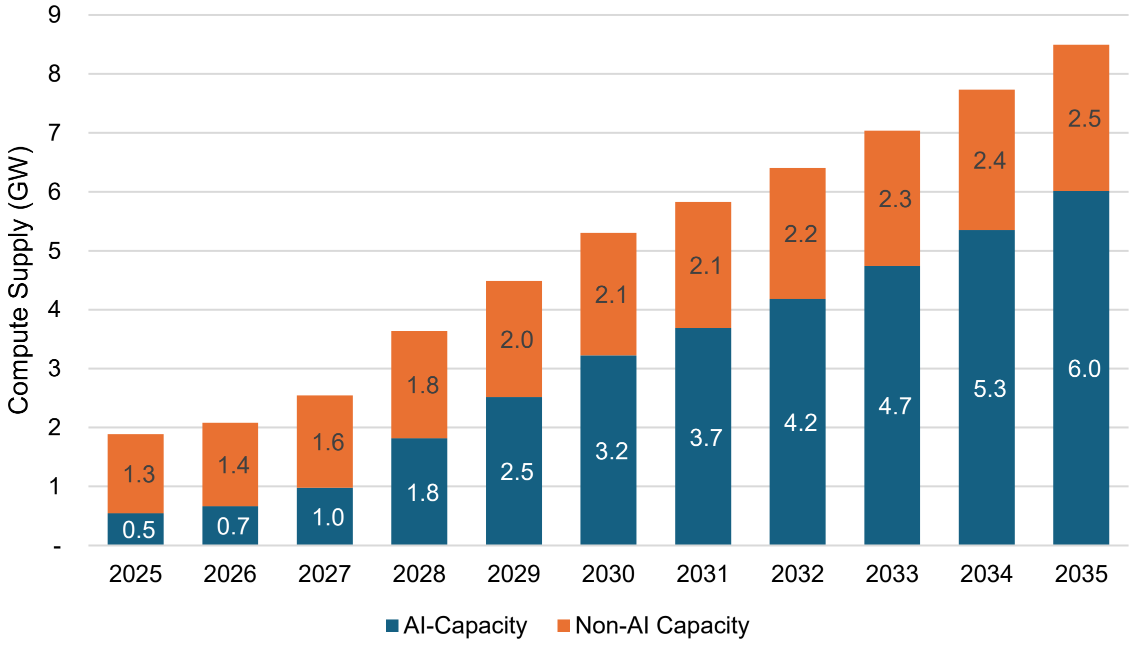
Future UK compute supply could increase to 5.4-11GW by 2035, driven by AI data centre growth (Table 2).[102] McKinsey's analysis models three UK data centre (DC) supply scenarios: low, medium, and high. The scenarios are driven by variations in investment flows, regulatory efficiency, and infrastructure readiness. This is preliminary analysis and does not represent a final government view of supply of compute in the UK: we will continue to refine this analysis.

Table 2: UK compute supply scenario-based projections by 2035 (GW, IT load capacity)

	High		Medium		Low	
	GW	% of total	GW	% of total	GW	% of total
Total Compute Demand	**11.0**		**8.5**		**5.4**	
AI	8.2	75%	6.0	71%	3.6	67%
Traditional	2.8	25%	2.5	29%	1.8	33%

In the low scenario, persistent planning and connection delays, coupled with declining market attractiveness, constrain capacity growth to 5.4 GW by 2035.[103] The medium scenario assumes resolution of connection delays but retains structural challenges such as high electricity costs and complex planning permissions, enabling capacity to reach 8.5 GW by 2035 (Figure 5).[104] The high scenario envisions significant policy and infrastructure reforms that enhance the UK's competitiveness, allowing it to attract a greater share of global hyperscaler capital expenditure and scale capacity to 11 GW.[105] Despite efforts to increase supply, a significant compute gap remains, posing challenges that need to be addressed.

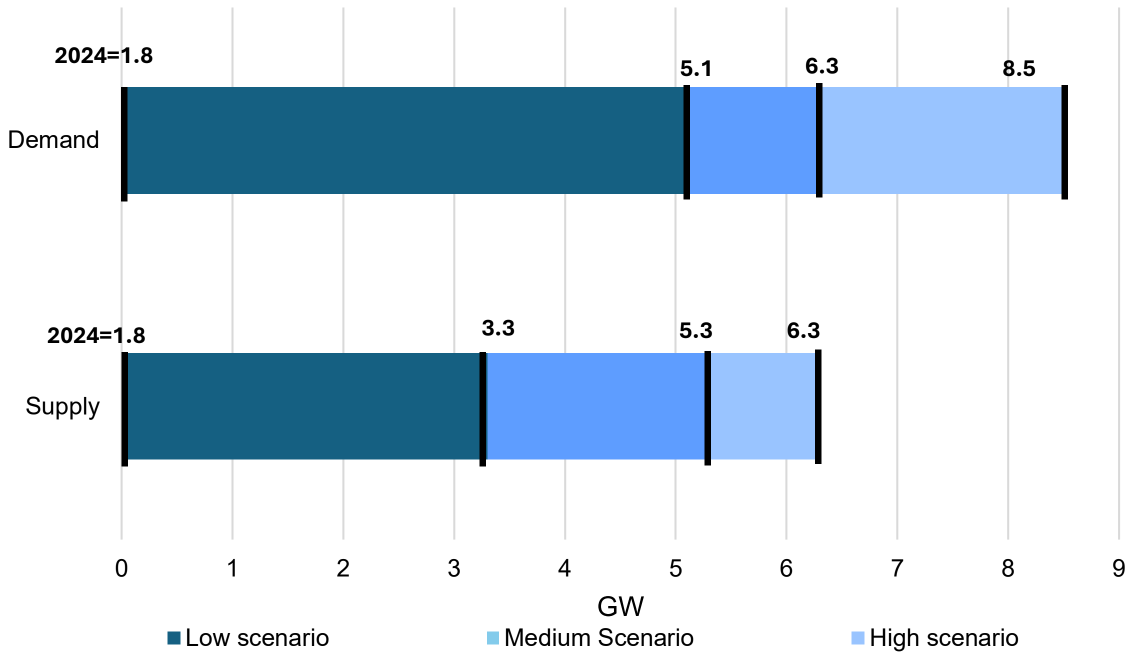
Figure 5: UK compute supply in the Medium Scenario, 2025-2035 (GW)



Compute Gap

Without intervention, analysis indicates that the UK could face up to a 5GW 'Compute Gap' within the next five years.[106] The UK's compute supply could nearly double to 3.3GW in 2030. This is based on the announced project pipeline and growth in data centre investment to satisfy growth in Gen-AI workloads.[107] This means the UK's market share of global data centre capacity could shrink from 3% to 1.5% from 2024-30. Coupled with UK compute demand expected to grow faster to 5.1-8.5GW that won't be met by UK supply, leads to a large compute gap by 2030. This excludes significant frontier model training growth in the UK, which would be added to this demand.

Figure 6: Compute demand and supply across low, medium and high scenarios by 2030 (GW)



By 2030, limited domestic compute capacity could seriously affect the UK economy. It may prevent the country from meeting sovereign needs like NHS or national security use cases that require secure, local data processing. Latency-sensitive technologies, such as autonomous vehicles and high-frequency trading, may be held back because they rely on real-time compute that can't be outsourced. A growing dependence on foreign infrastructure could increase exposure to global supply chain risks, potentially disrupting access and raising costs. It would also weaken the UK's ambition to be an "AI maker, not an AI taker," by limiting its ability to develop and deploy advanced AI models.

The UK public sector is facing a significant projected increase in AI compute demand, with forecasts suggesting up to a 22-fold rise between 2024 and 2030 in a high-demand scenario. This is equivalent to an added 420 exaFLOPs.[108] This surge underpins the AIRR programme's x20 target, which is based on scaling current capacity (Isambard-AI and Dawn supercomputers) to meet future needs. However, even with this planned expansion, a substantial supply gap stays. In the high-demand scenario, public sector AI demand alone would nearly match the entire proposed AIRR capacity, while in medium and low scenarios, the gap persists but is less severe. The AIRR investment is therefore positioned as a critical supply-side intervention to close this gap, particularly for public sector AI use cases. Yet, given the scale of demand from other sectors like education and SMEs, the programme will not be able to meet all needs. This highlights the necessity of a prioritised access model to ensure that the most critical public sector applications are supported.

The environmental impacts of compute must also be considered alongside supply challenges.

Environmental Impact of Compute

AI workloads are driving a sharp rise in electricity demand.[109] The escalation in energy demand is being driven by rising adoption of Gen-AI.[110] This increases demand for power-intensive processors, such as GPU and specialised AI chips, used to train and deploy advanced models. Gemini Ultra reportedly needed 35 MW to train, and future frontier models could demand gigawatt-scale power by 2029.[111] When including inference demand, energy requirements are even larger. Epoch estimates that a long input of 10k tokens (like a short paper or long magazine article) is estimated to cost per query around 2.5 watt-hours.[112] However, offsetting the large energy demand is an improvement in chip efficiency. For example, AMD's Radeon RX 6000 GPUs are 54% more energy efficient than the earlier generation of chips.[113]

Some projections show that by 2030, global energy consumption by AI-optimised data centres could surpass the total electricity currently used by Japan each year.[114] Specifically, these centres are expected to consume more than 945 TWh globally by 2030.[115]

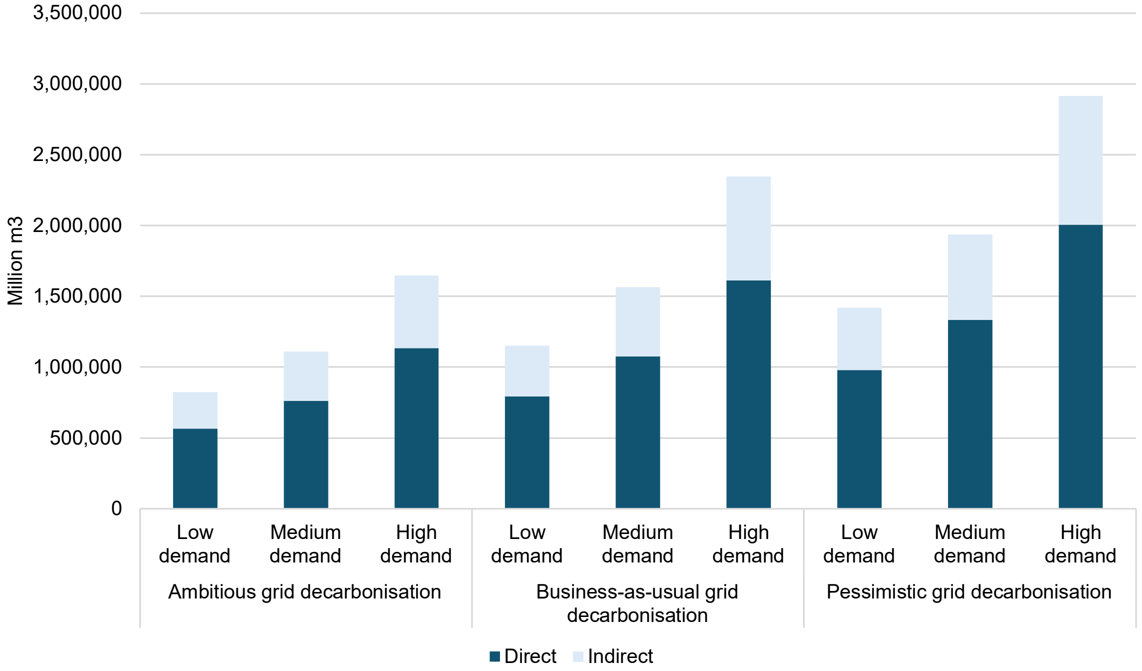
The original content of this section referring to DSIT's environmental impacts model has been temporarily removed. We keep analysis under routine review, and are updating this modelling to ensure it reflects the most up to date assumptions and analysis. This section will be updated by summer 2026.

Figure 7: GHG emissions under three decarbonisation scenarios, split by high, medium and low AI deployment (2025-2035)

The original content of this section referring to DSIT's environmental impacts model has been temporarily removed. We keep analysis under routine review, and are updating this modelling to ensure it reflects the most up to date assumptions and analysis. This section will be updated by summer 2026.

AI data centres also place heavy demands on water, primarily for cooling. Up to 40% of a data centre's total energy consumption can be for cooling. Factors like increased energy requirements for GPUs, higher density server arrangements, and AI processes running continuously makes it more of a challenge for AI data centres specifically.[120] This has spurred investment into new cooling solutions, such as using liquid immersion cooling where servers are submerged in coolant and can reduce energy use by 30% compared to traditional air cooling.[121] However, a recent study found that training GPT-3 in Microsoft US data centres (which are more efficient) consumed around 700,000 litres of water.[122] This means that global AI-related water withdrawal is expected to reach 4.2–6.6 billion m³ by 2027,[123] comparable to the annual water use of several countries.[124] Under a medium deployment scenario, cumulative water use could reach 1.6 trillion cubic metres by 2035, with a range of 1.1–1.9 trillion cubic metres depending on decarbonisation assumptions (Figure 8).

Figure 8: Water use under three decarbonisation scenarios, split by high, medium and low AI deployment (2025-2035)



Conclusion

The UK's AI compute needs are diverse and growing, particularly in latency-sensitive sectors like banking, retail, healthcare, and public services. While the current data centre supply meets sovereign demand, a potential compute gap could hinder economic growth, affect national security, and increase reliance on foreign resources.

Without intervention, the UK's compute supply could reach only 3.3GW, while demand could exceed 8.5GW. This gap risks undermining national security, economic growth, and the UK's ambition to lead in AI. Public sector demand, while smaller than private, has unique requirements (security, latency) that make domestic provision non-optional; planning must account for these needs to avoid mission-critical shortfalls (for example, in healthcare AI).

These findings reinforce the rationale for the actions set out in the UK Compute Roadmap, particularly the expansion of public compute infrastructure through the AIRR, the development of AI Growth Zones, and the prioritisation of sovereign and sustainable compute solutions. Together, these initiatives aim to ensure that the UK's compute ecosystem is equipped to meet future demand, support national missions, and secure long-term strategic advantage.

Annex A: AI Rapid Evidence Review Methodology

A rapid evidence review was conducted to find trends around hardware, data centre design and use cases. These were partially helped by four different Deep Research tools, Copilot, Gemini, Perplexity and EPPI Reviewer. Each AI tool was fed the same research questions, then the output was cross-referenced and merged to form the report.

- What are the latest developments in compute design? How will these affect future trends? How does this change between training and inference uses?
- How do specialised chips improve efficiency and performance of inference tasks compared to general purpose GPUs?
- To what extent are we likely to see a divergence in specialised hardware for training and inference?
- Is there evidence that certain hardware is, or will be, specialised for specific types of model training? (e.g. specialisation for specific architectures)
- What are the trends in the balance of inference vs training compute demand and how is this likely to play out in future?
- What are the trends in the balance of compute demand for narrow vs general AI models and how is this likely to play out in future?
- Do different domains (e.g. science, health, finance) require different hardware stocks? If so, why?
- What are trends in the layout of AI compute data centres? To what extent are we seeing more large, consolidated sites vs distributed sites?
- What are the recent trends and developments in distributed training techniques for AI models? How do these advancements impact the scalability and efficiency of training large AI models?
- For what use cases is distributed compute most useful?
- What do all these trends mean for future energy demands for AI compute data centres?
- What do all these trends mean for future water demands for AI compute data centres?
- What about water impacts?

Annex B: McKinsey Demand and Supply Model Methodology

DSIT commissioned McKinsey to provide a 10-year projection of future compute demand and supply. This model is crucial for understanding the potential growth in compute requirements driven by the rapid adoption of AI across various sectors. By forecasting both demand and supply, it aims to identify potential gaps and ensure that the UK can meet is compute needs, particularly for public sector AI uses.

The method involves forecasting both demand and supply based on various assumptions about AI adoption, technical progress for semiconductors and AI development, and other factors. The model uses historical data, expert interviews, and industry investment projections to inform the analysis. This analysis is intended to inform ongoing policy development and does not represent a final cross-cutting view of government. Scenario projections are subject to uncertainty and may evolve as further evidence and perspectives are incorporated.

Compute Supply

These forecasts are informed by semiconductor supply chain orders and global industry investment projections, accounting for ongoing construction delays and the UK's share of the global market given context like the UK being relatively unattractive given higher power costs.

To forecast the full potential of data centre growth, McKinsey estimated current expansions to data centres, planned and announced investments, and unannounced investments. For these first two categories, the S&P 451 Dataset was used to find announcements and when the data centres would be operational. For unannounced data centre construction, this was calculated by estimating the UK's share of global capital investment by the top five global hyperscalers, including the split by wholesale co-locators versus owned and operated buildouts. The UK's share was informed from expert interviews. Internal estimates were then applied to translate capital investments into gigawatt capacity, guided by the McKinsey Global Data Centre Model and global hyperscaler expert interviews.

The above provides a forecast of data centre construction between 2025 to 2030. To extend the forecast to 2035, the 2025-30 compound annual growth rate was extrapolated further. This was further informed through expert interviews and other limiting factors, such as the increase in UK grid power supply during the same period, specifically assessing the proportion of this power supply given to data centres using the McKinsey Global Energy Perspective.

Different supply scenarios were constructed to illustrate the effect of different growth constraints for data centres. This included the impact of connection queue delays - the time needed to power both planned and unplanned data centres. This involved evaluating the proportion of data centre announcements that may not materialise, and insights gathered from expert interviews. Additionally, an assessment of power supply expansion plans was also considered to decide the feasibility of increasing data centre capacity while ensuring adequate access for non-data centre users. This assessment relied on data sources from NESO Future Energy Scenarios.

Within the supply forecast, there is a breakdown for high-performance computing (HPC) supply. This involved first finding the current baseline of HPC capacity using the Top500 list of supercomputers. This is then combined with public press information to find new supercomputers that are either under construction or planned, along with their estimated capacities, to estimate how this might change over time.

Compute Demand

Compute demand consists of AI and traditional compute, both broken down further by public and private sector demand. AI compute demand can also be broken down further into training and inference needs, with the latter considering the needs for business-to-business (B2B) and business-to-consumer (B2C) AI inference compute demand. These were estimated separately but combined to provide an estimate of total compute demand.

To assess B2B use case inferencing for the period 2025-2035, McKinsey analysed the full potential of AI automation and value creation by industry. The analysis calculated Gen-AI spending by firms based on industry-specific adoption rates and the compute intensity of use case archetypes. This spending was then converted into implied compute demand, measured in power capacity, using an estimate of operational expenditure per watt.

For B2C use case inferencing from 2025 to 2030, McKinsey forecasted the number of daily interactions that a connected individual would have with a Large Language Model (LLM). This forecast was combined with the compute needed per interaction, which was derived from historical development trends. Compute estimates for the 2030-2035 period were extrapolated in line with global adoption trends.

In evaluating frontier training and small-to-medium model training for the period 2025-2030, McKinsey established a baseline using the known training costs of organisations such as OpenAI. Comparisons were drawn to similar industry players and their market positions. Historical growth rates in training costs were then used to forecast future costs, which were converted into gigawatts using OPEX cost per watt. These calculations were conducted for the pre-2030 period, while figures for 2031 to 2035 were extrapolated in line with global market expectations.

For other accelerated compute workloads between 2025 and 2030, McKinsey relied on global demand data for AI-ready servers, informed by Nvidia's chip shipment forecasts. The part of compute demand not addressed by Gen-AI was classified as 'other accelerated' compute workloads. Meanwhile, for traditional compute growth within data centres, this was based on global CPU shipment data. Calculations for 2025-2030 period were performed using a bottom-up methodology, with projections for 2030-2035 aligned with global market estimates.

In analysing public sector AI growth from 2025 to 2030, McKinsey applied a method like that used for B2B Gen-AI, incorporating general efficiency gains from automation and research and development (R&D) efforts. This analysis included sectors such as public services, healthcare, and defence. Projections for the period beyond 2030 were extrapolated based on global market trends.

Compute Gap

Comparing the compute supply and demand forecasts can indicate whether there will be a gap in the UK's capacity. Overall, there are nine combinations of forecast scenarios, though McKinsey expects the UK to be in a base-supply and medium-demand scenario. This suggests that total compute supply would rise from 1.8GW in 2024 to 3.3GW in 2030, while total compute demand would rise from 1.8GW in 2024 to 6.3GW in 2030, resulting in a 3GW in 2030 compute gap. For AI specific demand, the compute gap is slightly smaller at 2.7GW in 2030.

The compute demand and supply forecasts could be influenced by several factors that may alter their trajectory. A significant improvement in future model performance, surpassing benchmarks such as DeepSeek in 2025, might allow Gen-AI demand to be met with reduced compute requirements, resulting in a smaller gap. However, this improvement could indirectly increase Gen-AI adoption due to enhanced capabilities and reliability. Additionally, an increase in UK data centre investments beyond expectations, driven by improved economics or strategic/geopolitical motivations, could also reduce the gap. Conversely, if Gen-AI fails to scale whether globally or locally within the UK, potentially due to regulatory challenges, economic shortcomings, or supply chain obstacles—the compute gap could lessen as expected benefits and investment pipelines fail to materialise. These factors collectively highlight the dynamic and multifaceted nature of the compute gap's development.

Annex C: AI Environmental Impacts Model

The original content of this section referring to DSIT's environmental impacts model has been temporarily removed. We keep analysis under routine review, and are updating this modelling to ensure it reflects the most up to date assumptions and analysis. This section will be updated by summer 2026.

[1] https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use

[3] https://www.imf.org/en/Publications/WEO/Issues/2024/04/16/world-economic-outlook-april-2024

[4]https://www.ons.gov.uk/economy/economicoutputandproductivity/output/datasets/businessinsightsandimpactontheukeconomys Based on the assumption that a business is using at least one AI technology.

[5] https://www.ons.gov.uk/economy/economicoutputandproductivity/output/datasets/businessinsightsandimpactontheukeconomy Based on the assumption that a business is using at least one AI technology.
[6] https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey-wave-4/public-attitudes-to-data-and-ai-tracker-survey-wave-4-report
[7] Soon to be published AI Sector Study 2024
[8] Soon to be published AI Sector Study 2024
[9] https://epoch.ai/data/large-scale-ai-models, accessed on July 2025
[10] https://epoch.ai/data-insights/dataset-size-trend, accessed on July 2025
[11] An OpenAI analysis showed the required compute doubled every 3.4 months from 2012 to 2018. https://www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before/
[12] https://epoch.ai/data/machine-learning-hardware, accessed on July 2025
[13] https://epoch.ai/data/machine-learning-hardware, accessed on July 2025
[14] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
[15] https://epoch.ai/trends#investment
[16] https://epoch.ai/gate
[17] https://www.databricks.com/blog/llm-inference-performance-engineering-best-practices
[18] https://epoch.ai/data/notable-ai-models
[19] https://developer.nvidia.com/blog/floating-point-8-an-introduction-to-efficient-lower-precision-ai-training/
[20] https://thepycoach.com/ai-at-bargain-prices-but-whos-really-paying-the-bill-a1b92f1d7b1e
[21] Internal McKinsey Supply and Demand Analysis. Source: 'Situational Awareness' – Leopold Ashenbrenner, Epoch.ai, Nvidia specification sheets, OpenAI, Chip shipment data
[22] 'Situational Awareness' – Leopold Ashenbrenner, Epoch.ai, Nvidia specification sheets, OpenAI, Chip shipment data
[23] https://openai.com/index/ai-and-efficiency/
[24] https://www.xcubelabs.com/blog/real-time-inference-and-low-latency-models/
[25] https://www.tech.gov.sg/products-and-services/for-government-agencies/productivity-and-marketing/pair/
[26] McKinsey Global Data centre model, 451 data centre supply 2024, McKinsey Global Institute, expert interviews, ONS regional GDP data, expert interviews
[27] McKinsey Global Data centre model, 451 data centre supply 2024, McKinsey Global Institute, expert interviews, ONS regional GDP data, expert interviews: McKinsey Global Data centre model, 451 data centre supply 2024, McKinsey Global Institute, expert interviews, ONS regional GDP data, expert interviews
[28] https://blog.congenica.com/impact-of-the-100000-genomes-pilot-publication
[30] https://www.nostos-genomics.com/solutions
https://www.nostos-genomics.com/news/next-gen-diagnostics-ais-role-in-interpreting-genetic-variants
[31] https://www.astrazeneca.com/media-centre/medical-releases/astrazeneca-new-ai-technology-milton-predicts-more-than-1000-diseases-before-diagnosis.html
[33] https://nvidia.bsi.uk.com/products/nvidia-h100-pcie-gpu
[34] https://moorfieldsbrc.nihr.ac.uk/world-first-artificial-intelligence-foundation-model-for-eye-care-to-supercharge-global-efforts-to-prevent-blindness/
[35] https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/
[36] https://www.nature.com/articles/s41586-023-06555-x#Sec9
[37] https://www.nature.com/articles/s41467-024-55301-y
[38] https://www.chemify.io/
[39] https://www.imperial.nhs.uk/about-us/news/ai-model-can-predict-health-risks
[40] https://www.imperial.nhs.uk/about-us/news/virtual-biopsy-uses-ai-to-help-doctors-assess-lung-cancer
[41] https://cepa.org/article/france-pursues-an-ai-third-way/
[42] https://www.pmindia.gov.in/en/news_updates/cabinet-approves-ambitious-indiaai-mission-to-strengthen-the-ai-innovation-ecosystem/
[43] https://www.mckinsey.com/featured-insights/sustainable-inclusive-growth/charts/gen-ais-broad-reach
[44] https://www.ons.gov.uk/economy/economicoutputandproductivity/output/datasets/businessinsightsandimpactontheukeconomys

[45] Source: McKinsey Global Data centre model
[46] https://www.datacenterdynamics.com/en/news/new-nvidia-blackwell-accelerator-will-consume-1kw-of-power/. Assumes 100% of power is used for GPU operation and 1,000 watts is needed to power 1 GPU.
[47] "The economic potential of generative AI: The next productivity frontier" report & "AI power: Expanding data centre capacity to meet growing demand" report McKinsey Global Institute, McKinsey data centre benchmarks, "Situational Awareness" – Leopold Aschenbrenner, Epoch.ai, IHS Markit
[48] McKinsey Global Data centre model
[49] McKinsey Global Data centre model
[50] Internal McKinsey Supply and Demand Analysis. Source: IHS Markit, OpenAI, Nvidia specification sheets, McKinsey Global Institute, McKinsey data centre benchmarks, Epoch.ai, 'Situational Awareness' – Leopold Ashenbrenner, ICTCube, Statista, Expert interviews
[51] Internal McKinsey Supply and Demand Analysis. Source: IHS Markit, OpenAI, Nvidia specification sheets, McKinsey Global Institute, McKinsey data centre benchmarks, Epoch.ai, 'Situational Awareness' – Leopold Ashenbrenner, ICTCube, Statista, Expert interviews
[52] https://www.datacenterdynamics.com/en/news/new-nvidia-blackwell-accelerator-will-consume-1kw-of-power/ Calculation based on an energy efficiency rating of 68.83 GFLOP/s per watt
[53] https://top500.org/lists/top500/list/2025/06/
[54] https://awavesemi.com/to-gpu-or-not-gpu/
[55] https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/
[56] https://www.trgdatacenters.com/resource/gpu-vs-cpu-for-ai/
[57] https://www.thebusinessresearchcompany.com/report/ai-infrastructure-global-market-report
[58] https://www.computer.org/publications/tech-news/trends/ai-infrastructure-innovation/
[59] https://www.interglobixmagazine.com/the-evolution-of-hardware-for-ai/
[60] https://opentools.ai/news/ai-chip-wars-nvidia-faces-fierce-competition-from-inference-innovators
https://cloud.google.com/blog/products/gcp/quantifying-the-performance-of-the-tpu-our-first-machine-learning-chip
[61] https://www.ibm.com/think/insights/trends-hardware-gets-ai-updates-2024
[62] https://gorzim.com/blog/ai-hardware-trends-2025-to-2028/
[63] https://gorzim.com/blog/ai-hardware-trends-2025-to-2028/
[64] https://my.idc.com/getdoc.jsp?containerId=prUS53261225
[65] https://ukhsa.blog.gov.uk/2025/03/14/how-we-are-pioneering-artificial-intelligence-applications-in-public-health/
[66] https://www.mdpi.com/1999-4893/15/11/419
https://community.intel.com/t5/Blogs/Products-and-Solutions/FPGA/Using-oneAPI-for-Low-Latency-AI-Inference-with-Altera-FPGAs/post/1596452
[68] https://www.espjournals.org/IJACT/2023/Volume1-Issue1/IJACT-V1I1P107.pdf
[69] https://www.geeksforgeeks.org/hardware-requirements-for-artificial-intelligence/
[70] https://www.physics.ox.ac.uk/news/massive-dataset-accelerate-ai-research-astronomy
[71] https://www.physics.ox.ac.uk/news/massive-dataset-accelerate-ai-research-astronomy
https://github.com/MultimodalUniverse/MultimodalUniverse
[72] https://www.oecd.org/en/publications/a-blueprint-for-building-national-compute-capacity-for-artificial-intelligence_876367e3-en.html
[73] IDC Chip Shipment Data, S&P 451 Dataset
[75] https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ai-performance-ucs-c240-wp.html
https://docs.mlcommons.org/inference/
[76] https://cognitiveworld.com/articles/2025/4/1/rethinking-data-centres-in-the-era-of-ai-agents
[77] https://datacentrereview.com/2024/12/the-changing-focus-of-data-centres-in-the-ai-era/
[78] https://techedgeai.com/ai-infrastructure-in-2025-and-beyond-cloud-edge-and-decentralized-ai/
[79] https://www.cudos.org/blog/ai-inference-the-deepseek-disruption-and-the-future-of-compute
[80] https://www.digitalrealty.ie/resources/articles/decentralised-data-in-data-centres
[81] https://www.datacentreknowledge.com/hyperscalers/hyperscalers-in-2024-where-next-for-the-world-s-biggest-data-centre-operators-
[82] https://www.centreofyourdigitalworld.org/2025-impact-study
[83] https://www.digitalconstructiontoday.com/ai-data-centre-boom-threatened-by-power-availability-and-limited-supply-chain-capacity/2326/

[84] https://www.forbes.com/sites/emilsayegh/2025/04/21/cybersecuritys-red-alert-moment-hyperscalers-are-consolidating-power/
[85] https://incountry.com/blog/cloud-data-sovereignty-concerns-requirements-and-solutions/
[87] https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/how-high-tech-suppliers-are-responding-to-the-hyperscaler-opportunity
[89] Internal McKinsey Supply and Demand Analysis. Source: S&P 451 Dataset; McKinsey Global Data Centre Model, Expert Interviews, McKinsey AI Chip Model
[90] Internal McKinsey Supply and Demand Analysis. Source: S&P 451 Dataset; McKinsey Global Data Centre Model, Expert Interviews, McKinsey AI Chip Model
[91] Internal McKinsey Supply and Demand Analysis: S&P 451 Dataset; McKinsey Global Data Centre Model, Expert Interviews, McKinsey AI Chip Model
[92] Internal McKinsey Supply and Demand Analysis. Source: BofA Q4 Cloud CAPEX Preview, Expert interviews, McKinsey DC CAPEX model
[93] Internal McKinsey Supply and Demand Analysis. Source: BofA Q4 Cloud CAPEX Preview, Expert interviews, McKinsey DC CAPEX model
[94] Internal McKinsey Supply and Demand Analysis. Source: Customer Survey N=30, Pitchbook deal scan of Private Equity and Venture Capital deals from 2023-24, New Investment in Europe's Data Centre Markets to Hit New Heights in 2025 (CBRE), Expert interviews
[95] International Construction Cost Index, 2024
[96] Internal McKinsey Supply and Demand Analysis. Source: Customer Survey N=30, Pitchbook deal scan of Private Equity and Venture Capital deals from 2023-24, New Investment in Europe's Data Centre Markets to Hit New Heights in 2025 (CBRE), Expert interviews
[97] https://www.forbes.com/sites/katharinabuchholz/2024/08/23/the-extreme-cost-of-training-ai-models/
[98] Internal McKinsey Supply and Demand Analysis. Source: EPOCH AI, McKinsey CAPEX model, press search. It is assumed that ChatGPT-4 cost £32m to train and was completed in an area like North Virginia, US and was completed in one year. Scaling to other countries was done by taking the proportional difference in the implied OPEX cost per Watt between countries. The annual OPEX cost to operate a 1GW data centre in each country was then added.
[99] Internal McKinsey Supply and Demand Analysis. Source: EPOCH AI, McKinsey CAPEX model, press search
[100] Internal McKinsey Supply and Demand Analysis. Source: EPOCH AI, McKinsey CAPEX model, press search
[101] It is assumed that ChatGPT-4 cost £32m to train and was completed in an area like North Virginia, US and was completed in one year. Scaling to other countries was done by taking the proportional difference in the implied OPEX cost per Watt between countries. The annual OPEX cost to operate a 1GW data centre in each country was then added.
[102] Source: McKinsey Global Data centre model
[103] Source: McKinsey Global Data centre model
[104] Source: McKinsey Global Data centre model
[105]: McKinsey Global Data centre model
[106] Internal McKinsey Demand and Supply Analysis. Source: McKinsey Global Data Centre Model
[107] Internal McKinsey Demand and Supply Analysis. Source: S&P 451 DC Knowledge Base, BofA Q4 Cloud CAPEX Preview, Expert interviews, McKinsey DC CAPEX model, Press research
[108] Internal McKinsey Demand and Supply Analysis. Source: McKinsey Global Data Centre Model
[109] https://www.iea.org/reports/energy-and-ai
[110] https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2025/genai-power-consumption-creates-need-for-more-sustainable-data-centers.html
[111] https://arxiv.org/html/2405.21015v1
[112] https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use
[113] whaleflux.com/blog/gpu-compare-chart-mastery-from-spec-sheets-to-ai-cluster-efficiency-optimization/
[114] https://datacentremagazine.com/technology-and-ai/iea-asks-how-is-ai-reshaping-global-energy-demand
[115] https://www.iea.org/news/ai-is-set-to-drive-surging-electricity-demand-from-data-centres-while-offering-the-potential-to-transform-how-the-energy-sector-works
[120] https://airsysnorthamerica.com/understanding-the-design-cooling-of-ai-data-centers/
[121] https://www.datacenters.com/news/why-liquid-cooling-is-becoming-the-new-standard-in-hyperscale-facilities

[122] https://arxiv.org/pdf/2304.03271
[123] https://oecd.ai/en/wonk/how-much-water-does-ai-consume
[124] https://cee.illinois.edu/news/AIs-Challenging-Waters