technopolis group

October 2026

Quantifying UK Compute User Demand

Final Report

www.technopolis-group.com

October 2026

Quantifying UK Compute User Demand

Final Report

Cristina Rosemberg, Aphra Murray, Alexander Holmes, Shrishti Kajaria, Todd Cook, Sophie Young

Table of Contents

Executive Summary 1

Glossary of Key Terms 3

1 Introduction 5

2 Approach to Estimating Demand for Compute 8

3 Current Demand for High Performance Compute for AI 16

4 Estimating future demand for compute 29

5 Sovereign Demand for Compute 37

Tables

Table 1 Survey response rates across user groups 10

Table 2 Estimated HPC users across sectors 12

Table 3 Current UK compute demand estimates under alternative academic user population estimates 17

Table 4 Summary of current UK power demands for HPC in the UK 18

Table 5 Growth scenarios from the HPC user survey 31

Table 6 Estimated future power requirements 33

Table 7 Drivers of future demand for AI compute 47

Figures

Figure 1 Schematic overview of the study 8

Figure 2 Average annual demand for compute by sector (in H100e GPUh) 17

Figure 3 Schematic overview of run-time adjusted power demands 19

Figure 4 Comparison of estimates from this study with wider estimates of power demand 20

Figure 5 Types of data used by sectors 22

Figure 6 Shares of research publications using AI or CM by discipline 23

Figure 7 Share of total publications mentioning AI use in 2025 24

Figure 8 (Average) models by which respondents access compute (n=126) 25

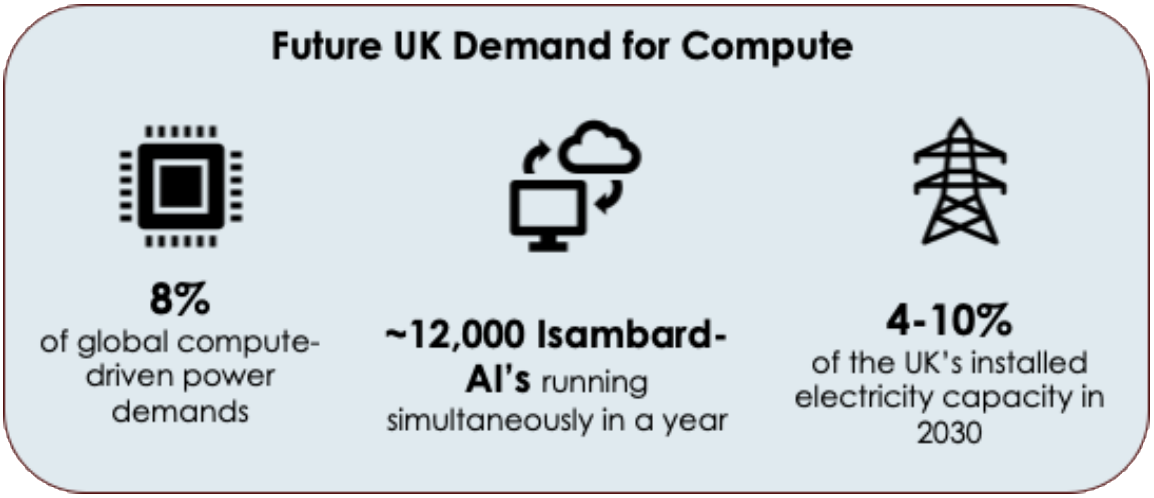
Figure 9 respondents' reasons for specific vendor requirements 27
Figure 10 Allocation of compute time by sector 28
Figure 11 Projected future demand (in GPUh) 32
Figure 12 Estimated future power demands for compute from only GPU cards 34
Figure 13 Projected demand for compute (in RFLOP) 36
Figure 14: Sovereign future demand for compute 39
Figure 15 Survey respondents' preferred access model (n=126) 40
Figure 16 Respondents' primary research application area 43
Figure 17 User latency requirements (n=134) 45
Figure 18 Drivers for latency requirements (n=121) 46
Figure 19 Top 10 fastest growing disciplines that mention use of AI (2021-2025) 49

Executive Summary

This study, commissioned by the Department for Business, Innovation, Science and Trade (BIST) -Formally DSIT- provides a systematic, bottom-up quantification of UK compute demand for AI research across academia, industry, and the public sector. It draws on a survey of UK HPC users, a bibliometric analysis of UK research publications, and demand modelling to provide current estimates and forward projections to 2031.

Current demand for compute is estimated at 33-40 billion H100-equivalent GPU hours per year – equivalent to approximately 750 Isambard AI's running simultaneously across a full year. In power terms, this translates to 5.1 – 6.1 GW of electrical demand, or roughly 12% of UK peak electricity demand and 5% of current UK installed electricity generation capacity. Although this study has not sought to address supply (i.e. where the demand is served) it is likely that a significant share of this current demand is served by overseas datacentres. Industrial users have the highest demand (in terms of average GPU hours per user), consistent with global trends showing the private sector's share of AI compute growing from ~40% in 2018 to ~80% by 2025.

Demand is projected to grow sharply. Under the most likely scenario (continued high growth in demand), UK compute demand is expected to reach up to 1,142 billion GPU hours by 2031. In power terms, this implies 11-21 GW by 2030 (taking into consideration both the power draw from the GPU cards and the full hardware system, and assuming no change in the efficiency of systems). These findings are supported by bibliometric data, which shows that the fastest-growing disciplines for AI adoption and use are recording average growth rates of over 110% per year, suggesting respondents' growth expectations may, if anything, be conservative.



A substantial share of this demand requires sovereign provision. Using requirements for Trusted Research Environments (TREs) (the most reliable measure of demand that is formally constrained to UK-hosted infrastructure) sovereign demand is estimated at approximately 76% of total projected demand, equivalent to 444 billion GPUh and 6.2 GW of power by 2030. This is driven largely by industry and the public sector, where users reported higher average TRE requirements (83% and 82% of users, respectively). This level of sovereign demand is unlikely to be met through national facilities alone; it will instead require a broader ecosystem of sovereign-compliant infrastructure, including UK-hosted commercial cloud.

Latency may also be an additional driver of domestic provision. Half of survey respondents (50%) describe low-latency compute as either critical or very important to their work, driven by the needs of interactive AI applications and simulation modelling — workloads that are particularly sensitive to geographic distance from infrastructure. These requirements may also drive the need for increased UK-based provision of compute.

Three further findings warrant additional attention from a policy-making perspective. First, AI is now a cross-disciplinary tool, not a specialist one: the share of UK research publications using AI rose from 9.5% in 2021 to 14.4% in 2025 (a 4.9 p.p. increase), with growth across every field. Future compute demand will likely be distributed, not concentrated in a small number of domains. Second, vendor dependency is high: 69% of users describe their workloads as fully or mostly tied to their current GPU manufacturer, driven by hardware capability constraints rather than software lock-in. Third, respondents across all demand scenarios identify the same core growth drivers (model scaling, inference deployment, and infrastructure availability) suggesting the direction of travel is not in dispute, only the pace.

Glossary of Key Terms

Term / Acronym	Definition
AI	Artificial Intelligence
AIRR	Artificial Intelligence Research Resource The UK's national AI compute infrastructure–a publicly funded, free-at-point-of-use network of GPU supercomputers (currently Isambard-AI and Dawn) intended to give researchers and companies the hardware needed to do serious AI work without paying commercial cloud rates, while building toward a 20x capacity expansion by 2030.
Dawn	Dawn is the Cambridge half of the UK's AIRR–an Intel/Dell-built AI supercomputer that, paired with Isambard-AI, gives UK researchers free-at-point-of-use access to frontier-scale AI compute.
Compute	The measurable computational work performed by a system, often quantified in FLOPs or, in the context of AI training, in total FLOPs used across a training run. In hardware terms, compute refers to the physical infrastructure that performs this work: GPUs, TPUs, specialised AI accelerators, and the data centre facilities that house and interconnect them. More holistically, compute can be understood as a full technology stack, encompassing hardware, networking, storage, software frameworks, and the operational capabilities required to deploy and manage these systems at scale.
FLOPs	Floating Point Operations A count of operations and a measure of the total amount of work that can be done (quantity)
FLOP/s	Floating Point Operations per Second This is a rate of computation, measuring how fast a processor can perform floating point maths (performance).
GPU	Graphics Processing Unit A specialised electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device.
GPUh	Graphics Processing Unit Hours Measures the active usage time of a GPU for tasks like AI training, rendering, or data processing, calculated as the number of GPUs multiplied by hours run.
GW	Gigawatt A unit of power equal to one billion watts, or 1,000 megawatts.

H100	NVIDIA H100 Tensor Core GPU is a premier, Hopper-architecture data centre GPU optimised for AI training, inference, and HPC.
HPC	High-Performance Compute Refers to the use of aggregated computing power–typically clusters of many tightly interconnected processors–to solve problems that are too large, complex, or time-sensitive for ordinary computers.
Isambard-AI	The flagship AIRR system–currently the UK's most powerful public compute facility. Named after Isambard Kingdom Brunel, it's operated by the University of Bristol at the National Composites Centre.
Latency	The time it takes for data to travel from its source to its destination or for a system to process a request.
LLMs	Large Language Models Type of AI designed to understand, generate, and process human-like text by analysing vast datasets using deep learning, specifically neural networks with billions of parameters.
Neural Networks	A machine learning model designed to simulate the way the human brain analyses and processes information
REA	Rapid Evidence Assessment
RFLOPs	Ronna FLOPs, meaning 10^{27} floating-point operations. A Ronna FLOP is the natural unit for talking about compute at scales beyond anything currently built.
Sovereign AI	The principle that a country has the capacity to independently determine the direction, governance and use of AI in its jurisdiction, free from undue independence or interference from external actors.
SME	Small- and Medium-sized Enterprise A category of business defined by size, typically measured by headcount. • Micro: fewer than 10 employees • Small: fewer than 50 employees • Medium: fewer than 250 employees
TPU	Tensor Processing Unit An application-specific circuit developed by Google specifically for machine learning and neural network tasks, designed to accelerate AI workloads.
TREs	Trusted Research Environments Highly secure, controlled digital spaces used to analyse sensitive data.

1 Introduction

1.1 High Performance Compute For AI

High-performance compute (HPC) has underpinned scientific research for decades. From climate modelling and genomics research to drug discovery and materials science, the ability to process vast quantities of data at speed has long been a prerequisite for frontier research. Today, a new and rapidly growing workload has emerged alongside these traditional scientific applications: artificial intelligence (AI). The rise of large language models, neural networks, and other machine learning architectures has transformed compute from a specialist scientific tool into a foundational input for economic competitiveness and national capability.

'Compute' is a term used in several distinct but related senses in policy and technical discourse. At its broadest, compute refers to the measurable computational work performed by a system, often quantified in floating point operations per second (FLOP/s) or, in the context of AI training, in total FLOPs used across a training run. In hardware terms, compute refers to the physical infrastructure that performs this work: graphics processing units (GPUs), tensor processing units (TPUs), specialised AI accelerators, and the data centre facilities that house and interconnect them. More holistically, compute can be understood as a full technology stack, encompassing hardware, networking, storage, software frameworks, and the operational capabilities required to deploy and manage these systems at scale. For the purposes of this report, HPC for AI refers to this full ecosystem of hardware, software, and infrastructure used specifically to train, fine-tune, and deploy AI models, with a particular focus on the GPU clusters that drive the most computationally intensive AI workloads.

Access to this infrastructure is increasingly understood as a strategic national asset. The cost and capability of AI systems are, to a significant degree, a function of the compute available to develop and run them. Nations and organisations with access to large, well-provisioned compute clusters are better positioned to develop and deploy frontier AI, attract AI research talent, and capture the productivity gains that AI is expected to generate across sectors including healthcare, financial services, defence, and the public sector. Conversely, constrained compute access risks locking research institutions and domestic industry out of the most capable AI systems, increasing dependence on foreign-owned cloud infrastructure, and limiting the UK's influence over the trajectory of this critical technology.

The strategic importance of compute has been recognised at the highest levels of UK government. The National AI Strategy (2021) established a ten-year ambition for the UK to become a global AI superpower, with HPC infrastructure identified as a key enabler.[1] The independent Future of Compute Review (2023) provided a comprehensive assessment of the UK's compute landscape and made recommendations on expanding national capacity and improving access.[2] Most recently, the AI Opportunities Action Plan (January 2025) committed the government to expanding the UK's sovereign compute capacity at least twentyfold by

[1] National AI Strategy, HMG (2022), Available at: https://www.gov.uk/government/publications/national-ai-strategy

[2] Independent Review of the Future of Compute, DSIT (2022), Available at: https://www.gov.uk/government/publications/future-of-compute-review

2030, backed by up to £2 billion in public investment through to 2030.[3] This commitment has since been operationalised through the publication of the UK Compute Roadmap (July 2025), a ten-point plan that sets out how government will build, allocate, and govern public compute infrastructure.[4] Alongside these strategies, the establishment of AI Growth Zones (dedicated sites designed to fast-track the development of AI-enabled data centres) signals a shift towards treating compute infrastructure with the same urgency as other critical national infrastructure (CNI).[5] In fact, in 2024, data centres were designated as CNI, allowing the government to support the sector in the event of critical incidents and therefore attempting to mitigate the impacts on the economy.[6] Despite this policy momentum, a fundamental question remains underexplored: how much compute does the UK actually need?

1.2 This study

The Department for Business, Innovation, Science, and Trade (BIST) commissioned Technopolis in partnership with London Economics (and with expert advice from Professor Wes Armour) to quantify UK demand for high-performance compute for AI. It provides a bottom-up quantification of UK demand, covering academia, industry and the public sector.

The high-level objectives were to:

- **Quantify current UK demand for HPC for AI**. Additionally, the study sought to provide an overview of how users are currently accessing provision;
- **Estimate and model future demand for HPC for AI**. There was also a particular interest in understanding the extent to which this demand would need to be sovereign (i.e. services that are bound geographically to UK infrastructure).

In addition to the two main objectives, there was also a desire to better understand the drivers of future demand. The term 'AI' is used colloquially (and in research and policy contexts) to cover a wide range of technologies and activities. In practice, it can refer to classical machine learning and statistical modelling through to large-scale deep learning and generative AI systems. Given the breadth of activity that falls under this label, and the significant public and policy interest in AI, it is important to be clear about the scope of this study.

[3] AI Opportunities Action Plan, DSIT (2025), Available at: https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

[4] UK Compute Roadmap, DSIT (2026), Available at: https://www.gov.uk/government/publications/uk-compute-roadmap/uk-compute-roadmap

[5] Policy paper: Delivering AI Growth Zones, DSIT (2025), Available at: https://www.gov.uk/government/publications/delivering-ai-growth-zones/delivering-ai-growth-zones

[6] Data centres to be given massive boost and protections from cyber criminals and IT blackouts, DSIT (2024), Available at: https://www.gov.uk/government/news/data-centres-to-be-given-massive-boost-and-protections-from-cyber-criminals-and-it-blackouts

This study is not about frontier AI model training. The training of the world's most powerful AI systems is conducted by a small number of large technology companies (largely based in the United States, China and South Korea) and is not captured in the study.[7]

What this study *has done* is construct bottom-up estimates of compute demand from the UK research community (including industry) – covering fine-tuning, inference, domain-specific model development, simulation and data analysis. This is, in many ways, the compute that underpins UK science, supports domestic AI adoption, and drives the case for sovereign infrastructure. These estimates are built from evidence of actual and anticipated usage across UK institutions and sectors, and are intended to characterise the scale of that demand. It does not make any assumptions or recommendations about how that demand should or will be supplied, or in what form sovereign or shared infrastructure might be configured.

1.3 This report

The study modelled compute demand across a range of scenarios, reflecting different assumptions about adoption rates, use cases and growth trajectories. For the purposes of this report, a single scenario (and associated estimates for demand) is presented, but this represents one point in a broader set of modelled outcomes. An accompanying interactive dashboard allows readers to explore the full range of scenarios and underlying assumptions.

The remainder of this report is structured as follows:

- **Section 3** – sets out a high-level methodological approach, including details on the survey approach
- **Section 4 –** presents our estimates of current demand and insights into how AI for HPC is currently being used and accessed
- **Section 5 –** presents our estimates of future demand, as well as the drivers of that demand

In a separate technical appendix, we provide a detailed methodological note. This includes a methods assessment for the variety of approaches that were considered for the study, and the methods that we eventually arrived at, in consultation with BIST, as well as further details on the data cleaning processes and bibliometric analysis that supported this study.

[7] Most large-scale models are developed by US companies (2024), Epoch AI: Available at: https://epoch.ai/data-insights/large-scale-models-by-country

2 Approach to Estimating Demand for Compute

2.1 Introduction

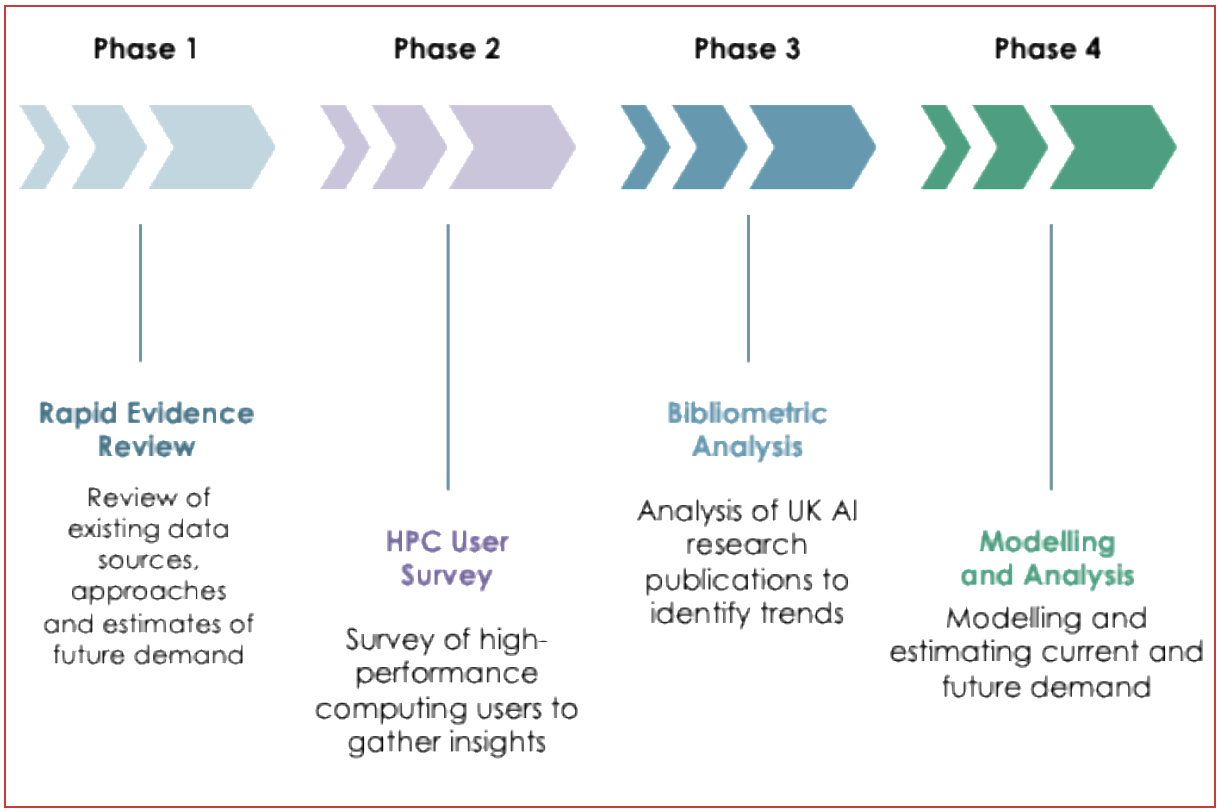
This section sets out a high-level view of the methodological approach used in this study to quantify current and future UK demand for AI compute. Estimating compute demand is inherently challenging: the landscape is technically complex, fast-moving, and distributed across a heterogeneous population of users spanning academia, industry, and the public sector. No single source of data is sufficient to produce a reliable estimate, and any attempt to calculate an estimate of total demand for compute requires a combination of primary data collection, secondary analysis, and transparent assumptions.

Given these constraints, the study adopted a bottom-up approach. The primary data source for the study was a survey, designed to capture how researchers and practitioners currently use HPC for AI, how they expect their usage to evolve, and what is driving those changes. Subsequent data treatment steps are then described in more detail below.

2.2 High-level Methodological Approach

The study was undertaken across four interrelated phases: a rapid evidence review, a survey of HPC users, a bibliometric analysis of UK AI research, and a modelling exercise. Figure 1 provides a schematic overview of the study.

Figure 1 Schematic overview of the study



A technical annex accompanies this report with step-by-step calculations, a description of approaches to estimating user populations (and therefore user demand) considered (and

those that were eventually discarded following a workshop with BIST), and full details of the assumptions underpinning the modelling. A separate, interactive, dashboard provides further detail on the various scenarios considered.

2.2.1 Rapid Evidence Review

The study began with a rapid review of existing literature, published studies and secondary data sources on compute demand. This served multiple purposes. First, it informed the design of the survey by identifying the dimensions of compute usage that are most relevant and measurable across different user types. Second, it surfaced existing methodological approaches and published demand projections that could subsequently be used to triangulate and sense-check our own estimates. The evidence review also identified several relevant international comparators and sector-level studies, which are referenced at appropriate points throughout this report. Lastly, the exercise identified relevant secondary data sources that were used for conversions of technical factors (e.g. GPUh to H100 equivalent GPUh) and are described in more detail later in the report and in the accompanying technical annex.

2.2.2 Survey of UK HPC Users

The primary data collection tool was a structured online survey, launched in January 2026 and held open for one month. The survey was designed to capture three main sources of evidence:

- **User profile** — including sector, organisation type, and research discipline, to enable segmentation of the results.
- **Current compute demand** — including volume of compute used, the systems and infrastructure accessed, and the nature of the workloads run and any specific sovereign compute needs.
- **Future demand expectations** — including anticipated changes in usage over a three- to five-year horizon, and the factors driving those changes.

The survey did not take a sampling approach, and instead sought to capture a broad range of perspectives and use cases as possible. It was distributed through a range of networks and channels likely to reach active HPC users, which means the sample may be skewed towards more intensive users. This should be borne in mind when interpreting the demand estimates as they likely reflect the higher end of the usage spectrum and may overstate average demand across the full population of HPC users.

The survey received 229 responses in total. The majority of respondents came from academia (61% of responses), with the remainder drawn from industry (21%) and the public sector (6%). Although the study originally intended to disaggregate data according to the size of the organisation (e.g. differentiating between SMEs, micro entities, etc.), the sample sizes were not big enough to support robust findings. As such, for the purposes of this study and for the remainder of the analysis, we have grouped all industry responses into a single 'industry' answer. As the total population of UK HPC users is not precisely known, it is not possible to calculate an overall response rate; the implications of this for the representativeness of our findings are discussed in Section 3.3.

For the purposes of estimating both current and future demand, the sample sizes are somewhat different (see Table 1). A key survey question asked respondents to report their current demand in terms of total GPU hours used over the past 12 months. About 61% (n=126) of respondents reported their GPU usage, which after data cleaning resulted in 118 usable responses (57% of

the total). A small number of responses were excluded as outliers – cases where reported usage was substantially higher or lower than the rest of the sample that including them would have distorted aggregate estimates.[8] These data form the basis for the demand modelling described in subsequent sections.

Table 1 Survey response rates across user groups

User Type	Count of Total Responses	% of total responses	Count of responses reporting current GPUh
Academia	143	62%	92
Public Sector	16	7%	6
Industry	49	21%	20
Total	230	-	118

Source: Technopolis (2026). Note that the responses reporting current GPUh excludes outliers.

Estimates of public sector demand (both current and future demand) should be treated with some caution, as it is based on the smallest sub-sample in the survey (n=6 organisations / departments in the public sector). This figure is likely influenced by a small number of respondents known to be particularly compute-intensive, and may not be representative of the broader public sector population.

2.2.3 *Bibliometric Analysis*

To complement the survey data, we conducted a bibliometric analysis using OpenAlex.[9] We identified papers with at least one author affiliated with a UK institution and screened abstracts for references to AI methods. This analysis served two functions: it provided an estimate of the proportion of UK academics engaging with AI and HPC in their research, and it yielded a disciplinary breakdown of that activity, which informed our subsequent segmentation of the survey data and the scaling methodology described below. Additional details on the bibliometric approach are provided in the technical annex.

2.2.4 *Modelling Current and Future Demand*

The primary unit of measurement used throughout this analysis is **GPU-hour (GPUh)** — specifically, hours on an NVIDIA H100-equivalent GPU. We selected this metric because it is workload-agnostic and operationally intuitive. It can be reported by users across sectors and institutions without requiring detailed knowledge of underlying system architectures, and it is broadly comparable across different compute environments. Calculations of GPUh also

[8] Following assessment of the underlying distribution, the preferred approach to outlier exclusion used a logarithmic transformation. Specifically, observations that exceeded 2 standard deviations from the mean following the transformation were excluded. This resulted in any observation below 10.8 H100e GPUh or above 10,557,466 being excluded. More detail on the treatment of outliers is given in Chapter 8 of the Technical Annex.

[9] OpenAlex is a comprehensive open catalogue of academic publications

enabled the study team to calculate alternative metrics (including FLOPs and energy consumption) where required. Full details on the approaches to these calculations and conversions are set out in the technical annex.

The modelling proceeded in two stages. First, survey responses were used to construct estimates of current compute demand for the survey sample. These were then scaled to UK-wide totals for three user populations (academia, the public sector, and industry) using secondary data sources to characterise the size and composition of each user group. A summary of the total number of estimated users, and a brief description of the approach used to arrive at these estimates are presented in the table below. In addition, we provide a reflection on the quality of the data available to make these estimates and an assessment of the impact of those assumptions on the subsequent modelling.

Table 2 Estimated HPC users across sectors and assessments on the quality of evidence

User Sector	Estimated Number of HPC Users	Data Sources	Assessment of the Approach
Industry	~47,000	• UK HPC market revenue (2024)[10] • Academia HPC market revenue share[11] • Public sector HPC market revenue share[12] • Average UK business expenditure on AI[13]	• **Data Quality**: **Medium** • **Impact on Modelling**: **Medium** The study used the share of the UK HPC market revenue (2024) and average industry expenditure on HPC to arrive at an estimated number of industrial users. The validity of this method implicitly relied on the assumption that all UK businesses who report an expenditure (of greater than 0) on AI are using HPC for AI. This includes adopters of proprietary AI products or services and introduces a risk of double counting through the inclusion of both the provider and the adopter. This method was chosen as a 'middle estimate' of industrial users following a workshop with BIST.
Academia (Method 1) (Middle)	~37,700	• Proportion of UK papers that make use of AI (identified via bibliometrics) • Estimates on the size of the UK research base (2013)[14]	• **Data Quality**: **Medium** • **Impact on Modelling**: **Low** There are no single sources of evidence (at the time of analysis) that set out both the size of the UK research base and then the number of academic users. The study

[10] UK High Performance Computing Market Size & Outlook, Grand View Horizon Research, 2025, Accessible at: https://www.grandviewresearch.com/horizon/outlook/high-performance-computing-market/uk/reports

[11] High Performance Computing Market, Emergen Research, 2026, Accessible at: https://www.emergenresearch.com/industry-report/high-performance-computing-market

[12] Compute Evidence Annex, DSIT, 2025, Accessible at: https://www.gov.uk/government/publications/uk-compute-roadmap

[13] AI Adoption Research, DSIT, 2026, Accessible at: https://www.gov.uk/government/publications/ai-adoption-research/ai-adoption-research

[14] International Comparative Performance of the UK Research Base – 2013, Elsevier/ Department of Business, Innovation and Skills (BIS), 2013, Accessible at: https://assets.publishing.service.gov.uk/media/5a7c256740f0b645ba3c6fae/bis-13-1297-international-comparative-performance-of-the-UK-research-base-2013.pdf

Academia (Method 2) (Mid-high)	~42,500	• Proportion of UK papers that make use of AI (identified via bibliometrics) • Estimates on the number of researchers in R&D - UK (2017)[15]	used five different approaches to calculate the size of the academic HPC user base, and numbers range from ~22k to ~62k. Given the range in the estimated numbers we have presented demand estimates across all estimated values of the academic user base. However, through the modelling process it became clear that the total number of academic users is not the primary driver of demand: it is the number of industry users (given the considerably greater scale of their individual demand) that has the most significant influence on total demand estimates." For these reasons, the study team concluded that the impact of these estimates on the modelling is minimal.
Academia (Method 3) (Low-low)	~22,500	• Proportion of UK papers that make use of AI (identified via bibliometrics) • Estimates on the number of FTE R&D staff in UK universities (2025)[16]	
Academia (Method 4) (Low-mid)	~23,400	• Proportion of UK papers that make use of AI (identified via bibliometrics) • Total employment in research & technical roles, Royal Society (2019)[17]	
Academia (Method 5) (High-high)	~62,400	• Proportion of UK papers that make use of AI (identified via bibliometrics) • Average publications per UK researcher per year - Technopolis derivation based on UK research	

[15] Researchers in R&D (per million people) – United Kingdom, World Bank Group Data, 2025 (most recent year 2017), Accessible at: https://data.worldbank.org/indicator/SP.POP.SCIE.RD.P6?locations=GB

[16] R&D skills supply and demand: long-term trends and workforce projections, DSIT, 2025, Accessible at: https://www.gov.uk/government/publications/rd-skills-supply-and-demand-workforce-trends-and-projections/rd-skills-supply-and-demand-long-term-trends-and-workforce-projections#executive-summary

[17] The research and technical workforce in the UK, the Royal Society, 2021, Accessible at: https://royalsociety.org/-/media/policy/Publications/2021/2021-02-12-research-and-technical-workforce-in-the-uk.pdf

		& innovation performance (2023)[18]	
Public sector	~640	• UKRI approved research organisations[19] • UK government departments, agencies and other public bodies[20]	• **Data Quality**: **Low** • **Impact on Modelling**: **Medium** There are a total of 183 UKRI approved research organisations. The study has assumed that at least one team uses advanced compute for AI in each of these research organisations. In addition, the study team reviewed a 2024 Research Briefing titled AI in UK government departments to identify additional government departments that would likely use HPC. The final approach (agreed together with BIST) assumed 1 research team per UK government department, agency and public body. A key limitation of this approach is that it assumes a uniform intensity of advanced compute deployment across all public research establishments and government departments, which is unlikely to reflect reality. In the absence of secondary data, it was not possible to account for variation in the intensity of advanced compute use across public sector organisations. After excluding outliers, public sector demand was the lowest average across the three user groups, and we therefore do not think that this has significantly impacted modelling results even if the total number of public sector users changes.

Source: Technopolis (2026)

[18] UK research and innovation performance: A Global Node for Collaboration and Innovation, Elsevier & Universities UK, 2025, Accessible at: https://elsevier.shorthandstories.com/uk-research-and-innovation-performance/index.html

[19] UKRI list of approved research organisations, UKRI, 2026, Accessible at: https://www.gov.uk/government/publications/ukri-endorsement-employing-or-hosting-institutions-global-talent-visa/ukri-list-of-approved-research-organisations

[20] Departments, agencies and public bodies, UK Government, 2026, Accessible at: https://www.gov.uk/government/organisations

2.2.5 A Note on Data Limitations and Assumptions

As with any study of this kind, a number of limitations should be borne in mind when interpreting the results. To mitigate the risk of methodological error, a workshop was held with BIST prior to finalising the modelling approach. The workshop reviewed a range of candidate methodologies (including several that were subsequently discarded) and arrived at a consensus approach. The technical annex documents the alternatives considered and the rationale for the choices made, and the main limitations of the study are highlighted below:

- **Survey sample size and representativeness.** With 231 responses, the survey provides a meaningful evidence base, but the sample is not large enough to support highly granular segmentation in all cases. While there was initially an intention to segment industry responses by industry type (e.g. large company vs SME), the sample sizes were not large enough to allow for meaningful and robust estimates. Where sub-group sample sizes are small, results are presented with appropriate caveats.
- **A fast-moving landscape.** AI compute demand is growing and evolving rapidly, making it difficult to obtain primary data that fully reflects the current state of play. Findings from the survey represent a snapshot taken in early 2026 and should be interpreted accordingly. Where secondary sources provide more recent data, these have been incorporated and noted.
- **Modelling assumptions.** The demand estimates presented in Sections 4 and 5 are built on a number of assumptions, most notably regarding the size of UK user populations and the representativeness of the survey sample. At each stage, we have clearly set out the assumptions made, the evidence on which they are based, and their associated uncertainties. Full details, including sensitivity analysis, are provided in an accompanying technical annex.

3 Current Demand for High Performance Compute for AI

3.1 Introduction

This section presents our findings on current demand for high-performance compute (HPC) for AI in the UK. It draws on survey data collected from industrial, academic, and public sector users, supplemented by a bibliometric analysis of AI-related research activity. We present estimates of current demand both in terms of compute consumption (measured in GPUh) and electrical power (measured in gigawatts, GW) and discuss the methodological assumptions underlying each estimate.

3.1.1 Demand in GPUh

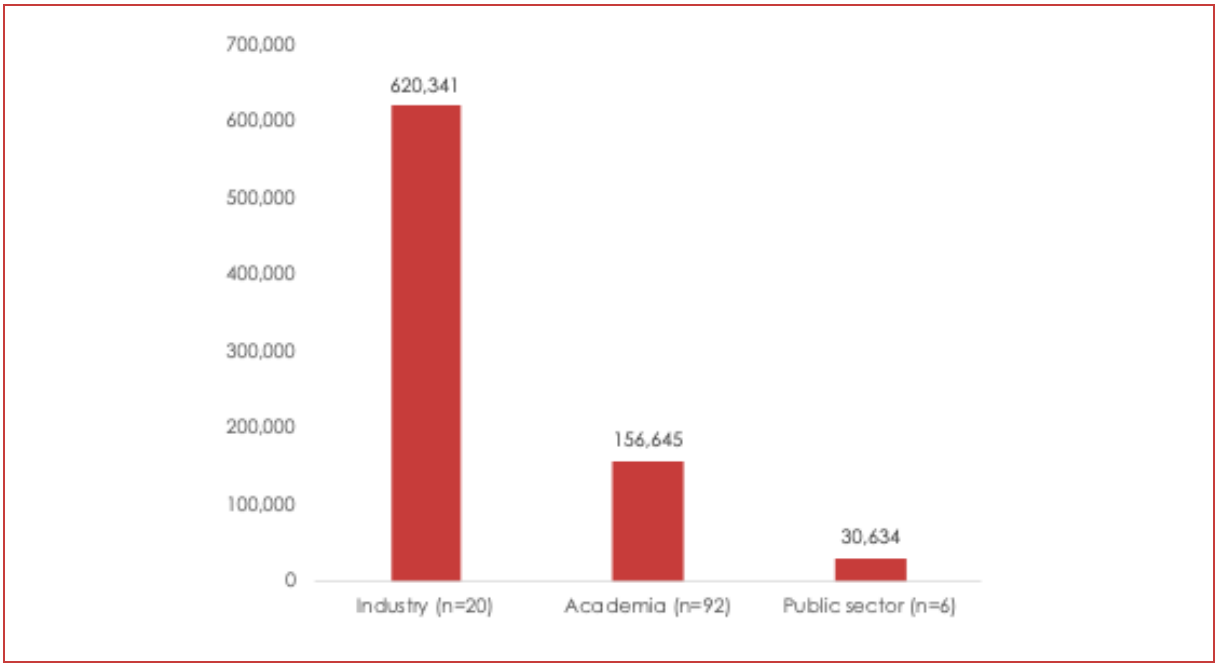
To estimate current demand, respondents were asked to report their total demand for HPC in the last 12 months in GPU hours (GPUh). Where respondents were unable to provide this figure directly, they were asked to estimate their last 12 months of expenditure on compute instead.[21] While this approach introduces additional complications and risk (particularly in converting expenditure figures into GPUh) this approach was deemed necessary to maximise survey responses, capturing data from a broad range of users, including those unable to express demand in GPUh terms. All responses have then been converted to H100-equivalent GPUh (H100e GPUh) to allow for consistent comparison across respondents using different hardware configurations.

Industrial users reported the highest average demand for compute, measured in H100e GPUh per organisation. This is consistent with findings in the wider literature, which suggest that industry's share of global AI compute has grown substantially from approximately 40% in 2018 to around 80% by 2025[22]. This may reflect the rapid commercialisation of AI and the scale of model training and inference workloads in the private sector.

[21] Conversions were made based on average prices to rent an hour on an H100 GPU (as of January 2026)

[22] Trends in Ai Supercomputers, Pilz, Sanders, Rahman and Heim, 2025, Accessible at: https://arxiv.org/abs/2504.16026

Figure 2 Average annual demand for compute by sector (in H100e GPUh)



Source: Technopolis analysis of survey data. Note that 'n' values here refer to respondents who provided estimates of current demand.

Survey responses have been scaled up to produce estimates of UK-wide compute demand, using the population estimates set out in Section 3.2.4. Because of the range of potential academic users (and the uncertainty associated with those estimates), we present a range of projected demand corresponding to different assumptions about the size of the academic user population (see Table 3 below).

Table 3 Current UK compute demand estimates under alternative academic user population estimates

Number of academic users	Total Demand (bn H100e GPUh)
~22k	33.3
~23k	33.5
~37k	35.7
~42k	36.4
~62k	40.0

Source: Technopolis analysis of survey data (2026)

Across these scenarios, current total demand for compute is estimated to fall in the range of **33.3 to 40.0 billion H100e GPUh.** Interestingly, tripling the assumed number of academic users does not result in a similar tripling of total demand. This suggests that demand for AI compute is primarily driven by industry, meaning that any variation in the academic population has a comparatively modest effect on aggregate demand.

To put this central estimate in context (using the mid-point assumption of approximately 37,000 academic users) the total UK demand of 35.7 billion H100e GPUh is equivalent to approximately 6.5 million GPU hours on Isambard AI, or roughly 750 Isambard-AI's running simultaneously across a full year.

3.1.2 Demand for power (GW)

Current UK demand for AI compute translates to an estimated **5.12 – 6.08 GW** of electrical power when accounting for both real-world usage patterns and full infrastructure requirements – equivalent to roughly 11.9% of current UK peak electricity demand[23].

Two dimensions determine where within the range power estimates fall: (1) whether estimates account for only the power demands from the GPU card or whether the full supporting infrastructure is included in estimates, and (2) whether demand is assumed to be continuous or adjusted for idle time. Table 4 summarises the resulting estimates across these dimensions, and further detail on the approach to calculations and their implications on the results is presented in subsequent sections.

Table 4 Summary of current UK power demands for HPC in the UK

	GPU card only	**Whole system**
Continuous (unadjusted)	2.66 – 3.16 GW	3.73 – 4.44 GW
Run-time adjusted	3.66 – 4.33 GW	5.12 – 6.08 GW

Source: Technopolis analysis of survey data (2026)

From GPU hours to card-level power

The baseline power estimate is derived by converting total annual GPUh into a continuous power draw, assuming an H100e maximum draw of 700W per graphics card.[24]

> On this basis, current UK demand for compute falls between **2.66 and 3.16 GW**, with the range reflecting the different estimates on the size of the academic user base.

This baseline assumes GPUs run at full utilisation continuously throughout the year. However, survey data suggests this overstates effective utilisation in practice. Respondents were asked

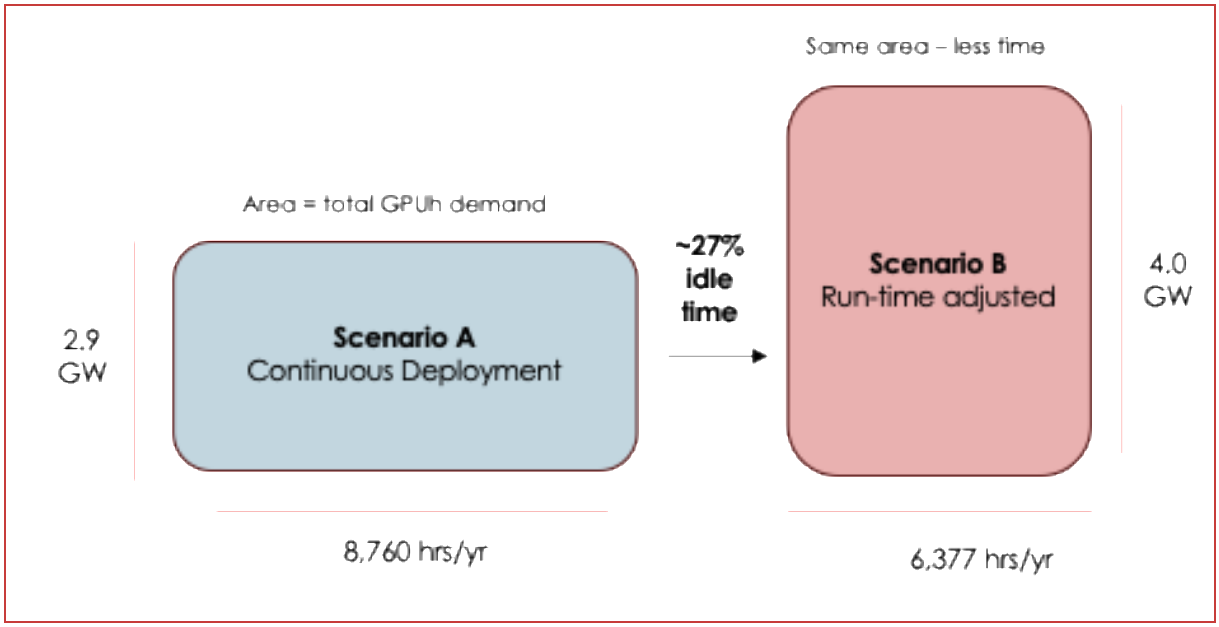
[23] In 2025, UK electricity demand peaked at 46GW. Source: Britain's Energy Explained: 2025 Review, NESO, Accessible at: https://www.neso.energy/news/britains-energy-explained-2025-review#:~:text=It%20hit%20an%20annual%20high,generation%20mix%20in%20real%20time

[24] An AI GPU (Graphics Process Unit) or "graphics card" is a high-performance, specialised GPU designed to accelerate AI workloads, such as machine learning, data analytics and deep learning. Accessible at: https://www.lenovo.com/us/en/knowledgebase/ai-graphics-cards-a-comprehensive-guide-for-2025/

to report (on average) how they split their allocated compute time between idle time and active use. On average, respondents reported spending approximately 27% of their allocated compute time idle, meaning GPUs are active for only 73% of the period for which capacity is provisioned.

This has important implications for estimating peak power demand. If total annual demand in GPU hours remains constant, but is consumed over a shorter effective window (because of idle time), then the peak power required to deliver that demand increases (see Figure 3), which underpins the calculations for run-time adjusted calculations.

Figure 3 Schematic overview of run-time adjusted power demands



Source: Technopolis (2026). Note that GW values reported in the figure are illustrative only.

From card-level to whole-system power

GPU cards do not operate in isolation. Delivering compute at scale requires supporting infrastructure (including memory, interconnects, storage arrays, and management servers) each of which draws additional power. To capture this, we apply a whole-system multiplier to the card-level estimates above.

It is important to note that this analysis focuses on the power demands of individual compute systems and does not extend to broader data centre infrastructure, nor does it account for the effects of interconnected GPU clusters, where workloads distributed across multiple systems may yield efficiency gains or introduce additional overhead depending on the configuration.

The multiplier is derived from Isambard-AI[25], used here as the reference infrastructure unit consistent with the H100-equivalent framework applied throughout this study. Based on its

[25]Isambard-AI is currently the UK's most powerful public compute facility.
https://www.gov.uk/government/publications/ai-research-resource/airr-advanced-supercomputers-for-the-uk

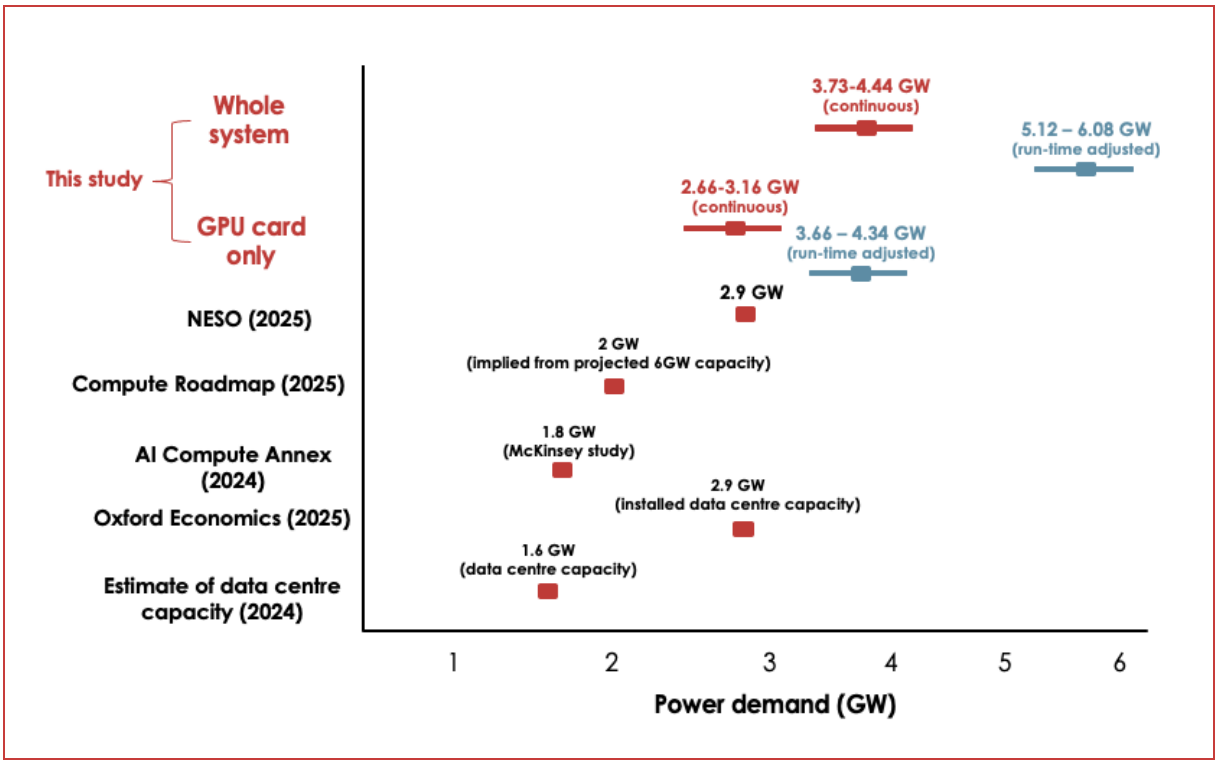
published peak power capacity of 7.63 MW across 5,448 chips, and accounting for the power characteristics of the NVIDIA GH200 Superchip at representative operating conditions, this yields a server-to-GPU ratio of approximately 1.4x.[26] In practice, this means that for every watt drawn by the GPUs, the surrounding infrastructure accounts for a further 0.4W.

It should be noted that Isambard-AI represents a large-scale, energy-efficient deployment; smaller or older facilities are likely to exhibit higher multipliers due to less efficient cooling and power delivery infrastructure. For example, the 8-GPU DGZ server, rated at 10,200 W with 8 H200s yields a server to GPU-ratio of around 1.8x.[27]

> Applying the multiplier from Isambard-AI to the card-level estimates produces whole-system power demands of **3.73 – 4.44 GW** on a continuous basis, rising to **5.12 – 6.08 GW** once adjusted for idle time.

Figure 4 places this estimate in the context of other recent studies.

Figure 4 Comparison of estimates from this study with wider estimates of power demand



Source: Technopolis (2026)

[26] According to the EPOCH AI GPU Clusters database. Available at https://epoch.ai/data/gpu-clusters?view=map&tab=point

[27] NVIDIA DGX H200 Power Consumption: What you Absolutely Must Know, Uvation (2025). Accessible at: https://uvation.com/articles/nvidia-dgx-h200-power-consumption-what-you-absolutely-must-know#conclusion-powering-the-next-wave-of-ai

This study consistently estimates demand values higher than recent studies, although this is perhaps unsurprising. The studies were conducted at different points in time; as the most recent estimate (drawing on data collected in January 2026) this study reflects a landscape in which AI has rapidly scaled and evolved, meaning earlier estimates may have become outdated relatively quickly or based on earlier technologies and AI development rates. All studies that seek to quantify both current and future demand for AI compute are inherently estimates, and therefore some divergence between them is to be expected. Specifically for these studies, the following differences might explain the different values reported between this study and others:

- **Different methodologies:** Each study employs different methodologies and underlying assumptions, which will naturally diverge. For example, the projections from the Compute Roadmap are inferred from projected installed capacities, while this study has adopted a bottom-up approach by directly consulting with both current and potential HPC users. This will necessarily introduce differences
- **Scope of the studies:** Where possible, comparisons have been drawn with studies of similar scope. In some cases, however, the studies identified for comparison have a broader remit (e.g. extending to full data centre infrastructure). The differences in the scope is likely to contribute to differences in the demand estimates between those studies and the results presented in this report.
- **Assumptions:** Finally, not all studies are fully transparent about the assumptions underpinning their figures. A notable example is the use of run-time adjustments: the adjustments applied in this study are specific to the sampled population, and it is unclear to what extent comparable adjustments have been applied (or disclosed) in other analyses.

Together, these factors should be borne in mind when comparing estimates across studies.

3.2 Insights into Current Use of HPC for AI

Beyond the aggregate demand estimates presented in Section 4.1, the survey and bibliometric data offer insight into how HPC for AI is current being used – across technologies, disciplines, access routes, and hardware configurations. Four key themes emerge, and are discussed below.

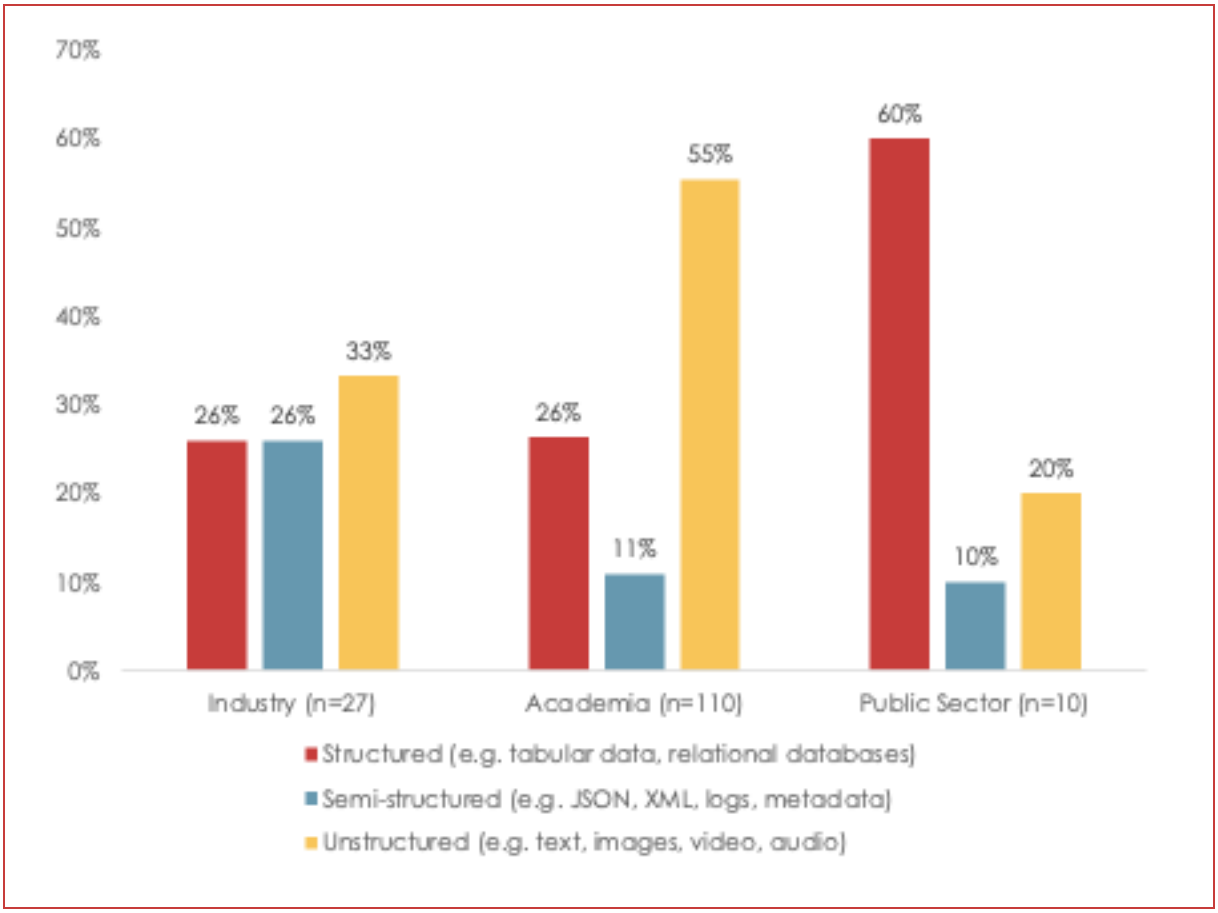
The most compute-intensive AI techniques dominate workloads

HPC demand is concentrated around the most computationally demanding AI approaches. Large Language Models (LLMs) are the most widely used, reported by 62% of survey respondents, followed by multimodal AI (50%) – both of which require substantial compute to train and deploy at scale. It is worth noting that the training of frontier LLMs – the most compute-intensive phase of model development – is largely conducted in a handful of countries (the United States, South Korea) by a small number of large technology companies, and would not be captured in estimates of UK demand. The compute demand reflected here largely relates to fine-tuning, adaptation, and inference workloads built on top of existing foundation models, rather than training from scratch. Both activities, nonetheless, require substantial compute resources at scale.

The nature of these workloads differs across sectors (see Figure 5). Industry users show a broadly even split between structured and unstructured data, while academic users rely considerably more on unstructured data (text, images, video, and audio) consistent with the dominance of deep learning and foundation model research in the academic literature. Public sector

workloads skew in the opposite direction, with a higher proportion relying on structured tabular data, reflecting the administrative and analytical character of most public sector AI applications.

Figure 5 Types of data used by sectors

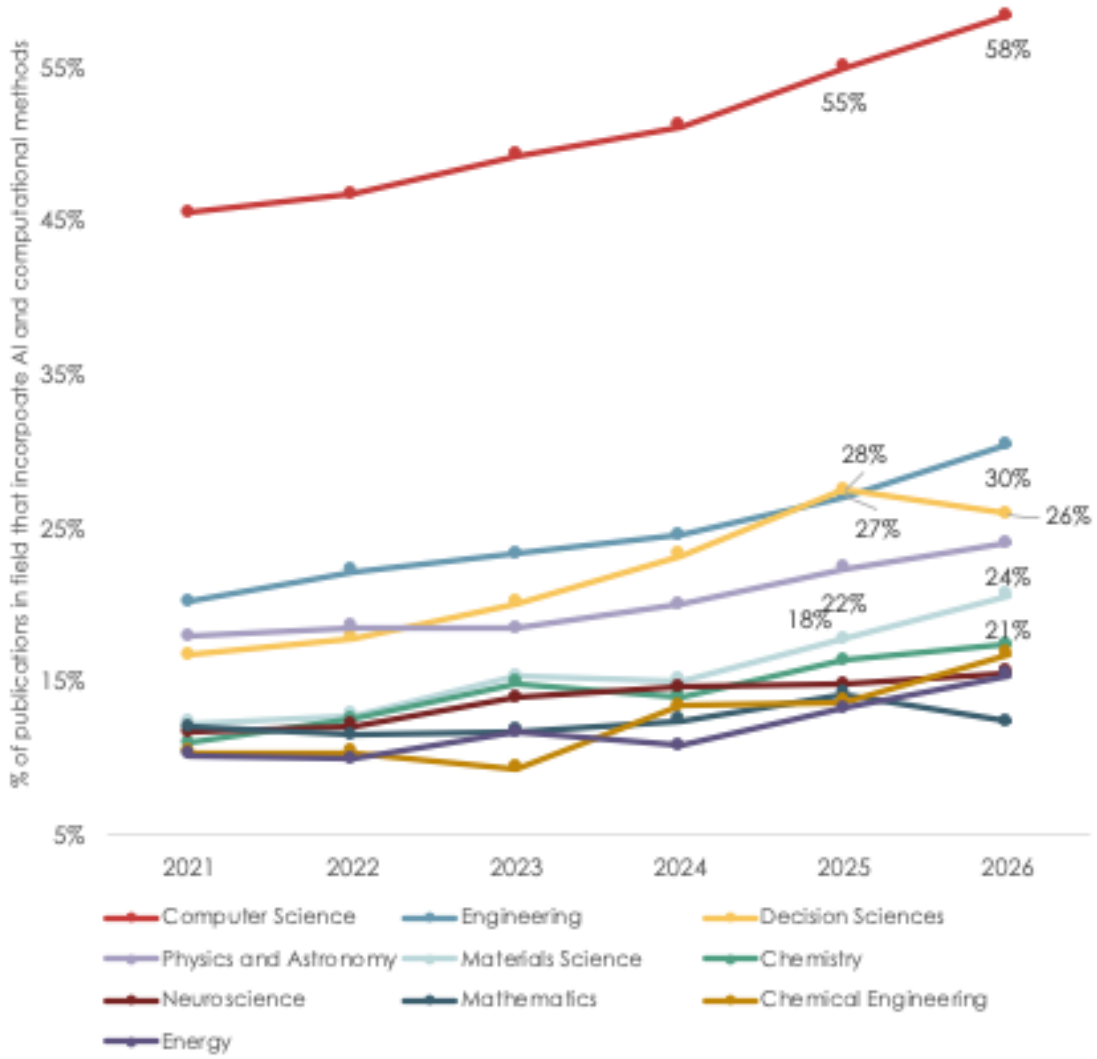


Source: Technopolis analysis of survey data (2026).

The use of AI in research is growing across all disciplines

The bibliometric analysis shows that AI is becoming a broadly adopted research tool, not merely a subject of study in its own right. The share of UK research publications incorporating AI rose from 9.5% in 2021 to 14.4% in 2025, with growth recorded across all disciplines. Figure 6 shows the proportion of publications that mention the use of AI or Computational Methods (CM) across the top 10 fields for AI use.

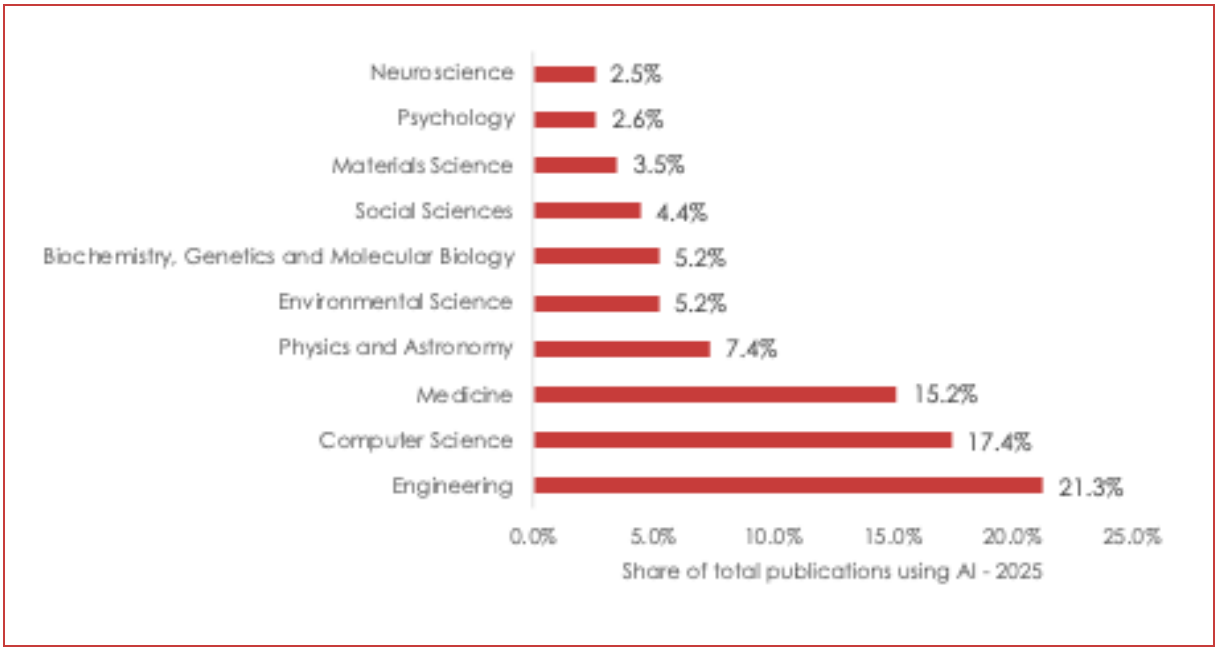
Figure 6 Shares of research publications using AI or Computational Methods by discipline



Source: Technopolis analysis of OpenAlex (2026). The % of publications in field that incorporate AI or computational methods (CM) are shown for the most recent years of analysis (2025 and 2026) for the five fields with the highest AI usage.

Computer Science has the highest intensity of AI use, with 55% of 2025 publications referencing AI or computational methods. Engineering (27%) and Decision Sciences (28%) follow. However, intensity of use and share of overall AI-related publications do not always align: while Computer Science leads on intensity, Engineering accounts for the largest share of total AI-related UK publications (21.3%), and Medicine (despite not featuring in the top 10 disciplines by intensity) holds a substantial share of the overall volume (Figure 7).

Figure 7 Share of total publications mentioning AI use in 2025



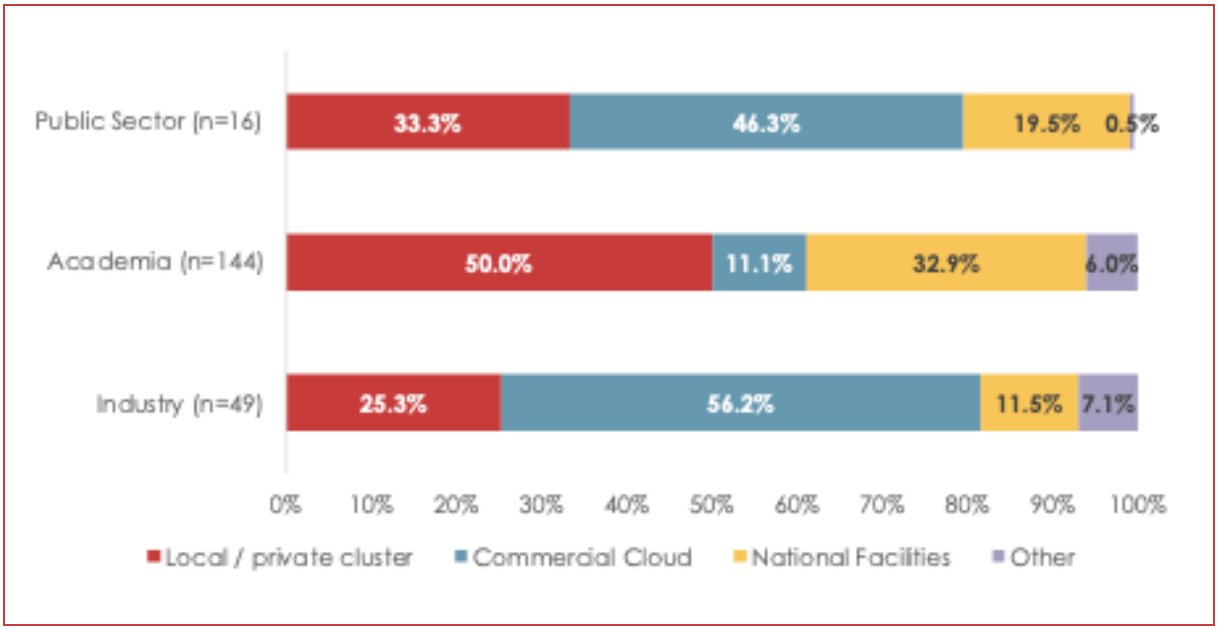
Source: Technopolis analysis of OpenAlex (2026)

This distinction matters for understanding where HPC demand for research originates: high-volume disciplines with moderate AI intensity may generate comparable aggregate compute demand to lower-volume, high-intensity fields.

National facilities meet a minority of total demand

Overall, 44.6% of reported compute time is accessed through local or private clusters, however there is a degree of variation across sectors. Across respondents, the academic sector tends to access compute resources through physical systems to a greater extent than Industry and the Public sector, who use significantly more cloud compute. Across all sectors, **28.3% of respondents indicated that they access compute through national facilities**. However, this figure again varies significantly when we disaggregate responses by sector (Figure 8).

Figure 8 (Average) models by which respondents access compute (n=126)



Source: Technopolis (2026). Note 'n' here refers to the number of respondents that reported their share of compute use.

As expected, academic users have the highest relative share of compute time accessed through national facilities, followed by the public sector. Industrial use of national facilities remains low at 11.5% of total compute time. However, not all surveyed companies are eligible for national facility use. For example, AIRR access is limited to start-ups and SMEs, with the understanding that larger companies likely have private access to clusters. However, it is important to note the potential for response bias in the survey findings. Although significant efforts were made to promote the survey widely (including through BIST and UKRI newsletters, LinkedIn promotion, and direct outreach to known or likely HPC users) the responses are still likely to disproportionately reflect the views of current users of national facilities and the responses from Figure 8 should be interpreted accordingly.

Combining sector weights with sector-level demand estimates allows us to estimate the share of total UK HPC demand that is currently met by national facilities.[28]

Approximately 15% of current UK advanced compute demand is currently met through national facilities.

It is important to distinguish between these two figures. The figure presented in the box above is notably lower than the 28.3% of total access because this estimates the share of total UK compute demand met by national facilities, taking into account the different levels of

[28] Specifically, we multiply the proportion of compute accessed through national facilities by UK-wide demand for each sector (industry, academia, public sector) using the middle estimate on the number of academic users.

compute demand across sectors. The industrial sector accounts for a large portion of demand, but only a minority of this is accessed through national facilities, with industrial users typically opting for cloud computing solutions or private infrastructure.

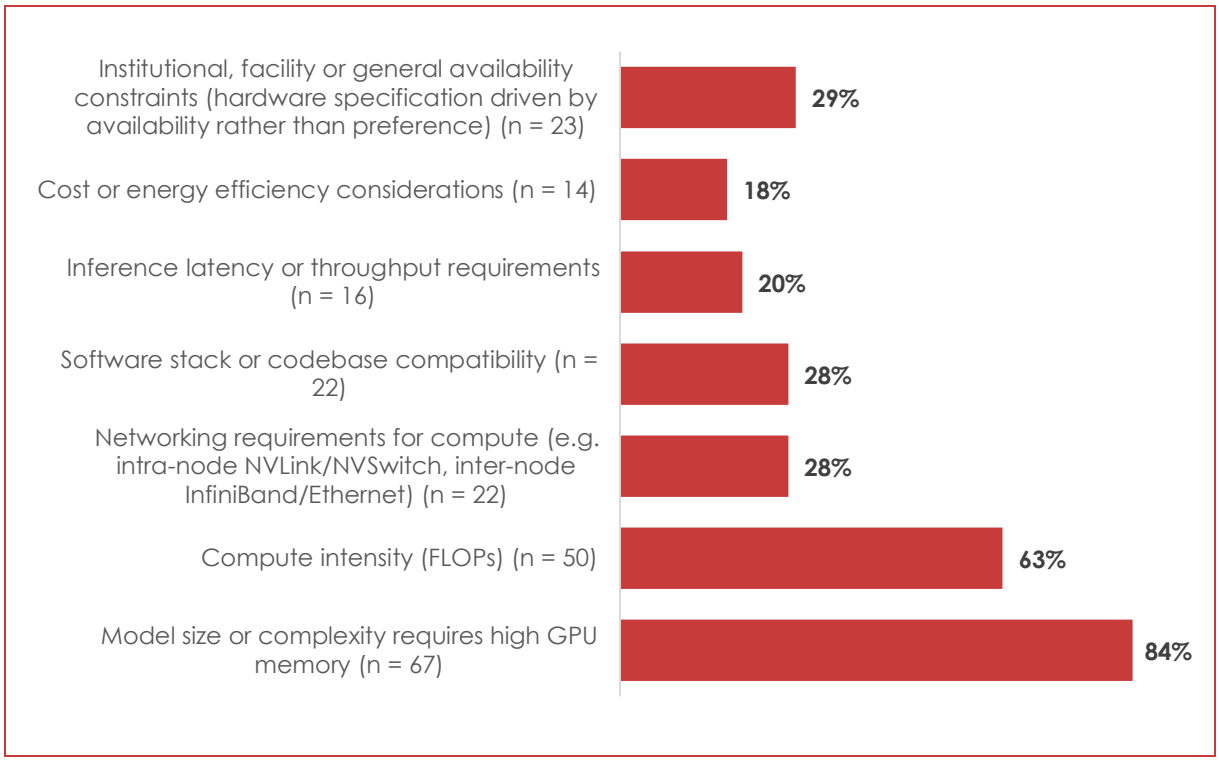
Vendor dependency is high and hardware-driven

A structurally significant feature of the current landscape is the degree to which workloads are tied to specific hardware vendors. Around two-thirds of respondents (69%) described their workloads as fully or mostly dependent on their current GPU manufacturer, with only 5% describing themselves as fully vendor-agnostic. While sample sizes do not support robust analysis about differences between user groups, academic respondents were more frequently able to answer the question about vendor dependency.[29] This may reflect how researchers procure access to compute. Academic users often apply for allocations through formal processes that tie them to their specific hardware requirements, whereas industry users purchasing cloud compute (for example) may simply select a service without needingto know or record the underlying vendor.

The primary driver of vendor dependency is hardware capability rather than software lock-in: 84% of respondents cited high GPU memory requirements as the principal constraint, followed by the raw compute intensity of their workloads. This dependency has implications for procurement flexibility and the resilience of the UK's compute supply, particularly as demand grows and hardware generations turn over (Figure 9).

[29] 46% of academic respondents, compared to 21% of public sector respondents and 19% of industry respondents.

Figure 9 respondents' reasons for specific vendor requirements (n = 80)



Source: Technopolis analysis of survey (2026)

Different compute requirements shape infrastructure needs

Figure 10 shows the allocation of compute time across different activities by sector, highlighting clear differences in how compute is used. Academic respondents devote the largest share of compute time to model training, reflecting their role in developing and advancing AI methods. In contrast, industrial users exhibit a more balanced distribution across training, inference, and data processing activities, consistent with the deployment and integration of AI into production environments – for example, running trained models against live data to generate predictions, recommendations, or automated decisions as part of an operational product or service. Public sector respondents report a relatively high share of compute time allocated to simulation activities, although this may be influenced by the smaller sample size for this group.

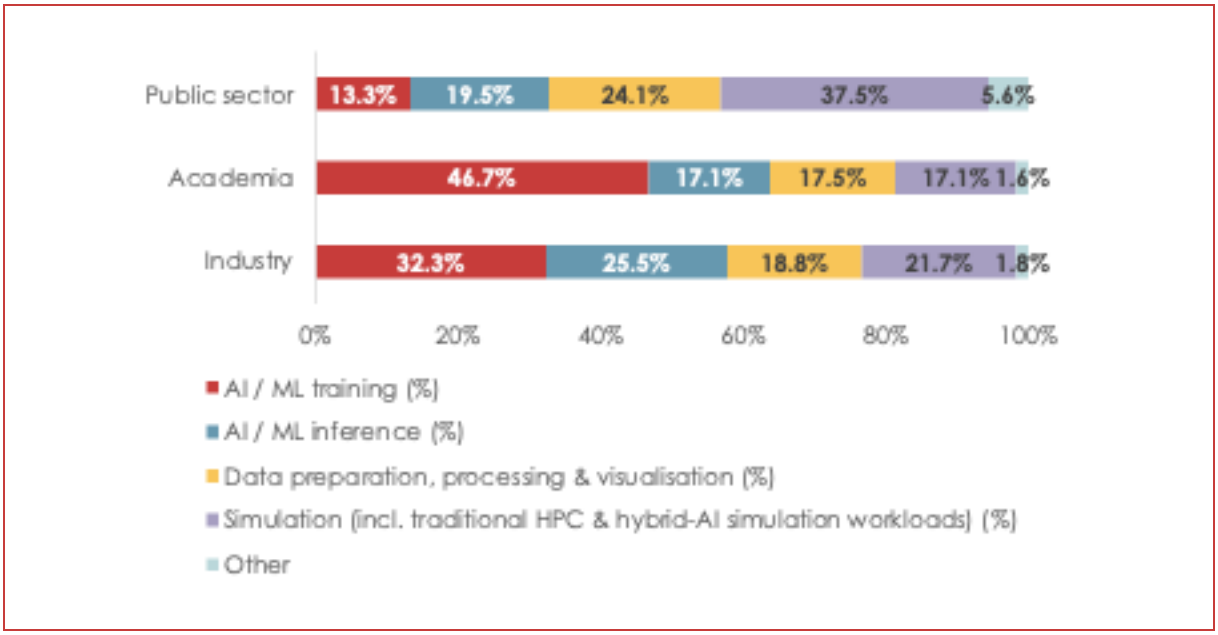
These differences in workload composition have important implications for infrastructure design and provisioning. Training workloads are typically the most compute- and energy-intensive, relying heavily on high-performance GPU accelerators and optimised interconnects.[30] By contrast, activities such as data preparation and processing tend to be more CPU-intensive and place greater demands on storage systems and data throughput,[31] although GPU

[30] The next big shifts in AI workloads and hyperscaler strategies, McKinsey (2025). Available at: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-next-big-shifts-in-ai-workloads-and-hyperscaler-strategies

[31] Efficient Tabular Data Preprocessing of ML Pipelines, Zhu, Jiang, and Alonso (2024). Available at: https://arxiv.org/abs/2409.14912

acceleration is increasingly being adopted in these areas.[32] A more inference-heavy workload profile, as seen in industry, may also require infrastructure that prioritises scalability, responsiveness, and cost-efficiency. This variation suggests that future compute provision (either publicly-funded or through private provision) will need to support a diverse mix of workload types, rather than being optimised solely for large-scale model training.

Figure 10 Use of compute time by sector



Source: Technopolis analysis of survey (2026). Note: the survey question asked respondents to estimate their average split across idle time and active compute use.

[32] For example, RAPIDS by NVIDIA accelerates the complete data science pipeline entirely on GPUs. See 'RAPIDS Accelerates Data Science End-to-End. Available at: https://developer.nvidia.com/blog/gpu-accelerated-analytics-rapids

4 Estimating future demand for compute

This section presents our findings on estimates for future demand for compute. Compute demand is estimated in terms of GPU hours, power capacity (Gigawatts) and FLOPs, and an analysis of the factors expected to shape demand trajectories over the next five to seven years.

As with the current demand estimates, a number of assumptions underpin the projections presented here. Principally, these relate to the use of predicted and derived growth rates, and the construction of future growth scenarios. In the interest of transparency, each of the key metrics has been assessed against the quality of the underlying data, with associated justifications, limitations and mitigations noted to help guide the interpretation of the results. Further methodological detail is provided in the technical annex.

It should also be noted that this report presents the results from a single central scenario. However, the modelling was developed across a range of growth assumptions and alternative scenarios, which can be explored interactively through an accompanying dashboard. Readers who wish to examine how demand estimates respond to different inputs are encouraged to use the dashboard alongside this report.

Table 5 Data quality and modelling impact assessment for future demand growth rates

Metric	Data Quality Rating	Impact on Modelling	Justification
Growth rate of compute in GPUh	**Medium**	**Low**	The HPC survey offered respondents six future scenarios, each comprising a short qualitative description and an associated year-on-year growth rate in GPU hours — for example, "Continued year-on-year growth at a similar rate to recent years (between 50% to 100%)". Respondents were asked to select the trajectory they considered most likely for their area of research, and a weighted average was applied to derive the central growth rate used in the modelling. It should be noted that growth rates vary significantly depending on sector, use case, and organisational context. Where possible, additional data points have been used for cross-referencing. Our bibliometric analysis, for example, found that the ten fastest-growing disciplines using HPC for AI grew at an average rate of 113% per year, providing a degree of external validation for the survey-derived estimates.
Growth rate of power and FLOPs	**Medium**	**Low**	Growth rates for power and FLOPs were derived from secondary literature, with a particular focus on industry reports. The available evidence was limited in volume and varied in its estimates. In response, conservative assumptions have been applied to avoid overstating future demand, and the sensitivity of results to these assumptions is explored through the scenario analysis.

Source: Technopolis (2026)

4.1 Future Demand in GPUh

To structure respondents' assessments of future compute demand, a Rapid Evidence Assessment (REA) was conducted to map the current AI growth landscape and identify plausible demand trajectories. The REA drew on existing literature and data sources to characterise a range of growth scenarios across the demand spectrum.

Findings from the REA were then used to develop six future scenarios, each comprising a short qualitative description and an associated year-on-year growth rate in GPU hours (see Table 6 below). These scenarios were presented to BIST at a dedicated methodological workshop, where the framing and range of options were reviewed.

Survey respondents were then asked to select the demand trajectory they considered most likely for their research area over the next five to seven years. The scenarios and their associated growth rates are set out in Table 5 below.

Table 6 Growth scenarios from the HPC user survey

Scenario	Description	Year-on-Year Growth rate (for GPUh) Y
1	Exponential year-on-year growth in demand	+100%
2	Continued year-on-year growth at a similar rate to recent years	+50 % to 100%
3	Continued year-on-year growth, but at a slower rate	+10% to 50%
4	Demand largely plateaus year-on-year	-10% to +10%
5	Demand declines (e.g. due to efficiency gains or reduced investment) year-on-year	-10% to -50%
6	Significant downturn in demand (e.g. "AI bubble" bursting) year-on-year	-50%

Source: Technopolis survey of UK HPC users (2026)

Based on a weighted average of responses, Scenario 2 (continued year-on-year growth at a rate comparable to recent years) was identified as the most likely outcome, followed by Scenario 1 (exponential year-on-year growth in demand).

The study calculated future demand for all scenarios presented in the table above. For the purposes of this report, estimates of future demand are presented only for Scenario 2 in the subsequent sections. Presenting all six scenarios in full (each with their underlying assumptions) risks diluting the key findings and significantly reducing the readability of the report. As Scenario 2 was identified as the most likely outcome by the respondents (i.e. the actual users of compute), it was judged most appropriate to present this scenario as the primary focus of the main report. Full calculations for each of the scenarios and the associated assumptions are presented in a separate dashboard for readers who wish to explore the full range of outcomes.

Drawing on the year-on-year growth rates associated with Scenario 2 (continued high growth), three trajectories were modelled to capture the range of plausible outcomes:

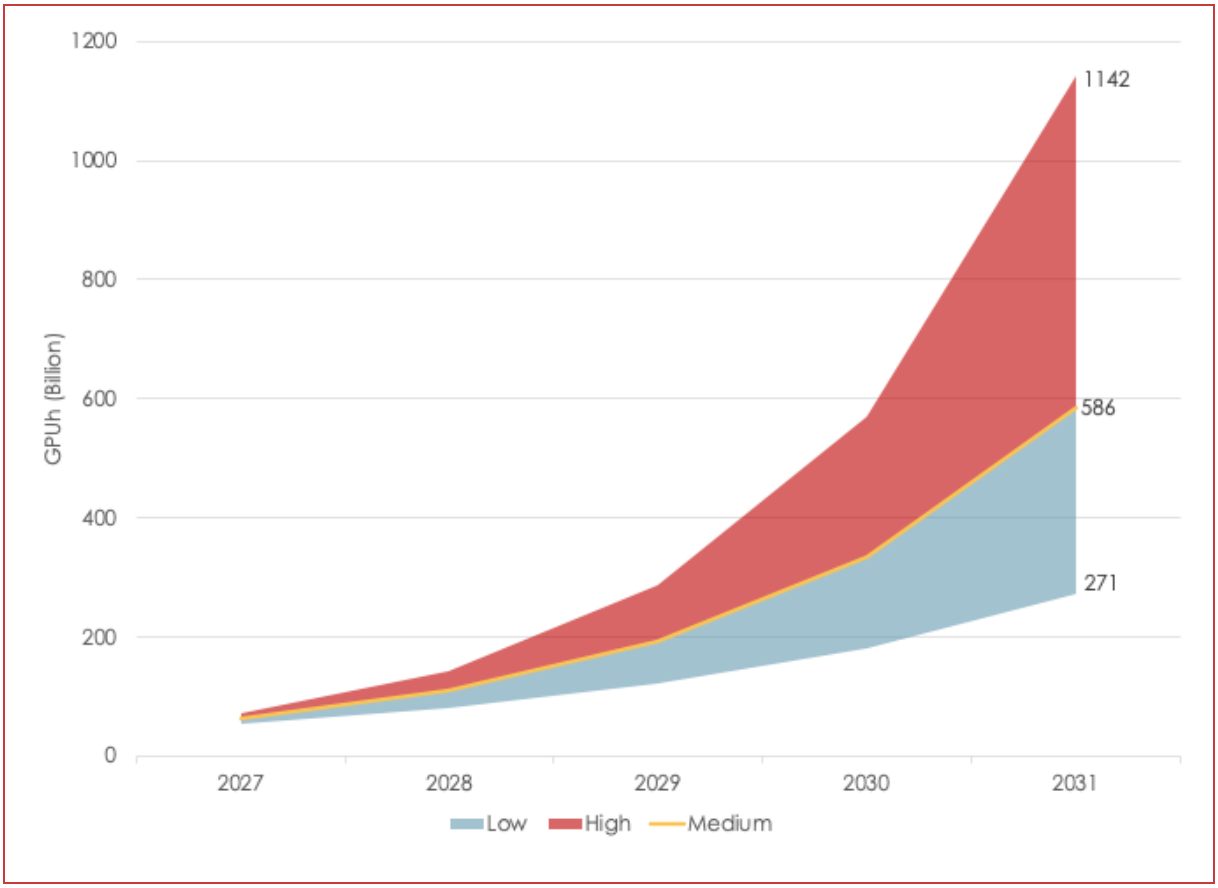
- High growth: 100% year-on-year growth

- Medium growth: 75% year-on-year growth
- Low growth: 50% year-on-year growth

These expectations are broadly consistent with observed trends in research activity: analysis of the ten fastest-growing disciplines for AI use shows average bibliometric growth of approximately 113% (109% when weighted by market share). This suggests that respondent expectations, while substantial, may in fact be slightly conservative relative to current trajectories (see Section 5.5 for further detail).

By 2031, projected demand ranges from 271 billion GPUh under the low growth trajectory, to 1,142 billion GPUs under the high-growth trajectory. The medium projection sits at 586 billion GPUh – equivalent to approximately 107.5 million GPUh on Isambard-AI, or roughly **12,275 Isambard-AIs**[33] **running simultaneously across a full year** (Figure 11).

Figure 11 Projected future demand (in GPUh)



Source: Technopolis (2026). Note that this scenario relies on the estimate of ~37K for the number of academic users, which changes demand projections.

[33] Isambard-AI (one of two supercomputers that make up the UK's AI Research Resource (AIRR)) is the UK's most powerful supercomputer. Note that this is being used as an illustrative metric, rather than these numbers suggesting that this is the capacity increased needed in the AIRR programme.

4.2 Future demand for Power (GW)

Alongside projected demand measured in GPUh, the study also estimated the associated electrical power requirements of meeting that demand. Two approaches are presented: one based on the power draw of the GPU card alone, and one based on the full infrastructure stack required to deliver that capacity.

Table 7 Estimated future power requirements

	GPU card only	Whole-system
Continuous (unadjusted)	5.91- 10.96 (8.15) GW	8.28 – 15.34 (11.4) GW
Run-time adjusted	8.12 – 15.05 (11.19) GW	11.37 – 21.07 (15.66) GW

Source: Technopolis (2026)

Unlike the GPUh projections, which drew on respondents' own growth estimates, the power demand projections rely on growth rates sources from existing industry reports (cited below). This approach was adopted to incorporate estimates of power (and associated efficiency) gains, which are an important factor in determining power consumption over time.

The medium estimate of current power demand (2.85 GW, see Section 4.1.2) was used as the baseline, with three growth rates then applied, derived from industry sources:

- High: 40% year-on-year
- Medium: 30% year-on-year
- Low: 20% year-on-year

These ranges reflect the inherent uncertainty in forecasting power demand. On one hand, the widespread adoption of AI chips and the growing scale of model training are likely to increase energy consumption. On the other, continuing improvements in chip efficiency may partially offset this.

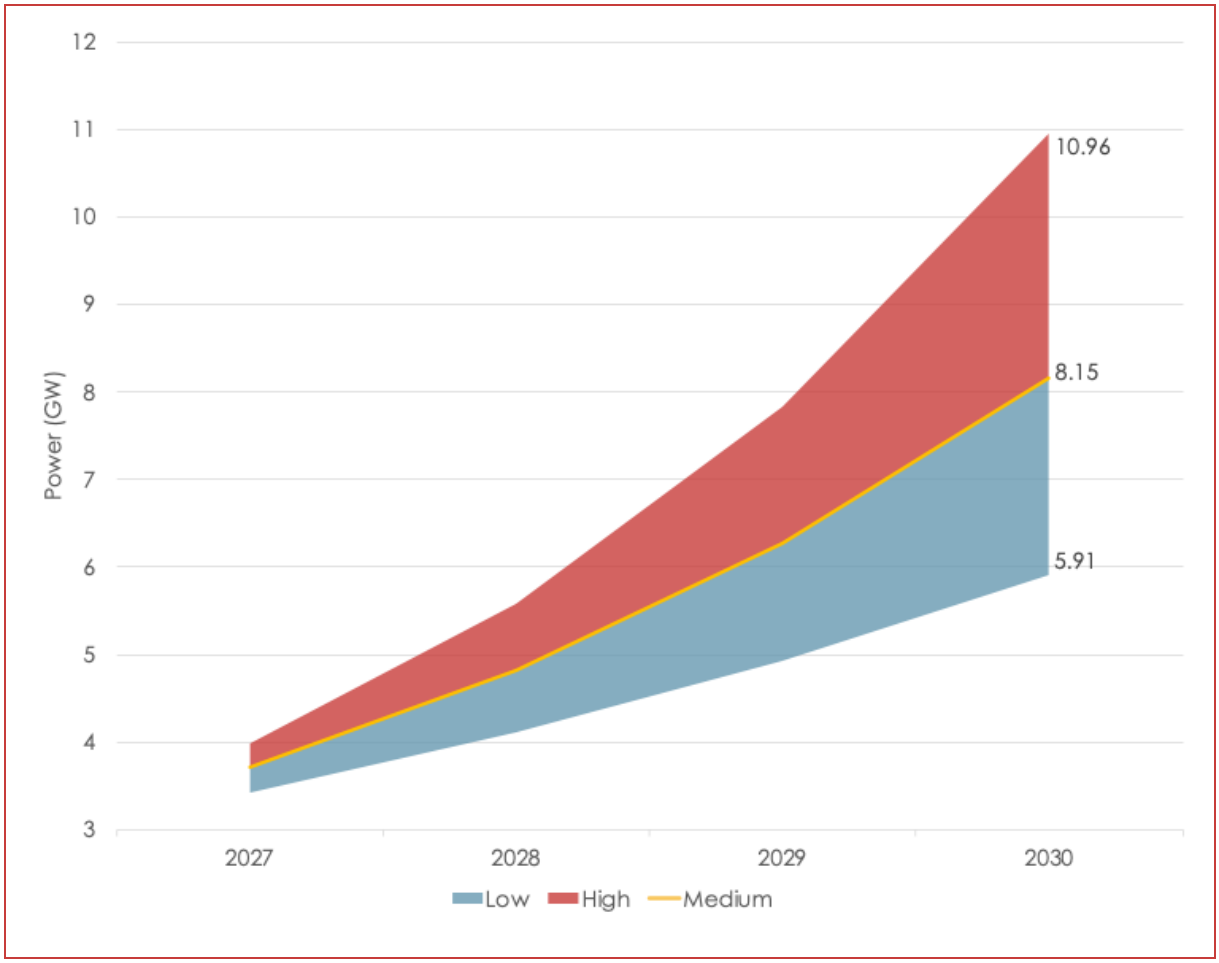
Industry estimates suggest that global data-centre power demand for all AI could increase at an annual rate of 19–27% from 2023 to 2030.[34] Separately, other sources estimate that frontier AI models may draw up to twice as much power in the future.[35] To account for uncertainty around the power demand of non-frontier models, broader data-centre usage, and GPU-level consumption, we adopt a deliberately conservative assumption of 40% annual growth for the high-growth scenario.

[34] AI power: Expanding data centre capacity to meeting growing demand, McKinsey & Company (2024), Accessible at: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand

[35] How much power will frontier AI training demand in 2030, Epoch AI (2025), Accessible at: https://epoch.ai/blog/power-demands-of-frontier-ai-training

On a GPU-card-only basis, power demand is projected to reach **5.91 – 10.96 GW** by 2030, with a medium estimate of 8.15 GW.[36] Accounting for full infrastructure and idle-time adjustments, the medium estimate rises to **15.66 GW,** with a range of **11.37 – 21.07 GW**.

Figure 12 Estimated future power demands for compute from only GPU cards



Source: Technopolis (2026). Note that these scenarios are based on a starting current power demand of 2.85 GW and uses the ~37K estimate of academic users.

To contextualise the medium unadjusted estimates of 8.15 GW: this would represent 8% of projected global compute-driven power demand in 2030 and 4% of US compute-driven power demand.[37] Based on the Department of Energy Security and Net Zero's (DESNZ) ambitions for future installed capacity of ~213GW, future demand for compute would be equivalent to

[36] Please note that we estimate future GW and FLOP demand through 2030, while GPU-hour demand is projected five years ahead, through 2031. This difference is due to availability of industry growth-rate projections, which extend only to 2030 for GPU-hours and FLOPs, and enable direct comparisons.

[37] Global data centre sector to nearly double to 200 GW amid AI infrastructure boom (JLL, 2026). Note that demand comparisons to the US are made as the US is a world-leader in demand for AI.

around 4% of that installed capacity.[38] More conservative market reports suggest that installed capacity in 2030 may be closer to ~150 GW meaning that demand for compute would instead be around 10% of installed demand.[39]

4.3 FLOP Capacity

Assessing demand for AI compute in terms of total floating point operations (FLOPs) provides a technology-neutral way of capturing the **overall volume of computation** required across different workloads. FLOPs measure the total number of mathematical operations needed to complete a task, and are widely used to quantify the compute requirements of AI models, particularly in the context of training.

Expressing demand in FLOPs enables aggregation across heterogeneous use cases and infrastructure, providing a consistent basis for comparing and projecting compute requirements. Unlike measures of compute capacity (e.g. FLOP/s), which reflect the rate at which computation can be performed, FLOPs capture the **total workload demand** placed on the system over a given period. In practice, FLOP-based estimates encompass a range of AI-related activities, including model training, fine-tuning, and inference. As such, the figures presented here should be interpreted as the **cumulative compute demand** generated across the UK ecosystem, rather than the requirements of any single model or training run.

Current demand for compute is estimated at approximately 6.57 RFLOPs.[40] Three growth rates were applied to forecast demand and capacity, drawing on observed trends in frontier model training – where compute has grown approximately 4-5x annually in the deep learning era – while applying more conservative assumptions to reflect the broader mix of workloads captured in UK demand:

- **High growth**: 350% year-on-year growth
- **Medium growth**: 250% year-on-year growth
- **Low growth**: 150% year-on-year growth

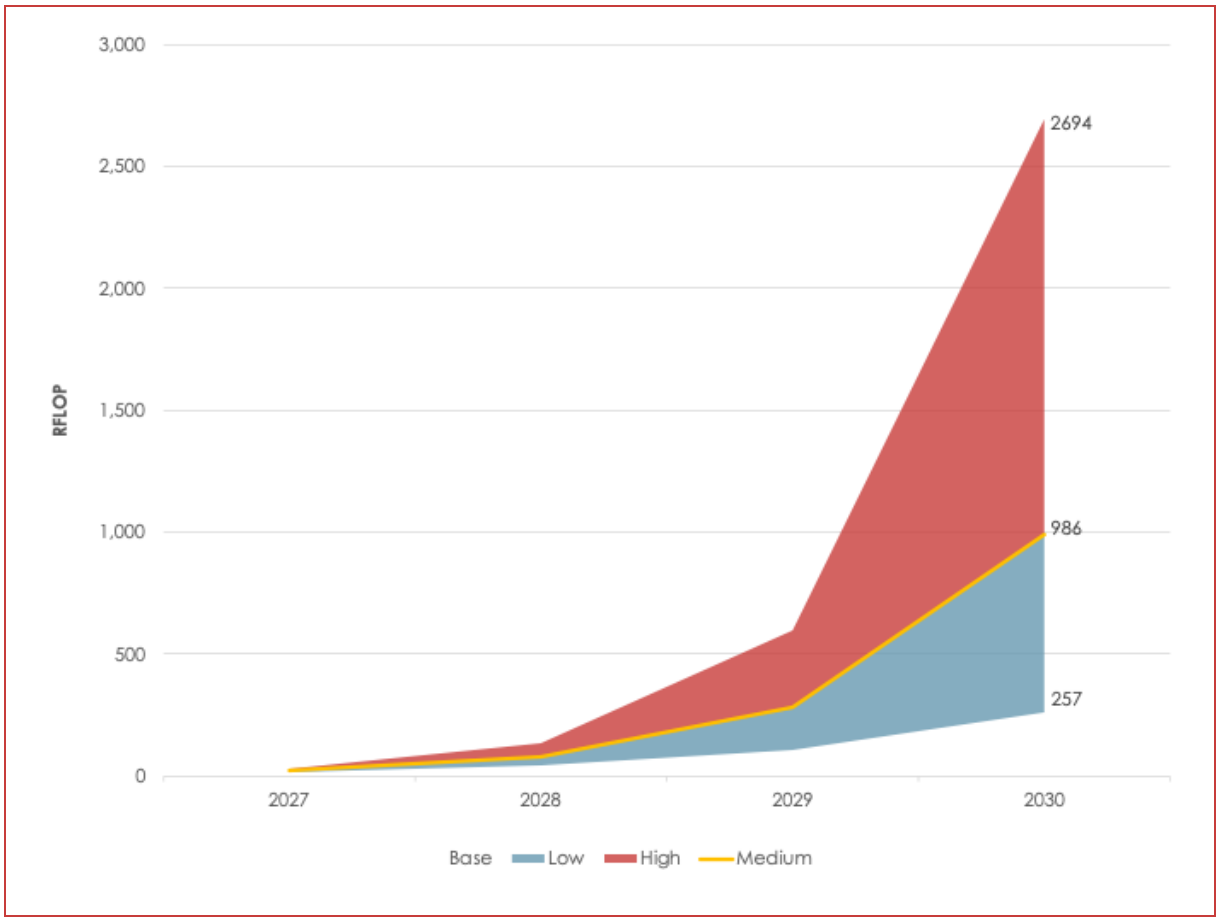
By 2030, demand for compute is projected to range from 257 RFLOPs to 2,694 RFLOPs, with a medium estimate of 986 RFLOPs (Figure 13).

[38] Clean Power 2030 Action Plan (DESNZ, 2025)

[39] United Kingdom Power Market Size & Share Analysis, Growth Trends and Forecasts (2025-2030), Mordor Intelligence

[40] 1 RonnaFlop = 10^{27} FLOP

Figure 13 Projected demand for compute (in RFLOP)



Source: Technopolis (2026). Note that these scenarios are based on a starting FLOP demand of 6.57 RFLOP and uses the ~37K estimate of academic users.

To contextualise these figures, it is important to distinguish between the compute requirements of individual frontier AI models and the aggregate demand estimated here. Frontier training runs can require extremely large volumes of compute, but these are typically concentrated in a small number of highly resourced organisations and represent discrete, one-off efforts.

By contrast, the demand estimated in this analysis reflects the **cumulative compute requirements of a large and diverse user base.** Rather than being driven by a small number of frontier-scale training runs, it captures the combined effect of many activities across the research ecosystem, including smaller-scale model development, fine-tuning, inference, and domain-specific applications.

This implies that future demand is likely to be **more distributed and heterogeneous in nature,** arising from the aggregating of compute needs across tens of thousands of users and disciplines. As a result, while frontier model trends provide a useful upper-bound reference point for growth, they are not directly representative of how compute demand is likely to materialise in the UK research context.

5 Sovereign Demand for Compute

The study captured total demand for compute, which in practice, is expected to be met through a combination of public resources (such as AIRR or any future expansions of this infrastructure) and through private sector infrastructure. A proportion of this demand will require sovereign access, specifically.

5.1 Introduction to Sovereign AI

The term "sovereign AI" is used in different ways across policy, industry, and research contexts. The rise of AI has been marked as a new technological era and it is widely understood by governments across the globe that those who deploy AI across their economies and institutions could gain a lasting advantage.

The concept of sovereign AI has risen rapidly up the policy agenda as governments grapple with the strategic implications of becoming dependent on a small number of foreign technology providers. Policy officials have recognised that AI is not simply another technological wave but a foundational governing infrastructure that will influence how states make decisions, deliver services and project strategic influence.[41] In this context, sovereignty is about more than ownership of hardware or data and instead is about agency and choice – the ability to determine and shape how AI is used domestically and to adapt as technology evolves.

The drivers of sovereign AI demand are varied. At the geopolitical level, advanced AI systems depend on scarce and highly concentrated resources including hyperscale compute, global cloud networks, semiconductors and highly skilled technical talent, all of which are inputs controlled by a small number of firms and jurisdictions.[42] This concentration creates strategic vulnerabilities for nations that rely entirely on foreign infrastructure. The choices governments make today about their AI systems will determine whether they can expand capability in the future or find themselves locked into systems that are prohibitively and politically costly to change.

The most direct constraint is legal. Data protection frameworks such as the UK General Data Protection Regulation (GDPR) place restrictions on the transfer of personal data to countries outside the UK unless adequate protections are in place.[43] While it may be economically attractive to process workloads using cloud infrastructure located in countries with lower energy costs, this is not always a viable option for organisations handling sensitive or regulated data. For research involving medical records, personal identifiable information, or other regulated datasets, sending data to overseas infrastructure may simply not be permissible,

[41] Sovereignty in the Age of AI: Strategic Choices, Structural Dependencies and the Long Game Ahead, Tony Blair Institute (2026), Accessible at: https://institute.global/insights/tech-and-digitalisation/sovereignty-in-the-age-of-ai-strategic-choices-structural-dependencies

[42] Compute sovereignty: The strategic importance of digital infrastructure, S&P Global (2026), Accessible at: https://www.spglobal.com/en/research-insights/special-reports/compute-sovereignty-strategic-importance-of-digital-infrastructure

[43] The UK's data protection legislation, HMG (2026), Accessible at: https://www.gov.uk/data-protection

regardless of the cost savings on offer. Similar restrictions apply in sectors such as defence and national security, where data sovereignty is an operational and strategic imperative rather than a matter of preference.

Sending data within the UK, by contrast, allows organisations to remain within a single legal and regulatory jurisdiction, reducing compliance complexity and risk. Domestic infrastructure also offers greater transparency and accountability — organisations can more readily audit where their data is stored, who has access to it, and under what legal framework it is protected.

Within this broad framing, sovereign compute demand can be understood to include:

- Research relating to national security or strategically important sectors
- Research or use cases involving data subject to regulatory compliance requirements (e.g. data protection frameworks such as GDPR)
- Research that relies on other forms of sensitive data, including medical data or projects with intellectual property and confidentiality constraints

5.2 Measuring Sovereign Demand for AI

Translating the concept of sovereign AI into a measurable demand signal presents several methodological challenges. As set out above, sovereignty in the AI context is not a fixed or universally agreed standard and encompasses a spectrum of considerations around data governance, jurisdictional control, infrastructure ownership, and strategic autonomy. This breadth makes it difficult to operationalise as a survey construct.

This introduces a risk that respondents interpret the term differently, leading to variation in how they assess their own requirements. Second, the need for sovereign access is often context-dependent: organisations may require sovereign compute for specific projects or datasets, rather than as a general or constant requirement. Third, considerations such as data sensitivity, regulatory compliance, and security constraints may overlap, making it difficult for respondents to clearly distinguish "sovereign" needs from broader secure or compliant compute requirements.

As a result, responses relating to sovereign compute should be interpreted with caution. Rather than representing a precise or uniform category of demand, they are better understood as indicative of a subset of use cases where control over infrastructure, data governance, and jurisdiction is particularly important. In practice, this suggests that sovereign demand is likely to be heterogeneous, episodic, and closely tied to specific applications, rather than a single, clearly bounded segment of overall compute demand.

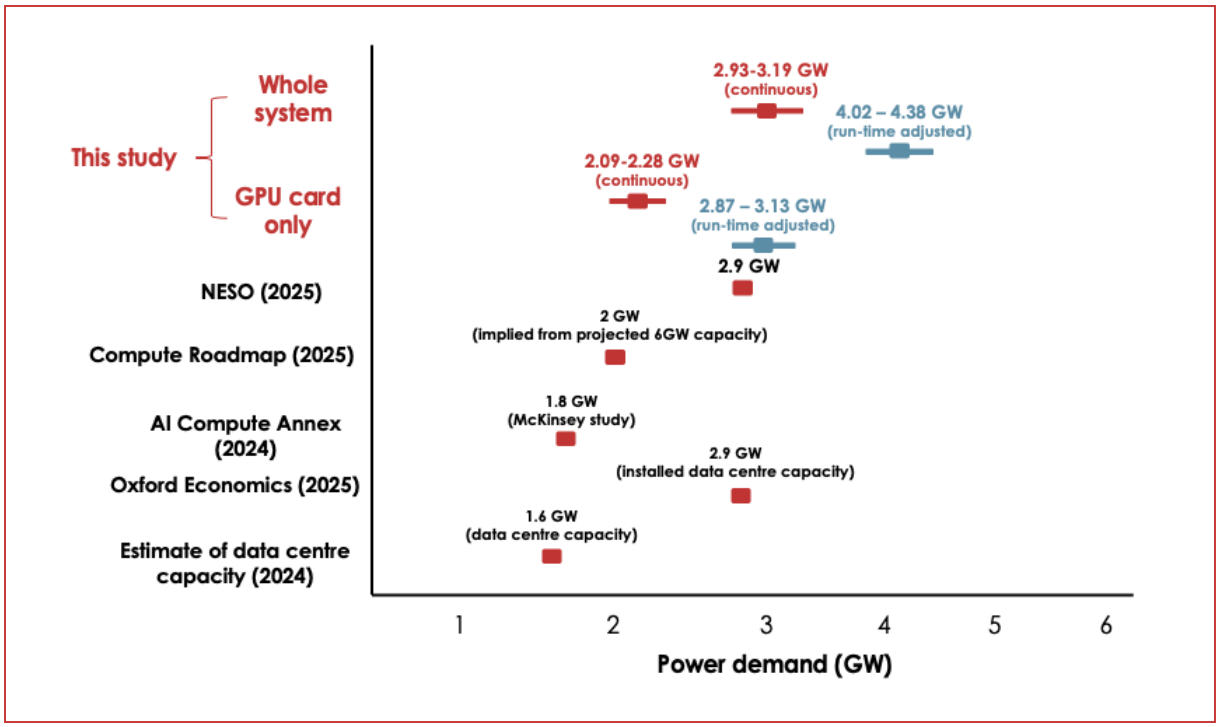
5.2.1 Trusted Research Environments

The approach to calculating sovereign demand for compute draws on user requirements for Trusted Research Environments (TREs). TREs are secure, controlled computing environments that allow researchers to access and analyse sensitive data without that data leaving a protected infrastructure. These environments are required for a range of reasons, including the use of sensitive medical data or IP confidentiality constraints. TRE usage is therefore used as the primary measure of sovereign compute demand – it provides a strict and reliable proxy, capturing demand that is formally bound to UK-hosted infrastructure by data sensitivity, regulation or institutional requirement.

Reported TRE usage varied considerably by sector. Among industry respondents, 83% (n=24) reported requiring TRE access, as did 82% of public sector respondents (n=11). Academic

respondents reported markedly lower usage, at 38% (n=111). The study has then used this distribution of TRE requirements to project future demand for sovereign compute access. Sovereign demand was estimated by sector, calculated as the product of each sector's requirements for TRE's and its current compute demand. Aggregate sovereign demand was then derived by summing across the three sectors. Figure 14 shows estimates of sovereign power demand in the context of other recent studies.

Figure 14 Estimates of sovereign power demand compared to other studies

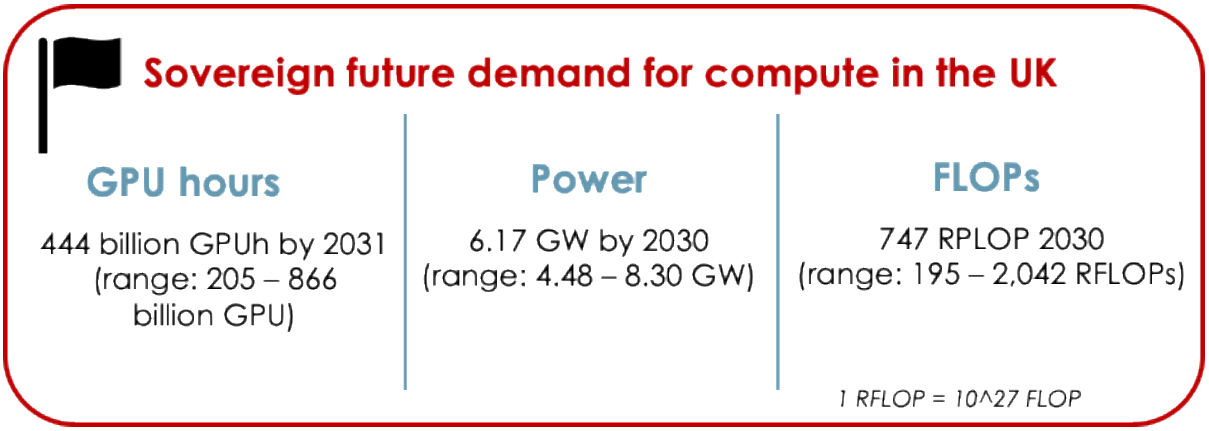


Source: Technopolis (2026)

5.2.2 *Future sovereign demand*

Using TRE usage as the primary proxy, and applying the same growth rates as Section 5.1, future sovereign demand is estimated as follows for the medium growth scenario:

Figure 15 Sovereign future demand for compute

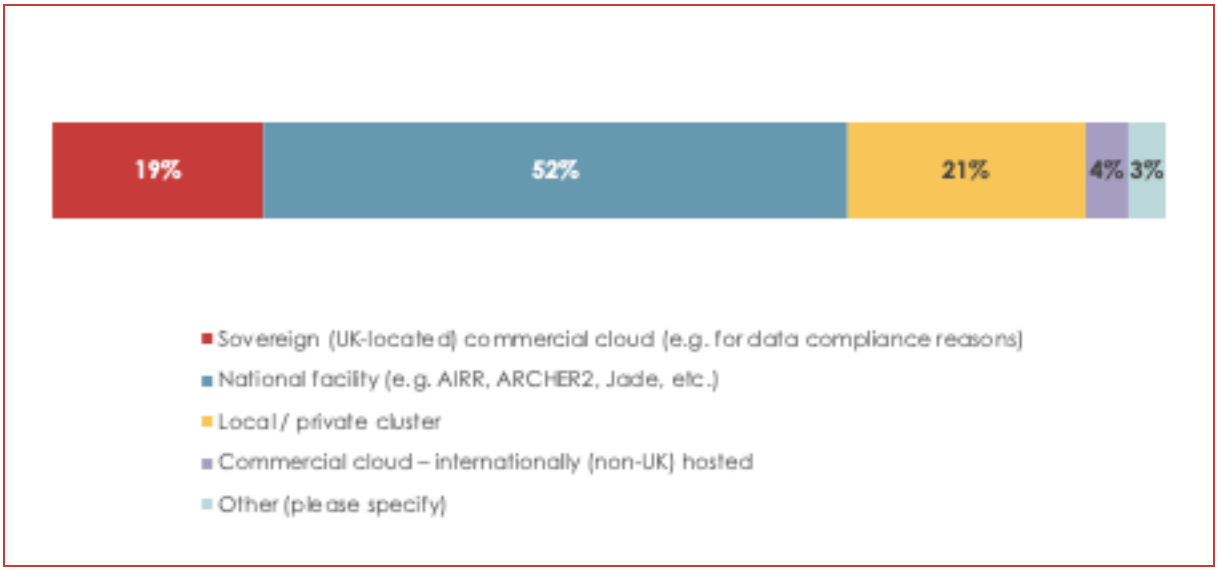


Based on these estimates, sovereign demand accounts for approximately 76% of total projected compute demand. This figure should be treated with caution, given the small samples of industry and public sector respondents, both of which reported very high TRE usage. On the one hand, increasing availability and usability of TREs alongside growing confidence in secure data access frameworks, may enable a wider range of users and workloads to operate within sovereign-compliant settings. This would tend to increase the share of demand classified as sovereign. On the other hand, changes in regulatory requirements, data governance standards, or organisational risk tolerance could either expand or constrain the types of data and workloads that must be processed within sovereign environments. Regardless of the precise share, this level of demand is unlikely to be met through national facilities alone, and will require a broader ecosystem of sovereign-compliant infrastructure, although the scope of this study does not make any suggestions as to the way in which this demand could or should be met (i.e. this is not a supply-side study).

5.3 Preferred Access Models

Another approach to conceptualising sovereign demand for AI draws on respondents' preferred compute access model, distinguishing between on-shore capacity (local/private clusters), national facilities and sovereign UK-hosted cloud) and off-shore capacity (internationally hosted commercial cloud). Compared to TRE usage, this is a 'softer' measure of sovereign demand; it captures preferences rather than formal requirements, and is therefore treated as a secondary measure of sovereign demand. We have not used this measure to calculate or quantify sovereign demand, but instead to provide context as to how users think about their own usage patterns and to highlight the challenges in accurately capturing sovereign demand.

Figure 16 Survey respondents' preferred access model (n=126)



Source: Technopolis survey of UK HPC users (2026)

Overall, the majority of survey respondents expressed a preference for access via national facilities (52%), followed by local / private clusters (21%) and sovereign (UK-located)

commercial cloud (19%). This distribution is likely influenced by the composition of the respondent sample, which includes a significant proportion of academic users who are more accustomed to (and potentially more reliant) on publicly funded compute infrastructure compared to industry users.

Qualitative responses about why this was viewed as a preferred access model reveal a nuanced picture, with preferences shaped by a range of practical, regulatory and strategic considerations rather than a single dominant concern, including data security and regulatory compliance (as with TREs). Additional motivations for choosing various access models includes:

- **Cost and sustainability:** A recurring theme across responses (from both SMEs and academics) was the perceived long-term cost advantage of national facilities over commercial cloud. Several respondents noted that while commercial cloud offers flexibility, it becomes "more costly" over time as individual facilities must remain up to date with the latest requirements, and that national infrastructure (where compute is pooled and centrally managed) offers better value. It should be noted however, that AIRR is free to access for researchers and this may have skewed responses on costs.

"I think building national facilities for public access is more sustainable compared to distributed compute in siloes." - Survey Respondent

"Allows for greater pooling resources for a fast, state-of-the-art facility. Local/private clusters can be costly and/or unreliable." - Survey Respondent

"National facilities are essential to academia, given the increasing pressure on university finances. Smaller institutions cannot compete globally if they don't have access to world class infrastructure and they simply don't have the budget to build and maintain them. National infrastructure is far more cost effective, even if this requires buy-in." - Survey Respondent

- **Flexibility and accessibility:** Commercial cloud providers were acknowledged for their advantages in flexibility, availability, and access to the latest hardware — often being "first to bring on new chips." Local and private clusters were similarly valued for their responsiveness and the ability to support interactive, iterative research workflows. Several respondents noted tension between the flexibility of local or commercial options and the scale and reliability of national facilities, with one describing UK-based cloud as "the sweet spot" for fast-moving research environments.

"UK-located commercial cloud provides the flexibility to scale compute on demand while meeting data residency, security, and regulatory requirements. It supports bursty analytical and forecasting workloads without the overhead of maintaining dedicated on-premise infrastructure." - Survey Respondent

"We don't have power for local systems of the size we need. We need [facilities] UK-based because of HMG security requirements. National facilities are not set up for 24/7 resilient operations with sufficient bounds on security for defence customers." - Survey Respondent

- **Support for national facilities:** Despite the challenges respondents reported with accessing national facilities (for example high and competitive application rates, unclear eligibility for SMEs), there was broad support for their role in the research ecosystem. Respondents emphasised their importance for democratising access, particularly for smaller institutions unable to fund their own infrastructure. Positive experiences with AIRR (and Isambard-AI in particular) were cited by multiple respondents, with one describing AIRR infrastructure as "remarkable" and unlike anything encountered across other countries. National facilities were also seen as essential for reducing dependency on international commercial providers and advancing UK compute sovereignty.

> **"Having a national facility with dedicated personnel is essential. It becomes easier to maintain and sustain over longer periods of time. The EuroHPC model seems to be a very effective way to support this." - Survey Respondent**

- **Sovereignty and strategic independence:** A number of respondents framed their preferences in explicitly strategic terms, expressing concern about vendor lock-in and the risks of relying on international cloud providers. Several noted that the UK "cannot control" international commercial infrastructure, and that sovereign capacity is essential not only for data security but for long-term national competitiveness in AI.

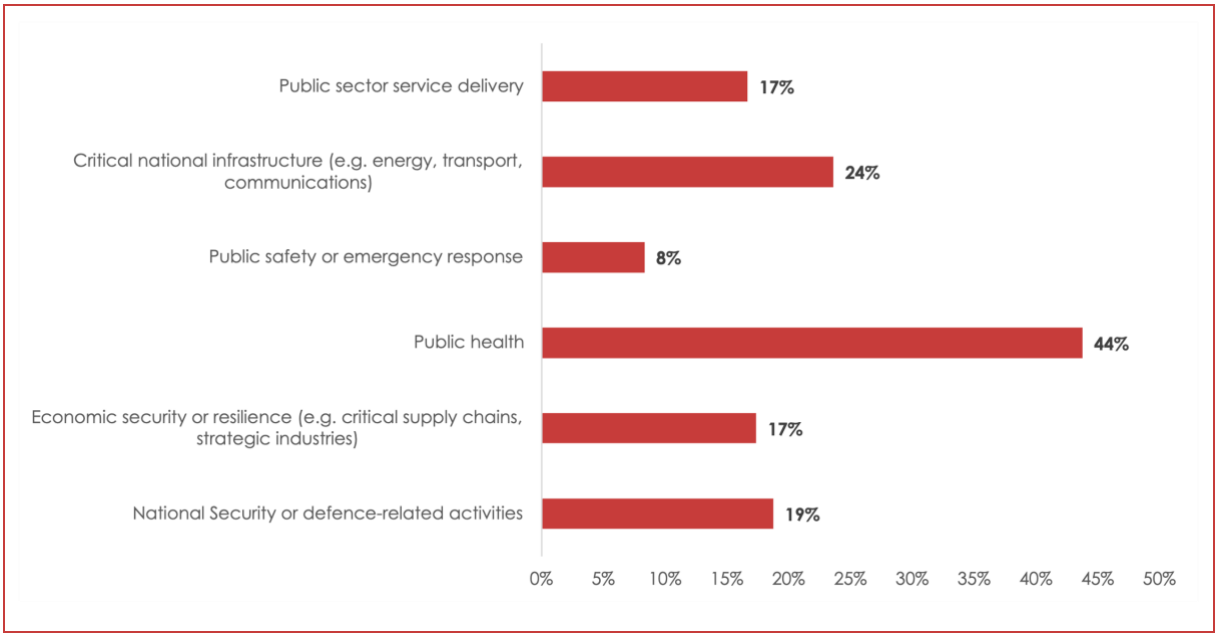
> **"We want UK residency [for our data] and we don't want to rely on international commercial cloud companies which the UK cannot control." - Survey Respondent**

While the access model preferences outlined above provide a useful indication of demand for on-shore compute, they should be treated with caution as a measure of sovereign need. Unlike TRE requirements, access model preferences capture a broader and more heterogeneous set of motivations. Some of these, such as regulatory compliance and data protection obligations, do constitute genuine sovereign requirements. However, others, particularly those relating to strategic independence, national competitiveness, or a general sense of security, reflect considered preferences rather than hard constraints. A researcher who "feels safer" using UK-hosted infrastructure, or who favours national facilities on principled grounds, may not be in a position where sovereign access is strictly necessary. This distinction matters: conflating preference with requirement risks overstating the share of demand that must be met through sovereign infrastructure, as opposed to the share that would benefit from it, although this may still be a consideration for policy-makers when conceptualising sovereign AI demand. For planning purposes, TRE usage therefore remains the more reliable lower-bound estimate of sovereign compute demand, while the preferred access model provides a useful (if broader) upper bound.

Although it was not used for modelling purposes, respondents were also asked to map their current research against key areas that might also be considerations for sovereign access.

Public health was the most frequently reported research area (44% of respondents), followed by critical national infrastructure (24%).

Figure 17 Respondents' primary research application area (n = 157)



Source: Technopolis analysis of survey (2026). Note that respondents could choose multiple answers.

5.4 Demand for low-latency compute

A further dimension of future demand concerns latency sensitivity. Latency refers to the delay between a system receiving data and producing a response, typically measured in milliseconds. For many AI workloads, this is inconsequential, but for a growing class of applications, even small lags can degrade performance, introduce safety risks, or fundamentally break the user experience.

It is useful to distinguish workloads by their sensitivity to delay. Training workloads (focused on developing and refining AI models) are largely insensitive to latency and can tolerate delays of 100 milliseconds or more.[44] Because no real-time interaction is required, these can be sited in remote or power-rich locations where grid capacity is more readily available, without meaningful performance penalty.

Inference workloads, by contrast (where a trained model is queried to generate outputs) are increasingly distributed to be as close to the end user or application as possible.[45] As AI

[44] The next big shifts in AI workloads and hyperscaler strategies, McKinsey & Company (2025), Accessible at: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-next-big-shifts-in-ai-workloads-and-hyperscaler-strategies

[45] Understanding AI deployment methods and locations, Data Centre Dynamics (2025), Available at: https://www.datacenterdynamics.com/en/opinions/understanding-ai-deployment-methods-and-locations/

becomes embedded in consumer and enterprise products, even modest delays perceptible to users (generally above 200–300 milliseconds) can meaningfully degrade experience.

Beyond inference, several categories of application place still more demanding requirements on latency. Real-time decision-making and control systems, including autonomous vehicles, industrial robotics, and algorithmic trading platforms, require near-instantaneous processing of sensor or market data to support split-second decisions. [46] In these contexts, delay is not just an inconvenience: it can create physical safety risks or result in significant financial losses. Live data processing and streaming applications, which continuously ingest and analyse data as it is generated, place similar demands on infrastructure. [47] Human-in-the-loop systems, where humans supervise or intervene in AI workflows, depend on sufficiently fast response times to make that oversight meaningful in practice. [48] Interactive AI systems (those that respond dynamically to users, environments, or simulations) degrade noticeably when latency pushes response times beyond what feels immediate to an operator or end user.[49]

The practical implication is that as AI moves from back-office computation into operational and consumer settings, the geographic proximity of compute infrastructure becomes a material constraint. Relying on distant or overseas data centres introduces network latency that cannot easily be engineered away. Looking further ahead, smaller edge data centres connected by high-speed networks may help to distribute compute closer to where it is needed, reducing both latency and bandwidth demands. In the near term, however, demand for low-latency AI applications is likely to strengthen the case for domestically available compute capacity.

In total, 17% of survey **respondents reported that low latency is critical to their work, while a further 35% indicated that it is very important**. Criticality varies by user type, as shown in Figure 18. The findings suggest that that latency is less of an issue to academic users than those in industry or the public sector, with only 45% reporting latency to be "critical" or "very important", compared to 70% and 71% respectively. This likely reflects differences in activities between sectors - academics allocate a higher proportion of time to training than industry or the public sector (Figure 10). AI/ML training is typically batch-orientated and is focussed on maximising throughput (total work completed over long durations). Inference is often focussed on minimising response time (how quickly an output is produced after an input is received)[50] [51].

[46] What is a real-time system?, Intel. Available at: https://www.intel.com/content/www/us/en/learn/what-is-a-real-time-system.html

[47] What is stream processing? IBM. Available at: https://www.ibm.com/think/topics/stream-processing

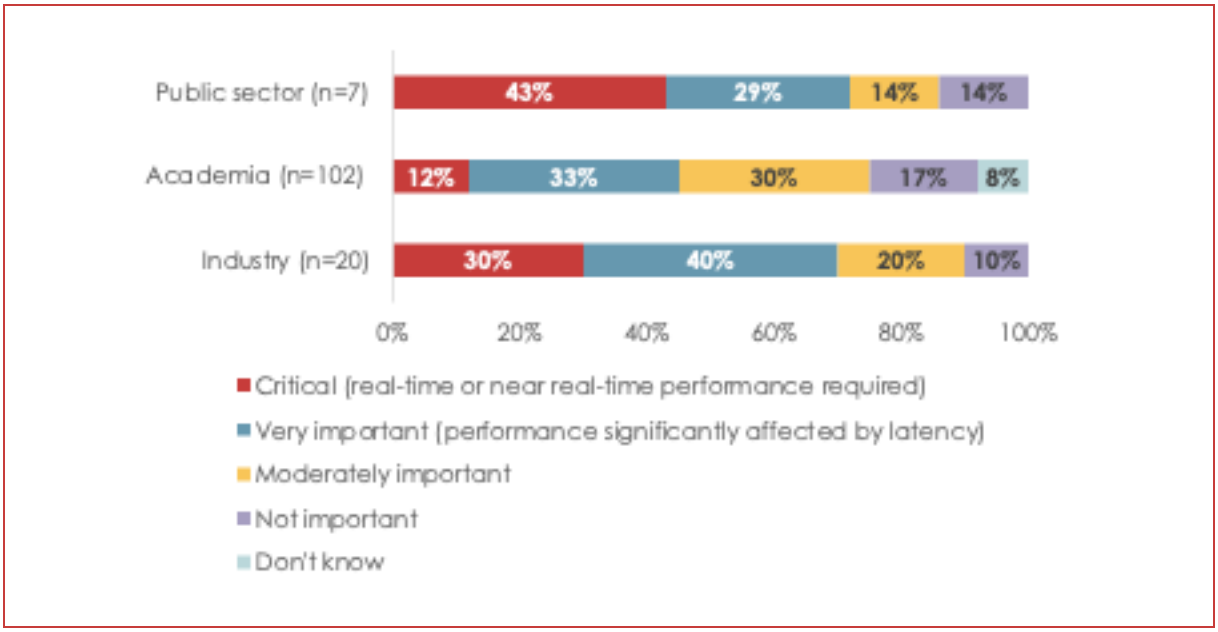
[48] What is human-in-the-loop? Stanford University HAI. Available at: https://hai.stanford.edu/ai-definitions/what-is-human-in-the-loop

[49] Interactive AI: The definition, use case, and relevance for enterprises, Iterate.ai. Available at: https://iterate.ai/ai-glossary/what-is-interactive-ai

[50] AI inference vs training: Key differences explained, Digital Ocean. Available at: https://www.digitalocean.com/resources/articles/ai-inference-vs-training

[51] Not all inference is latency sensitive. Batch inference tends to have similar constraints to training, where throughput is prioritised over response time.

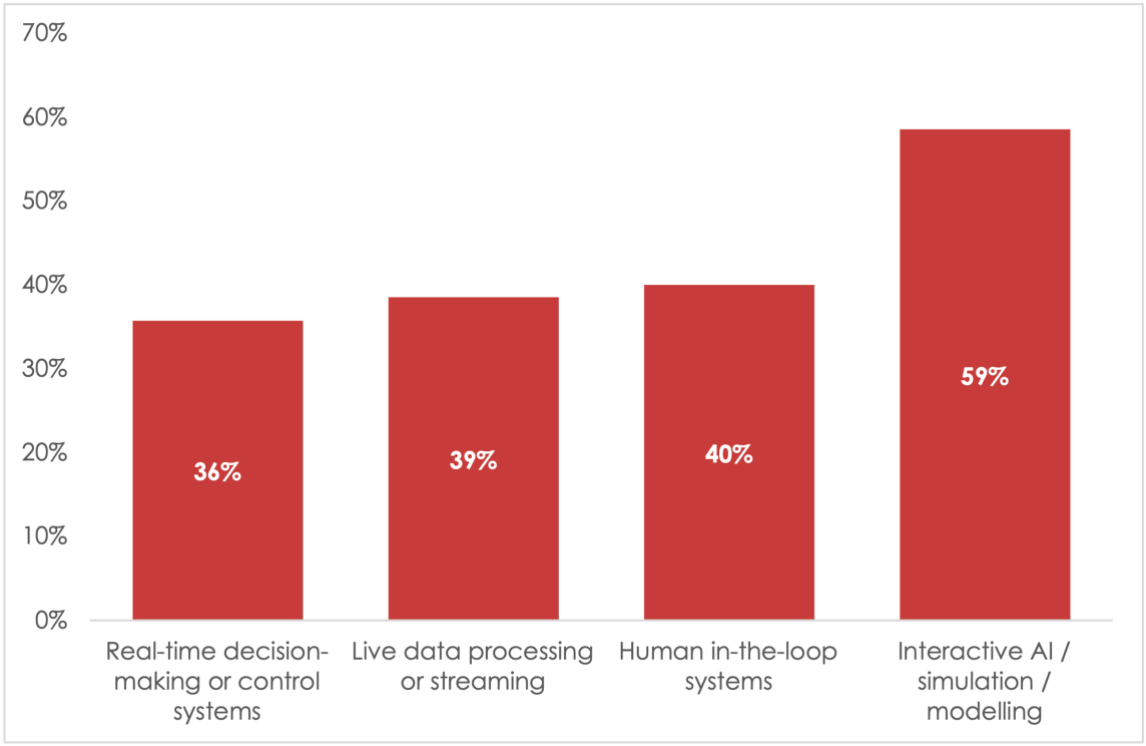
Figure 18 User latency requirements (n=134)



Source: Technopolis analysis of survey (2026). Note 'n' here refers to the number of respondents who reported their latency requirements.

To mitigate the risk of over-reporting (i.e. respondents defaulting to stating that low latency is important), the survey also asked respondents to identify the underlying drivers of these requirements. The results show that the need for interactive AI applications and simulation modelling is the most frequently cited driver (Figure 19). These types of workloads typically require rapid feedback loops, real-time processing, or user interaction, all of which are highly sensitive to delays in compute response times.

Figure 19 Drivers for latency requirements (n=121)



Source: Technopolis analysis of survey (2026). Survey question: "What are the primary drivers of your low-latency requirements? (Select all that apply)". Note 'n' here refers to the number of respondents that reported their low-latency requirements. Real-time decision making or control systems are systems that process incoming data and generate decisions or actions to support immediate or time-critical operation. Examples include autonomous vehicles, algorithmic trading or AI-enabled manufacturing systems[52]. Live data processing or streaming refers to the continuous processing and analysis of data as it is generated, enabling real-time insights or system actions[53]. Human in-the-loop systems are systems in which humans participate in the operation, supervision, or decision-making of AI workflows[54]. Interactive AI refers to systems that dynamically interact with users, environments, or simulations in near real time, rather than running as static batch processes.

5.5 Drivers of future demand

In addition to quantitative estimates of future demand, the survey explored stakeholders' views on the factors shaping future demand for AI compute. Respondents were asked to rate the importance of a set of potential drivers of demand on a scale from 1 ("not a driver") to 5 ("a

major driver"). Table 7 presents the average scores of each of the drivers based on the scenario for future growth that respondents chose.

Table 8 Drivers of future demand for AI compute

	Exponential year-on-year growth (n=72)	Continued year-on-year growth (n=86)	Demand largely plateaus (n=7)	Demand declines (n=4)	Significant downturn in demand (n=4)
Scaling of AI models and algorithms (use of more compute-intensive models)	**4.5**	**4.2**	**2.9**	**4**	**4.3**
Rapid growth in AI inference and real-world deployment / mass adoption of generative AI by consumers and enterprises	**4.3**	**4**	**3.7**	**4**	**4**
Rapid advancements in open-weight model capabilities	**3.9**	**3.6**	**3.3**	**3.8**	**3.3**
Industry digital transformation and automation needs - demand for real-time forecasting, decision-making and process automation	**3.7**	**3.5**	**3.5**	**3.3**	**3.3**
Growth in data availability and complexity - larger datasets (e.g. weather, economic, behavioural data)	**4**	**4**	**3.4**	**3.8**	**3.5**
Greater access to GPUs (advances in availability of infrastructure)	**4.1**	**4**	**3.8**	**4.3**	**3.8**
Economic incentives and expected returns from AI (anticipated productivity gains)	**3.9**	**3.5**	**3.4**	**4**	**4**
Competitive and strategic pressures (race to build proprietary AI capabilities and infrastructure)	**4.1**	**3.9**	**4**	**3.5**	**3.3**
Policy / geopolitical factors (national AI strategies, industrial policy)	**3.4**	**3.3**	**3.2**	**2.8**	**3.3**

Energy, power, and infrastructure constraints (rising electricity demand from AI workloads, slowing efficiency gains)	**3.2**	**3.1**	**3.6**	**2**	**2.8**
Driver difference (for a given scenario)	**1.3**	**1.1**	**1.1**	**2.3**	**1.5**

Source: Technopolis survey of UK HPC users (2026). Values presented in the table are average scores across different demand scenarios and 'n' values refer to the number of respondents who ranked a given scenario either 1 or 2 (likely scenarios). In total, 115 respondents ranked scenarios.

Despite selecting different demand trajectories (e.g. exponential growth, continued growth, plateaus), respondents identified a largely consistent set of underlying drivers, as shown by the marginal difference in drivers for a given scenario (bottom row). This suggests that divergence in expectations about future demand is not driven by fundamentally different beliefs about *what* is shaping demand, but rather by differing interpretations of *how these drivers will evolve and interact* in practice.

Across all scenarios, the scaling of AI models and algorithms (particularly the increasing use of compute-intensive models) was consistently identified as a major driver of demand. This was closely followed by the rapid growth in AI inference and real-world deployment, including the widespread adoption of generative AI by both consumers and enterprises. Similarly, increased access to GPUs and improvements in infrastructure availability were seen as key enablers of rising demand.

Other factors, including the growth in data availability and complexity, competitive pressures to develop proprietary AI capabilities, and broader economic incentives linked to expected productivity gains, were also rated as important across most scenarios. In contrast, policy and geopolitical factors, as well as energy and infrastructure constraints, were generally viewed as less influential, although still relevant.

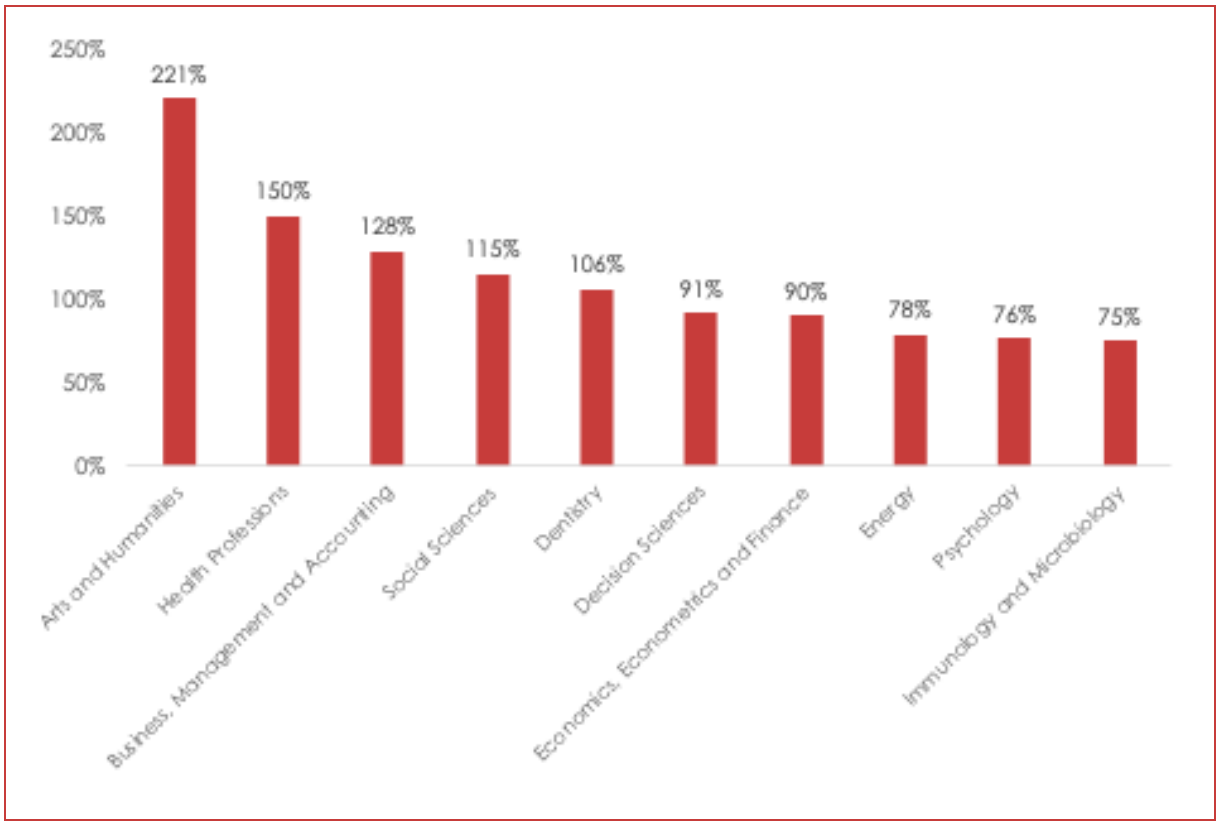
Importantly, the consistency in identified drivers across scenarios indicates that uncertainty about future demand is less about disagreement over the direction of travel and more about the magnitude and pace of change. For example, respondents who anticipate plateauing or declining demand still recognise the importance of model scaling and AI deployment, but may place greater weight on countervailing forces such as efficiency gains, cost constraints, or market saturation.

A small number of respondents selected scenarios in which demand declines (including plateau, decline, significant downturn), making interpreting the rating of potential drivers challenging. This reflects a key limitation of understanding root causes through a survey. Future work may consider additional qualitative work, such as foresight studies, to understand these scenarios and the pathways by which these drivers are expected to impact HPC for AI's demand trajectory.

Disciplinary trends in AI adoption

While the survey provides insight into the drivers shaping future demand, the bibliometric analysis offers an indication of where this demand may emerge across the research landscape. Specifically, the analysis highlights the disciplines in which AI adoption (and by extension, potential compute demand) is growing most rapidly (Figure 20).

Figure 20 Top 10 fastest growing disciplines that mention use of AI (2021-2025)



Source: Technopolis analysis of over 700,000 publications In OpenAlex (2026)

The results show that growth is led by the Arts and Humanities (221%), followed by Health Professions (150%), Business, Management and Accounting (128%) and the Social Sciences (115%).

The prominence of these disciplines should be interpreted with some caution, as it likely reflects both actual growth in demand and the relatively low baseline, which can inflate percentage growth rates. Beyond the top-ranked disciplines, growth rates converge within a narrower band (approximately 75% - 110%), suggesting a more even diffusion of AI across a wide range of fields. This indicates that future demand for compute is unlikely to be concentrated in a small number of domains, but instead distributed across an increasingly diverse set of disciplines.

technopolis group

www.technopolis-group.com