

Hard facts. Clear stories.

Copenhagen  
Economics

CE

# COMPUTE FOR GEN AI

Assessment of competitive conditions

CLIENT: AMAZON WEB SERVICES (AWS)  
AUGUST 2025

#### **ABOUT COPENHAGEN ECONOMICS**

Copenhagen Economics is an expert-driven consulting company built on a deep knowledge of applied economics, and one of the leading economics firms in Europe.

We believe that sound economic analysis can equip decision makers with hard facts and clear stories to make better choices for the benefit of society.

#### **A brief note on consultancy research**

As is standard in our field of professional services, research is designed so that:

- the client chooses the research question;
- we analyse and address the question to the best of our knowledge;
- findings and conclusions are our own.

For further information, see [www.copenhageneconomics.com](http://www.copenhageneconomics.com). We remain available for and appreciate any questions or comments.

### EVIDENCE SUGGESTS THAT COMPUTE PROVIDERS FOR GEN AI OPERATE IN AN INCREASINGLY COMPETITIVE ENVIRONMENT



#### MANY PROVIDERS OF COMPUTE FOR GEN AI

Evidence suggests that there are **> 100 providers of compute for Gen AI**, including new AI-specialised providers (e.g., CoreWeave, Cerebras, Lambda) and on-premises providers<sup>1</sup>



#### BUYERS MAKE USE OF MULTIPLE OPTIONS

Buyers often use various options for accessing compute for Gen AI workloads, including **multiple cloud providers or a combination of on-prem and cloud services**



#### SEVERAL EMERGING PLAYERS ARE GROWING FAST

Several emerging AI-specialised players and on-premises providers are experiencing **triple-digit revenue growth**



#### INDICATIVE EVIDENCE SUGGESTS DECLINING PRICES

Available evidence indicates that the inflation-adjusted **price of GPU processing power has decreased** over the recent years



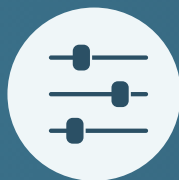
#### SUBSTANTIAL FURTHER INVESTMENT IS EXPECTED

Total worldwide **investment in data centres** is forecasted to reach **USD 1 trillion annually by 2029**, half of which is expected to support AI workload<sup>2</sup>



- Competitive conditions could vary between users (e.g. developers, deployers) and types of Gen AI workload (pre-training, fine-tuning, and inference). For example, pre-training requires more intensive workloads than inference.
- Further research would be required to capture any differences in competitive conditions.

### VARIOUS EXPECTED DEVELOPMENTS MAY FURTHER STRENGTHEN COMPETITION



SHIFT OF DEMAND TOWARDS **LESS COMPUTATIONALLY INTENSIVE WORKLOADS** (FINE-TUNING AND INFERENCE)



ONGOING INNOVATION IN COMPUTE USAGE TOWARDS **SMALLER, MORE EFFICIENT MODELS, AND ON-DEVICE INFERENCE**

POLICYMAKERS SHOULD CAREFULLY ASSESS COMPETITIVE CONDITIONS BEFORE CONSIDERING ANY POTENTIAL INTERVENTION

The research underlying this report was carried out between April and June 2025. Given the fast-moving nature of this market, subsequent developments may not be fully reflected.

## TABLE OF CONTENTS

---

<b>Executive summary</b>	<b>5</b>
<b>1 Background</b>	<b>10</b>
1.1 <b>Compute is an important input to Gen AI</b>	<b>11</b>
1.2 <b>Some policymakers have raised competition concerns</b>	<b>15</b>
<b>2 Evidence suggests that providers of compute for Gen AI operate in an increasingly competitive environment</b>	<b>17</b>
2.1 <b>There is a large and increasing number of providers of compute for Gen AI</b>	<b>18</b>
2.1.1 There are many players and specialised new entrants offering compute solutions for Gen AI	18
2.1.2 Gen AI developers and deployers frequently use multiple viable alternatives for compute	21
2.2 <b>Providers of compute for Gen AI are growing across the board</b>	<b>25</b>
2.3 <b>Substantial investment in compute capacity</b>	<b>28</b>
2.4 <b>Evidence suggests declining prices and an expanding service offer</b>	<b>30</b>
<b>3 Various expected developments may further strengthen competition</b>	<b>35</b>

<b>3.1</b>	<b>Industry trends suggest demand is shifting towards less computationally intensive workloads</b>	<b>35</b>
<b>3.2</b>	<b>Innovation in compute usage</b>	<b>38</b>
3.2.1	Smaller models are becoming more capable and inference costs are falling	38
3.2.2	Model efficiency is driven by continued innovation in model design and hardware	40
3.2.3	On-device is emerging as a viable alternative for some inference workloads	42
<b>4</b>	<b>Concluding remarks</b>	<b>43</b>
	<b>References</b>	<b>44</b>

## EXECUTIVE SUMMARY

---

- (1) According to a report by the European Commission’s Joint Research Centre (JRC), Generative Artificial Intelligence (hereafter ‘Gen AI’) has the potential to deliver “*substantial productivity gains*”, and to “*transform industries across the EU, acting as a critical driver of innovation and economic transformation*”.<sup>1</sup>
- (2) Compute is a critical input to Gen AI. The term refers to specialised hardware, such as ‘accelerators’<sup>2</sup> (i.e. high-performance processors or chips, such as Graphic Processing Units, GPUs), that can perform many operations in parallel. Access to such specialised hardware is necessary both for developing, i.e. designing and building, new Gen AI models<sup>3</sup>, as well as deploying, i.e. adapting and running, Gen AI models in end-user applications. Compute can be accessed through various means, including cloud solutions and on-premises infrastructure.
- (3) There are three types of Gen AI ‘workloads’ which have distinct compute needs:
  - **Pre-training** refers to the workloads associated with building a new Gen AI model from scratch. This is the most computationally intensive type of workload, often requiring the prolonged and parallel use of accelerators.
  - **Fine-tuning** refers to the workloads associated with adapting an existing Gen AI model to specific domains or tasks. Fine-tuning is also computationally intensive but, compared to pre-training, typically requires the use of fewer and/or less powerful accelerators for a shorter period of time.
  - **Inference** refers to the workloads associated with using Gen AI models, i.e. the workloads that are generated when running the model to produce outputs in response to user inputs. Inference workloads are significantly less computationally intensive per task. However, these workloads are recurring, and widespread use of a Gen AI model can lead to significant cumulative compute needs.
- (4) It is estimated that projected growth of Gen AI will lead to increase in demand of 39 per cent per year on average for compute for Gen AI between 2023-2030.

---

<sup>1</sup> Abendroth Dias, K., et al. (2025). Generative AI Outlook Report - Exploring the Intersection of Technology, Society and Policy. Publications Office of the European Union, Luxembourg. Pages 54 and 44. Available [here](#).

<sup>2</sup> An accelerator is a dedicated processor, chip or hardware unit optimised to perform a high number of calculations often in parallel and to accelerate specific workloads—such as AI training, inference, or graphics rendering. Examples of accelerators are Graphic Processing Units (GPUs) originally designed for graphic rendering tasks but now used extensively for Gen AI model training and Tensor Processing Units (TPUs) designed for large matrix calculations, or Neural Processing Units (NPUs) designed for AI inference on devices.

<sup>3</sup> In this report, we use the term ‘Gen AI models’ to refer to so-called foundation models, such as OpenAI’s GPT-4 or Meta’s Llama, which are large-scale models trained on broad datasets and capable of performing a wide range of tasks.

- (5) Some competition authorities and policymakers, such as the European Commission<sup>4</sup> and the UK's Competition and Markets Authority (CMA)<sup>5</sup>, have raised concerns about competitive dynamics in the provision of compute for Gen AI. In particular, there is a concern that a large share of compute capacity is concentrated in the hands of a few vertically integrated providers,<sup>6</sup> potentially limiting access for third parties and hindering Gen AI adoption.<sup>7</sup> To date, concerns have been framed as forward-looking risks that warrant continued monitoring as the market develops.
- (6) Against this backdrop, Amazon Web Services (AWS) has commissioned Copenhagen Economics to examine, based on available evidence, (i) whether providers of compute for Gen AI operate in a competitive environment and (ii) how expected developments in the sector may affect competition going forward.
- (7) In this report, we provide a preliminary assessment based on publicly available information. We assess whether market outcomes, such as trends relating to market entry, growth, investment, and pricing, are broadly consistent with a well-functioning market. Although available evidence is not granular enough to capture potential differences across workloads or buyer types, it is sufficient to support a high-level assessment. The research underlying this report was carried out between April and June 2025. Given the fast-moving nature of this market, subsequent developments may not be fully reflected.
- (8) We reach two main findings: (i) first, **providers of compute for Gen AI operate in an increasingly competitive environment**, and (ii) second, **various expected developments may further strengthen competition**.

### **Evidence suggests that providers of compute for Gen AI operate in an increasingly competitive environment**

- (9) First, **there is a large and increasing number of compute providers for Gen AI**. We have found evidence of over 100 providers of compute for Gen AI, including several new AI-specialised providers that have launched services in recent years.<sup>8</sup> In addition, well-established on-premises providers<sup>9</sup> offer servers optimised for Gen AI workloads, supporting on-premises solutions (hereafter 'on-prem'). As of May 2025, there were 48 on-premises providers that integrated NVIDIA's accelerators, popular for Gen AI workloads, in their products.<sup>10</sup> In general, developers and deployers appear to use several options to access compute for Gen AI workloads, including public clouds and

---

<sup>4</sup> European Commission (2024). Competition policy brief, Available [here](#).

<sup>5</sup> CMA (2024). AI Foundation Models, Available [here](#).

<sup>6</sup> One source suggests that the three largest compute providers hold a combined share of around 75 per cent of the overall public cloud sector but might account for nearly 96 per cent of all new public cloud AI projects. See IOT Analytics (2024). Who is winning the cloud AI race? Microsoft vs. AWS vs. Google, Available [here](#).

<sup>7</sup> Vertically integrated players may choose to prioritise their own downstream Gen AI services and distort competition downstream.

<sup>8</sup> SemiAnalysis (2024). AI Neocloud Playbook and Anatomy, Available [here](#). (Accessed 9 June 2025). We cannot determine whether all of these players offer substitutable services from a buyer perspective. However, it appears as if many offer comparable services: access to similar types of accelerators (e.g. GPUs) via public or private clouds.

<sup>9</sup> We define on-premises providers as firms that supply server, storage, and networking systems that are physically located and operated within a customer's own data centre or colocation facility, enabling organisations to manage AI compute locally. In the context of Gen AI, this infrastructure may include built-in accelerators and may be optimised for pre-training, fine-tuning and/or inference workloads.

<sup>10</sup> These providers are listed as original equipment manufacturers in the compute category of NVIDIA's partner network. NVIDIA (2025). NVIDIA Partner Network, Available [here](#). (Accessed: 9 June 2025)

on-prem solutions. For example, according to a recent IDC global survey, almost half of respondents use primarily on-prem for both development and deployment of Gen AI.<sup>11</sup> Relatedly, Google Cloud's 2025 State of AI Infrastructure report shows that 93 per cent of AI developers and deployers multi-source for their Gen AI workloads, either from multiple cloud providers or by combining on-prem with one or more cloud providers.<sup>12</sup>

- (10) Second, **compute providers are growing across the board**. While it is not possible based on publicly available data to determine which players offer substitutable services, e.g. for specific types of workloads, it is apparent that many of the new AI-specialised players are securing major contracts with Gen AI developers, and achieving triple-digit revenue growth, in several cases outpacing the three largest cloud providers (AWS, Microsoft Azure, Google Cloud).<sup>13</sup> Independent benchmarks suggest that their service quality can rival that of more established firms, and their high valuations indicate strong expectations for continued expansion and competitive relevance.<sup>14</sup> Additionally, established on-premises providers, such as Dell, HPE and Supermicro, and other cloud providers, such as Oracle, are also registering high levels of revenue growth driven by their AI-related offerings.<sup>15</sup>
- (11) Third, **substantial investments in compute capacity are being made by private firms and public entities**, which suggests a fast-moving sector that is undergoing significant transformation. Total worldwide investment in data centres is forecasted to reach USD 1 trillion annually by 2029, half of which is expected to support AI workloads.<sup>16</sup> The three largest cloud providers, AWS, Microsoft Azure, and Google Cloud announced capital expenditure ranging from USD 75 to 100 billion each in 2025 alone. In parallel, several other initiatives, both private and public, are contributing to this momentum. In Europe, the European Commission's AI Continental Action Plan, launched in April 2025, sets out to mobilise EUR 200 billion (via a mix of public and private sources) to develop up to five AI "gigafactories" across strategic sectors, such as healthcare and science, while in the United States, the Stargate Project, driven by, among others, OpenAI and Oracle, aims to mobilise over USD 500 billion in AI-related infrastructure over the next five years.
- (12) Fourth, there is **indicative evidence suggesting declining prices and an expanding service offer**. According to industry experts, the inflation-adjusted price per FP32 FLOP (a standard measure of GPU processing performance)<sup>17</sup> has decreased by approximately 74 per cent since 2019,<sup>18</sup> with rental prices for GPUs often employed for Gen AI workloads falling from peaks of USD

---

<sup>11</sup> IDC (2024). On-Premises AI Infrastructure Balances Innovation and Security p.11, Available [here](#). (Accessed: 10 June 2025)

<sup>12</sup> Google Cloud (2025). State of AI infrastructure p.43, Available [here](#). (Accessed: 10 June 2025)

<sup>13</sup> Growth rates may not be directly comparable due to differences in baseline revenues and differences in service portfolio, i.e. inability to isolate growth in Gen AI business for firms with a broader cloud offering.

<sup>14</sup> SemiAnalysis (2025). The GPU Cloud ClusterMAX™ Rating System, Available [here](#). (Accessed: 20 May 2025)

<sup>15</sup> Copenhagen Economics based on firms' financial statements, public announcements and industry articles: Oracle (2024). Oracle Announces Fiscal 2025 Second Quarter Financial Results, Available [here](#); Dell Technologies Inc. (2025). Consolidated Statements of Income and Related Financial Highlights, Available [here](#); Supermicro (2025). 2024 Annual Report, Available [here](#); HPE (2024). Condensed Consolidated Statements of Earnings, Available [here](#) (All accessed: 11 June 2025)

<sup>16</sup> Dell'Oro Group (2025). Data Center capex to surpass 1 trillion by 2029, Available [here](#). (Accessed: 20 May 2025)

<sup>17</sup> FLOP stands for floating point operations per second. As FLOPs capture the amount of computational power delivered, this metric can be seen as a proxy for quality-adjusted prices of compute.

<sup>18</sup> Li, C (2025). The Evolution of GPU Pricing: A Deep Dive into Cost per FP32 FLOP for Hyperscalers, Available [here](#). (Accessed: 18 June 2025)

8 in 2023 to less than USD 2 in 2024. Similarly, according to S&P Global Market Intelligence's Cloud Price Index, the three largest cloud providers, AWS, Microsoft Azure and Google Cloud, have been reducing their prices for Gen AI services. Additionally, existing survey evidence suggests that in many cases Gen AI developers and deployers do not see the costs of compute as the main barrier to developing and deploying Gen AI solutions. In parallel with these pricing trends, during the past five years, seven out of the eight largest cloud providers globally according to Statista<sup>19</sup> (AWS, Microsoft Azure, Google Cloud, Alibaba, Oracle, IBM and Tencent) have introduced or expanded their service offer to Gen AI developers and deployers by: (i) offering direct access to accelerators, (ii) providing tools to support developers in training and fine-tuning, and (iii) offering an AI platform to access existing off-the-shelf models.

### **Various expected developments may further strengthen competition**

- (13) While the growth of Gen AI is expected to increase overall demand for compute, various industry and technological trends suggest that the needs of individual Gen AI developers and deployers may increasingly be met by a broader set of compute providers, which could strengthen competition in future.
- (14) First, **industry trends suggest demand is shifting towards less computationally intensive workloads**. While attention in recent years has focused on the training of new and improved Gen AI models, experts suggest that attention is now turning towards (i) fine-tuning of existing models for specialised tasks, and (ii) inference activity. S&P's Global Market Intelligence listed "*models [becoming] more specialised and domain-specific*" as one of the main "2025 Trends in Data, AI and Analytics", reporting that 67 per cent of firms using Gen AI technologies are fine-tuning an existing Gen AI model.<sup>20</sup> In parallel, according to Gartner, by 2028 over 80 per cent of AI workloads in data centres will be related to inference tasks – up from 40 per cent in 2023. The shift of demand towards less computationally intensive workloads may expand the range of compute options available to Gen AI developers and deployers.

---

<sup>19</sup> Statista (2025). Amazon and Microsoft Stay Ahead in Global Cloud Market, Available [here](#). (Accessed: 9 June 2025)

<sup>20</sup> S&P Global Market Intelligence (2025). 2025 Trends in Data, AI & Analytics, Available [here](#). (Accessed: 20 May 2025)

- (15) Second, **there is continuous innovation in compute usage**. Models are becoming more efficient and less computationally intensive – for pre-training, fine-tuning and especially inference workloads. Most leading AI developers are developing increasingly capable models trained on a smaller set of parameters (Small Language Models), while, according to Epoch.ai, inference costs have fallen dramatically across the board by 9x to 900x.<sup>21</sup> These trends are driven, among others, by advancements in (i) model design with the adoption of techniques such as model distillation, mixture-of-experts architectures and (ii) hardware advancements, with the consistent release of new generations of accelerators and performance per dollar improving by around 30 per cent annually. Additionally, on-device inference, e.g. running a Gen AI model on a smartphone, is expected to become more viable in the years to come, which will create an additional source of compute for some workloads. Combined, these innovations could mean that a broader range of compute options may be able to serve Gen AI developers and deployers.

**Policymakers should carefully assess competitive conditions before considering any potential intervention**

- (16) As outlined, some policymakers have voiced concerns about the potential risks of market concentration and limited access to compute for Gen AI. There have been some suggestions to intervene in order to regulate the supply of cloud services.<sup>22</sup>
- (17) Current evidence suggests an evolving landscape which appears consistent with an increasingly competitive environment. New and diverse players are entering and expanding, prices appear to be declining, and demand is shifting toward less computationally intensive workloads, which can potentially be more easily served by a wider range of options. Against this backdrop, regulatory intervention may be premature. Continued monitoring of market outcomes is warranted to ensure that access to compute remains open, competitive, and supportive of innovation.

**Structure of this report**

- (18) The remainder of this report is structured as follows:
- Chapter 1 provides an overview of the role of compute and the growing demand for Gen AI workloads.
  - Chapter 2 presents evidence on the functioning of the market for compute, including an overview of providers of compute for Gen AI, Gen AI developers and deployers' use of available options, investments and price trends.
  - Chapter 3 presents evidence on expected developments in compute needs and their potential implications for competition in the provision of compute for Gen AI.
  - Chapter 4 provides concluding remarks, summarising the key insights from the preceding chapters.

---

<sup>21</sup> The range in inference cost decline reflects differences in the performance benchmarks used. The slowest decline (9x) corresponds to the cost of achieving GPT-3.5 Turbo (March 2023 -level performance on general knowledge tasks. The fastest decline (900x) applies to achieving GPT-4o (May 2024)-level performance on Ph.D.-level science questions. Epoch AI (2025). LLM inference prices have fallen rapidly but unequally across tasks, Available [here](#). (Accessed: 9 June 2025)

<sup>22</sup> European Commission recently launched a public consultation for an impact assessment on a Cloud and AI development Act which is considering, among the policy options, measures to “address the computational capacity deficit”. See: European Commission (2025). Call for evidence for an impact assessment. AI Continent – new cloud and AI development act. Page 2. Available [here](#). (Accessed: 20 May 2025)

# 1 BACKGROUND

## Key findings

- Compute is an important input throughout the value chain of Gen AI development and deployment.
- Compute can be accessed in several ways: public cloud, private or virtual cloud, on-premises solutions, or even on-device for some inference workloads.
- Compute is required in all three main types of Gen AI workloads: pre-training, fine-tuning, and inference.
- As a rule of thumb, compute needs decline for workloads further down the value chain: pre-training workloads are more computationally intensive whilst inference workloads are less so. However, inference workloads are recurring, and widespread use of a Gen AI model can lead to significant cumulative compute needs.
- Demand for Gen AI workloads is expected to increase by an average of 39 per cent per year between 2023-2030.
- Some competition authorities and policymakers, such as the European Commission and the UK's Competition and Markets Authority (CMA), have raised concerns about competitive dynamics in the provision of compute for Gen AI.

- (19) Gen AI refers to a subset of AI systems that are capable of generating new content — such as text, images, code, or audio — based on learned patterns in data. These systems are typically built on foundation models (hereafter “Gen AI models”), such as OpenAI’s GPT or Meta’s Llama, which are complex probabilistic models adaptable to a wide range of tasks.
- (20) Gen AI is expected to contribute to substantial economic growth by increasing productivity and stimulating innovation. Research by Goldman Sachs suggests that Gen AI could support a 7 per cent increase in global GDP over a ten-year period.<sup>23</sup> Similarly, J.P. Morgan projects that this impact could be even greater, with potential gains of up to 10 per cent of global GDP.<sup>24</sup> There is ongoing debate regarding the precise magnitude of these effects, with some experts suggesting that the benefits may be more modest.<sup>25</sup> Productivity gains may arise from the automation of routine tasks, improvements in decision-making, and enhancements in business processes. Innovation can be stimulated by enabling the development of new products, services, and business models as well as scientific discovery (e.g. discovery of new proteins and materials<sup>26</sup>).
- (21) In this chapter, we describe:
- The role of compute as an important input to Gen AI and the different compute needs in the development and deployment of Gen AI (Section 1.1).

<sup>23</sup> Goldman Sachs (2023). Generative AI could raise global GDP by 7%, Available [here](#). (Accessed: 20 May 2025)

<sup>24</sup> The exact timeframe for this projected GDP increase is not specified. J.P. Morgan (2024). Is generative AI a game changer? Available [here](#). (Accessed: 20 May 2025)

<sup>25</sup> Professor Daron Acemoglu estimates that AI could raise US GDP by only 1%–1.5% over a decade. Acemoglu, D. (2024). The Simple Macroeconomics of AI. *NBER Working Paper* No. 32487.

<sup>26</sup> OECD (2023). Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research, OECD Publishing, Paris, Available [here](#).

- The concerns raised by some policymakers in relation to competitive conditions for the provision of compute (Section 1.2).

## 1.1 COMPUTE IS AN IMPORTANT INPUT TO GEN AI

- (22) The Gen AI value chain, that ranges from developing the models to integrating them into end-user applications, relies on compute.<sup>27</sup> In this study, we distinguish between the development and deployment phases of the value chain and introduce the main types of computational tasks (or “workloads”) associated with each that require compute.
- (23) Development consists of the design and building of new models. This process typically involves processing large volumes of unstructured data in order to identify patterns which are then encoded into billions of parameters (or “weights”) representing the model’s learned knowledge.<sup>28</sup> Gen AI models can differ along several dimensions, such as in modality, specialisation, their degree of openness and more, see Box 1.

### Box 1

#### Different types of Gen AI models

#### Gen AI models differ in:



**Modality** (language, visual, multi-modal). Language models (e.g. GPT, Llama) generate text; visual models (e.g. DALL-E, Stable Diffusion) create images; and multi-modal models integrate text, image, or audio inputs. Emerging AI agents combine these capabilities with task planning and decision-making.



**Specialisation.** Models may be general-purpose or domain-specific, with rising investment in specialised models tailored to particular industries. Examples of such specialised models include Med-PaLM 2 for healthcare, BloombergGPT for finance, and Harvey for legal applications.



**Size.** Model size, often measured in billions of parameters, affects performance and resource needs: larger models perform better on complex tasks but are costlier to run; smaller models offer efficiency and are increasingly effective when fine-tuned.



**Openness.** Open weight (also referred to as “open source”) models (e.g. Llama) can be downloaded and customised by developers; fully open weight models may also share training data and code. In contrast, closed weight (also referred to as “proprietary”) models (e.g. GPT-4, Gemini) are only accessible via APIs, with compute managed by the provider.

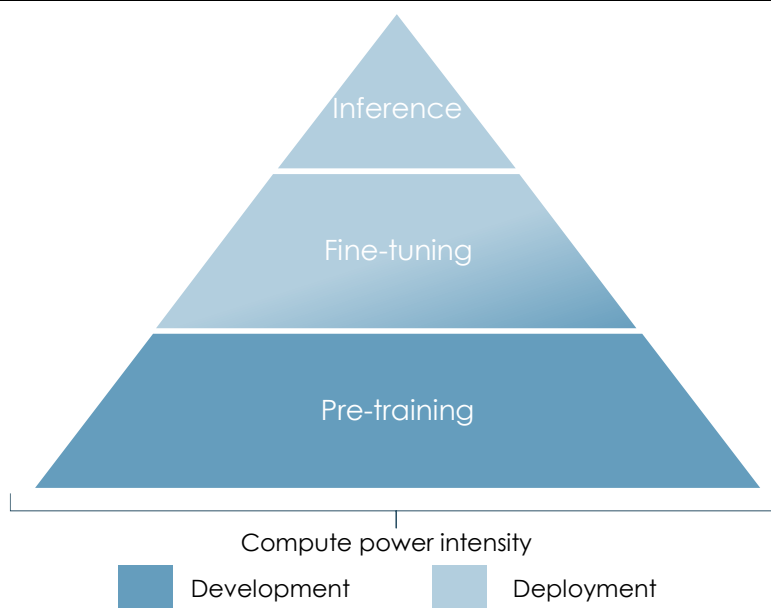
Sources: <sup>1</sup>Forbes, <sup>2</sup>DeepSeek, <sup>3</sup>Unite.Ai, <sup>4</sup>The Science Survey, <sup>5</sup>Unite.Ai, <sup>6</sup>What went into training DeepSeek-R1? | Epoch AI, <sup>7-9</sup>arXiv: A Review of DeepSeek Models’ Key Innovative Techniques, pages 2,4,5-6

<sup>27</sup> Besides compute, other factors such as data and expertise, are also important inputs to Gen AI and together form what other studies has defined as the ‘infrastructure layer’ of the Gen AI value chain. See for further details CMA (2024). AI Foundation Models, Available [here](#).

<sup>28</sup> Microsoft Community Hub (2024). Differences between Pre-Training and Supervised Fine-Tuning (SFT), Available [here](#). (Accessed: 9 June 2025)

- (24) Once developed, a Gen AI model can be made available in several ways, such as via the model developer's platforms/APIs, via cloud providers' platforms, or, in some cases, by making the model parameters directly available for download.<sup>29</sup>
- (25) Deployment consists in the adaptation and integration of a Gen AI model into applications and services for end-users. This could, for instance, be via a conversational chatbot (e.g. OpenAI's ChatGPT), specialised tools for coding (e.g. GitHub Copilot) or image generation platforms (e.g. Midjourney). Deployment may be delivered either by a vertically integrated Gen AI developer or by third parties which build on an existing model.
- (26) Within this development-to-deployment process, there are three main types of Gen AI workloads: pre-training, fine-tuning, and inference – each placing different demands on compute, see **Figure 1**.

**Figure 1**  
**Compute needs vary by type of Gen AI workload**



Note: We consider fine-tuning to be part of development when performed extensively but it can also be considered part of deployment when it is more limited and performed by downstream players. In a typical value chain, inference would be considered part of the downstream of pre-training, while here it is shown at the top of the pyramid.

Source: Copenhagen Economics

<sup>29</sup> These parameters consist of the weights and biases that are learned during training which, along with the model architecture, determine the model's behaviour. See, for instance, Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning, MIT press.

- (27) **Pre-training** is the most computationally intensive workload.<sup>30</sup> It is part of the initial development phase and consists in training the model from scratch on unstructured data to establish its general capabilities. While exact compute needs depend on the complexity of the model, pre-training typically requires dense clusters of high-end hardware accelerators (equivalent to advanced semiconductor chips or processors, e.g. Graphic Processing Units or ‘GPUs’, or Tensor Processing Units or ‘TPUs’) operating in parallel over extended periods.<sup>31</sup> This type of workload is concluded once the model has been developed. The amount of compute needed for pre-training therefore does not vary with subsequent use of the model. As such, the cost to source compute could be considered fixed for a specific model – although AI developers need to innovate constantly, thus “training” costs at the firm level may be recurring.
- (28) **Fine-tuning**, sometimes referred to as post-training or refinement, involves further training of an existing model using targeted datasets to customise it for specific tasks (e.g. legal summarisation or customer support).<sup>32</sup> Fine-tuning is also part of development when performed extensively (in some cases effectively resulting in a “new model”), but can be considered part of deployment when it is more limited and performed by downstream players. Fine-tuning typically requires less compute resources (e.g. a smaller number of accelerators) than pre-training and can be executed on more flexible infrastructure over shorter timeframes.<sup>33</sup> Some AI developers and cloud providers offer deployers the ability to fine-tune models using their own data (OpenAI’s fine-tuning API for GPT models<sup>34</sup>, AWS’s SageMaker<sup>35</sup>). Just as with pre-training, the costs associated with procuring compute for fine-tuning do not depend on how much the model is subsequently used.
- (29) **Inference** is the process of applying a trained model to generate outputs in response to user inputs (e.g. answering a query). It is generally less computationally demanding per workload but places requirements on responsiveness and scalability as the source of compute must support real-time performance and simultaneous requests from many users. In contrast to pre-training and fine-tuning, the costs associated with procuring compute for inference depend on how much the model is used – and compute needs thus scale with usage, therefore best being categorised as variable costs. As such, while the per-unit compute need is low, aggregate compute usage for inference can be high.

---

<sup>30</sup> “Pre-training is the most computationally intensive step of developing a FM, often requiring hundreds of accelerator chips for many days.” CMA (2023). AI Foundation Models: Initial Report. Page 126, Available [here](#). (Accessed: 20 May 2025).

<sup>31</sup> Accelerators are specialised processors, chips or hardware units designed for accelerated computing which enable a high number of calculations to be conducted often simultaneously. Graphic Processing Units (GPUs) are examples of such hardware, originally designed for graphic rendering tasks but now used extensively for Gen AI model training. Another example is Tensor Processing Units (TPUs) designed for large matrix calculations. The pre-training process may require dense clusters of high-performance accelerators – in some cases numbering in the thousands – working in parallel over extended period that may range from days to multiple weeks or even months. These clusters need to be tightly interconnected with high bandwidth to enable efficient training. NVIDIA (2021). What Is Accelerated Computing? Available [here](#). (Accessed: 9 June 2025).

<sup>32</sup> We note that fine-tuning is only a form of post-training which include any modification of the base model after pre-training (e.g., reinforced learning, model distillation, etc.). For simplicity here we refer to fine-tuning as post-training more generally.

<sup>33</sup> IBM (2024). What is fine-tuning? Available [here](#). (Accessed: 9 June 2025).

<sup>34</sup> OpenAI. Fine-tuning, Available [here](#). (Accessed: 9 June 2025).

<sup>35</sup> Amazon SageMaker AI. Fine-tune a Model, Available [here](#). (Accessed: 9 June 2025).

Although inference is triggered by end-users, the process of sourcing compute for these workloads is typically managed by deployers.<sup>36</sup>

- (30) Overall, compute requirements per workload tend to decline through the value chain – from pre-training to inference – but each workload imposes different demands. The nature of these workloads determines which compute providers can best serve them, with pre-training typically requiring advanced large-scale infrastructure and inference being more distributed and widely contestable.<sup>37</sup>
- (31) Notwithstanding these differences, Gen AI developers and deployers have different options to access compute for Gen AI workloads, from public cloud services to private or virtual clouds, and on-prem solutions, with the possibility to even execute certain workloads, within inference, on-device, see **Table 1**.

**Table 1**  
**Compute for Gen AI can be accessed from multiple sources**

OPTION	DESCRIPTION
Public cloud	Compute is available either directly or via an AI development platform. On-demand access is typical but fixed-term contracts are possible as well. High flexibility to scale supply to meet demand as needs evolve.
Private/virtual cloud	Cloud resources are dedicated to one firm via long-term contracts. Users have more control over the infrastructure, but scaling quickly can be more challenging.
On-premises	AI firms can acquire their own hardware and operate it either on-premises or in a colocation centre, where several tenants can rent space for their own compute hardware in a third-party data centre. This option grants full control over infrastructure but scaling and updating may be more costly.
On-device (a subset of "edge" computing)	Certain inference workloads with lower compute requirements can be run directly on AI-compatible devices, such as certain smartphones, which can be beneficial for tasks requiring low latency, or in cases where there may be connection issues. This option is widely available to users but is limited in terms of compute.

Source: Copenhagen Economics

- (32) In some cases, the source of compute is pre-determined by the type of model and mode of access. Gen AI developers usually decide for themselves how to buy or allocate compute to train their models. Gen AI deployers, however, may find that their preferred Gen AI model and/or mode of access for model inference (e.g. via a specific cloud provider's platform) is 'pre-packaged' with a specific cloud provider.<sup>38</sup>

<sup>36</sup> There are various forms of vertical integration in the industry. The same entity can be a developer or a deployer and the same deployer can also be the end-user of a model application or service.

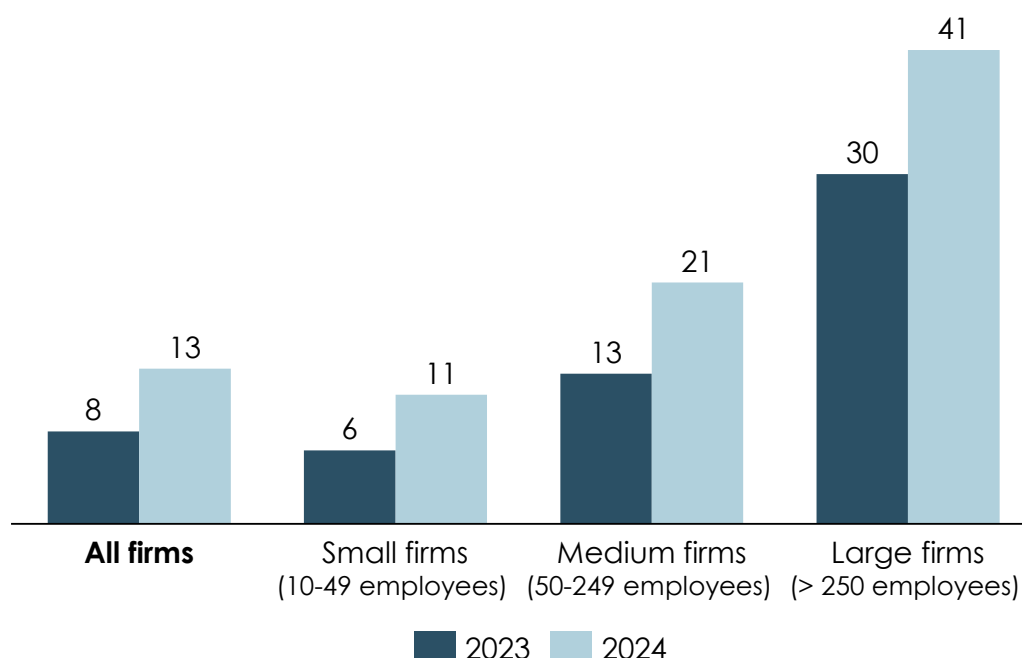
<sup>37</sup> CERRE (2025). A Competition Policy for Cloud and AI, Available [here](#). (Accessed: 19 June 2025). See also Financial Times (2025). How 'inference' is driving competition to NVIDIA's AI chip dominance, Available [here](#). (Accessed: 10 June 2025).

<sup>38</sup> Gen AI deployers or users who download open weight models can choose their compute sources, such as their own servers or platforms like Hugging Face. In contrast, accessing closed weight models via APIs or cloud platforms means that the Gen AI developer manages the computational resources, while the Gen AI deployer or user typically pays a usage fee to access the model.

## 1.2 SOME POLICYMAKERS HAVE RAISED COMPETITION CONCERNS

- (33) The adoption of Gen AI services among firms is currently experiencing significant growth. In the European Union ('EU'), the share of firms using AI technologies increased from 8 to 13 per cent between 2023-2024, with substantially higher adoption rates among large firms, see **Figure 2**.<sup>39</sup>

**Figure 2**  
**AI adoption in businesses has increased in the EU**  
Percentage of firms



Note: Eurostat refers to firms as "enterprises", which we assume is a synonymous term.

Source: Eurostat (2025). Use of artificial intelligence in enterprises, Available [here](#). (Accessed 20 May 2025)

- (34) Growth in the use of Gen AI services is driving increased demand for compute, with demand for Gen AI workloads expected to grow by an average of 39 per cent per year between 2023-2030.<sup>40</sup>

<sup>39</sup> Eurostat (2025). Use of artificial intelligence in enterprises, Available [here](#). (Accessed: 20 May 2025). The report analyses AI generally and not just Gen AI.

<sup>40</sup> McKinsey (2024). AI power: Expanding data center capacity to meet growing demand, Available [here](#). (Accessed: 20 May 2025).

- (35) Because compute is such an important input, sufficient and competitive access to compute is important for the growth in Gen AI. Some competition authorities and policymakers in Europe and UK have expressed concerns about competitive dynamics in the provision of compute for Gen AI. In particular, there is a concern that a large share of compute capacity is concentrated in the hands of a few vertically integrated providers, which may choose to prioritise their own downstream Gen AI services, potentially limiting access for third parties and hindering broader Gen AI adoption.<sup>41</sup> One source suggests that the three largest computer providers (AWS, Microsoft Azure, Google Cloud) hold a combined share of around 75 per cent of the overall public cloud sector but might account for nearly 96 per cent of all new public cloud AI projects.<sup>42</sup>
- (36) In the September 2024 EU Competition Policy Brief on competition in Gen AI, the European Commission outlined, as one potential competition risk to monitor, “*the risk that incumbent large digital players, which may currently enjoy preferential access to generative AI’s key components, grant it to third parties on an exclusive basis, or prevent competitors from accessing it. This may affect access to any of the key inputs [...], including computing infrastructure [...]*.”<sup>43</sup>
- (37) Similarly, the CMA has reported a concern that firms which control critical inputs, such as compute, “*may restrict access to them to shield themselves from competition.*”<sup>44</sup> Specifically, “*they could do this to: (1) prevent other firms from building new, competitive [Gen AI models] that might challenge their own (or their partners’, where relevant); and/or (2) protect their position in related markets, by making it harder for potential rivals in those markets to develop or deploy capable [Gen AI models]*”.<sup>45</sup>
- (38) We are not aware of any in-depth investigations by competition authorities that have found evidence of these concerns materialising currently. Rather, concerns have reflected a forward-looking perspective and underscore the need to closely monitor market developments in this rapidly changing and strategically important area.

---

<sup>41</sup> When authorities refer to “incumbent large digital players”, they may be referring to the three largest cloud providers in Europe: AWS, Microsoft Azure, and Google Cloud. AWS, Microsoft Azure and Google Cloud are all active in the downstream market for Gen AI development via vertical integration and/or partnerships.

<sup>42</sup> IOT Analytics (2024). Who is winning the cloud AI race? Microsoft vs. AWS vs. Google, Available [here](#).

<sup>43</sup> European Commission (2024). Competition policy brief, p.7, Available [here](#). (Accessed: 20 May 2025).

<sup>44</sup> CMA (2024). AI Foundation Models. Page 77, Available [here](#). (Accessed: 20 May 2025).

<sup>45</sup> CMA (2024). AI Foundation Models. Page 77, Available [here](#). (Accessed: 20 May 2025).

## 2 EVIDENCE SUGGESTS THAT PROVIDERS OF COMPUTE FOR GEN AI OPERATE IN AN INCREASINGLY COMPETITIVE ENVIRONMENT

### Key findings

- More than 100 players, including newer AI-specialised players and established on-premises providers<sup>46</sup>, offer compute for Gen AI.
- Evidence shows that Gen AI developers and deployers frequently employ multiple providers for compute, including extensive use of on-premises solutions. This suggests that developers and deployers may have severable viable alternatives when sourcing compute.
- Emerging AI-specialised players and on-premises providers are exhibiting rapid revenue growth, in many cases growing faster than the three largest cloud providers (AWS, Microsoft Azure, Google Cloud).
- Dell'Oro forecasts that global investments in data centres will reach USD 1 trillion annually by 2029, with about half of this investment directed towards servers optimised for AI workloads.
- Prices appear to be declining, with the inflation-adjusted price per FP32 FLOP (a standard measure of GPU processing performance)<sup>47</sup> decreasing by approximately 74 percent since 2019.

- (39) In this chapter, we seek to assess whether competition concerns regarding the provision of compute for Gen AI are currently valid. We do this by examining whether market outcomes are broadly consistent with those that would be expected in a competitive environment.
- (40) In a well-functioning, competitive environment,<sup>48</sup> we would expect to observe a large and/or growing number of providers offering compute, that Gen AI developers and deployers make use of multiple options to access compute and can substitute between them, that there are low or declining levels of concentration, sustained levels of investment, expanding service offers, and signs of declining prices. By contrast, persistent barriers to entry for the provision of compute, limited choice, high concentration, and rising prices would suggest the need for closer scrutiny.

<sup>46</sup> We define on-premises providers as firms that supply server, storage, and networking systems that are physically located and operated within a customer's own data centre or colocation facility enabling organizations to manage AI compute locally. In the context of generative AI, this infrastructure may include built-in accelerators and be optimized for training and inference workloads.

<sup>47</sup> FLOP stands for floating point operations per second. As FLOPs capture the amount of computational power delivered, this metric can be seen as a proxy for quality-adjusted prices of compute.

<sup>48</sup> A competitive environment can be linked to a well-functioning market. The benchmark of a well-functioning market is somewhat abstract – especially in cases where there may be economies of scale and/or network effects. The CMA has described a “well-functioning” market as one that “*displays the beneficial aspects of competition which make markets work well for customers [...] but not an idealized, perfectly competitive market*”. CMA (2024) [Draft] Markets Guidance, Available [here](#). (Accessed: 20 May 2025).

- (41) Based on our assessment of available evidence, we find that market outcomes are broadly consistent with an increasingly competitive market. Specifically, we find that:
- There is a large and increasing number of providers of compute for Gen AI-related use cases, and that Gen AI developers and deployers source from multiple viable alternatives (Section 2.1)
  - Compute providers are growing across the board (Section 2.2)
  - There is substantial investment in expanding compute capacity (Section 2.3)
  - Evidence suggests declining prices and an expanding service offer (Section 2.4)

## **2.1 THERE IS A LARGE AND INCREASING NUMBER OF PROVIDERS OF COMPUTE FOR GEN AI**

### **2.1.1 There are many players and specialised new entrants offering compute solutions for Gen AI**

- (42) There is a growing range of providers offering different solutions to access compute, suggesting that the market is increasingly competitive.
- (43) According to research from early 2025 by SemiAnalysis, there are more than 100 players offering access to accelerators (mostly GPUs) for Gen AI workloads via public or private cloud. This includes large established cloud providers, such as AWS, Microsoft Azure and Google Cloud, as well as many other large firms, and newer AI-specialised providers, see Figure 3.

**Figure 3**  
**More than 100 players offer compute for Gen AI**

Compute power providers - established cloud providers and AI-specialised players				Brokers and platforms
ACECloud.ai	Denvr Data	Hut8	Ows Cloud	Aethir
Airon AI	DigitalOcean	IBM Cloud	Patmos	Akash Network
Akamai	Dihuni	Ionos	puzi.cloud	Atlas Cloud
Aligned.co	E2E Networks	ionstream.ai	Qubrid AI	Clore.ai
Aolani Cloud	Elastx	Iris Energy	Runsun Cloud	Cudo Compute
Alibaba Cloud	Engage Stack	Jarvis Labs AI	Runpod	Dataoorts Cloud
Applied Digital	EscherCloudAI	klustr.ai	Salad	DGX Cloud
Aperia Cloud	Evroc	Lambda labs	Scaleway	Fluidstack
Arc Compute	Exoscale	Lamini	Scott Datacenter	gpucompare.com
AtNorth Compute	Fast GPU	Lafitude.sh	Seeweb	GPU.net
Atlas Cloud	FastWeb	LeaderGPU	Sesterce	gpumart.com
Atos	FlexAI	LeptonAI	Sify Technology	GpuList.ai
Ax3 AI	Genesis Cloud	Liquid-Web	Soluna Computing	hostedAI
AWS	Gcore	Massed Compute	SMC	Hydrahost
Baionity Cloud	GMI Cloud	Megaspeed AI	Stack IT	Lightning AI
Backprop Cloud	Google Cloud	Microsoft Azure	Taiga Cloud	ML Foundry
banana.dev	GreenAI Cloud	ML Foundry	Telogiva Qunex Cloud	Modal
Bitdeer Cloud	Hetzner GPU Cloud	Nebius	Tencent Cloud	NetMind AI
Cirrascale	Horizon Compute	NeevCloud	Tensordock	Prime Intellect
CivoCloud	Hot Aisle Cloud	NexGen Cloud	Tensorwave	Shadeform
Clear.ml	hpc-ai.com	Nscale	Together AI	Salad
Contabo GPU Cloud	HPE Greenlake	Oblivus Cloud	Turboscale	valdi.ai
Coreweave	Huawei Cloud	Outscale	Ubitus	Vast.AI
Corvex.ai	Hyperstack	Oracle	Utho	
CR8DL	Hypertec Cloud	OVHcloud	VNG Cloud	
Crusoe			Vultr	
Datacrunch				

Source: Copenhagen Economics based on SemiAnalysis (2024). AI Neocloud Playbook and Anatomy, Available [here](#). (Accessed 9 June 2025)

AI-specialised providers — sometimes referred to as AI Neoclouds<sup>49</sup> — focus on offering access to accelerators (mostly GPUs) as a service, i.e. renting the accelerators to Gen AI developers and deployers for both pre-training, fine-tuning and inference workloads. CoreWeave, Lambda, Crusoe, and Nebius are examples of leading AI-specialised providers globally, each with established operations in Europe, see Box 2.

<sup>49</sup> “An AI Neocloud is defined as a new breed of cloud compute provider focused on offering GPU compute rental”. While their services may be used for AI workloads overall beyond Gen AI (e.g. more traditional machine learning and predictive techniques), most of them have become popular with the rise and adoption of Gen AI. SemiAnalysis (2024). AI Neocloud Playbook and Anatomy, Available [here](#). (Accessed 9 June 2025).

**Box 2**

**Leading global AI-specialised cloud providers**



CoreWeave is the largest AI-specialised cloud provider with expertise in running large-scale workloads. Established in 2017, they completed an IPO in March 2025. They have provided compute to OpenAI, Mistral and Cohere, among others.<sup>1</sup>



Designed exclusively for AI, Lambda delivers seamless, enterprise-grade access to compute power. Founded in 2012, Lambda's first compute offering for Gen AI was GPU powered desktop workstations. Since then, they have expanded to a full range of compute services including public cloud, private cloud, and on-premises solutions.<sup>2</sup>



Crusoe offers accelerators both on-demand and through 6-month to 3-year contracts via its fast and scalable cloud platform.<sup>3</sup> It is also a leading builder of energy-efficient data centres for Gen AI, which earned it a role as a key infrastructure partner in OpenAI's Stargate initiative.<sup>4</sup>



Nebius offers accelerators through both on-demand access and reserved capacity via its AI-optimized cloud platform. It is positioning itself as a full-stack AI infrastructure provider across Europe and North America and they operate data centres in Finland, France, and the United States. Nebius is publicly traded on Nasdaq.<sup>5</sup>

Sources: <sup>1</sup>Forbes (2025). CoreWeave Cofounder Discusses Cloud AI Firm's IPO, Available [here](#); <sup>2</sup>Lambda. The AI Developer Cloud, Available [here](#); <sup>3</sup>Crusoe (2024). Empowering the AI Revolution, Available [here](#); <sup>4</sup>Forbes (2025). Meet The Tiny Startup Building Stargate, OpenAI's \$500 Billion Data Center Moonshot, Available [here](#); <sup>5</sup>TechCrunch (2024). The curious case of Nebius, the publicly traded AI infrastructure 'startup', Available [here](#). (All accessed: 17 June 2025)

- (44) Recently, there are also examples of Gen AI developers that are integrating upstream and starting to supply compute as part of their Gen AI offer. Leading French Gen AI developer Mistral, in partnership with NVIDIA, has announced Mistral Compute, a *“new AI infrastructure offering that will provide customers a private, integrated stack—GPUs, orchestration, APIs, products, and services in whatever form factor they need, from bare-metal servers to fully-managed PaaS”*.<sup>50</sup>
- (45) Additionally, besides compute offered on public cloud, a significant number of firms offer AI-optimised servers and related infrastructure for on-premises setups (“on-prem” or private/virtual cloud environments). Examples include well-known companies such as Dell, HPE, Lenovo, NVIDIA, Supermicro, Cisco, VMware, IBM, Huawei, and Fujitsu. As of May 2025, there were 48 on-premises providers that integrated NVIDIA’s accelerators, popular for Gen AI workloads, in their products.<sup>51</sup> As discussed in the following section, these solutions represent an additional avenue, often used by Gen AI developers and deployers, through which compute can be sourced.

<sup>50</sup> Mistral (2025). Mistral Compute, Available [here](#). (Accessed: 17 June 2025).

<sup>51</sup> These providers are listed as original equipment manufacturers in the compute category of NVIDIA’s partner network. NVIDIA (2025). NVIDIA Partner Network, Available [here](#). (Accessed: 9 June 2025).

- (46) Finally, new semiconductor startups, such as Cerebras, Groq, and SambaNova, are developing AI accelerators designed specifically for AI workloads and mostly for inference.<sup>52</sup> These firms now offer services similar to those of cloud providers, allowing Gen AI deployers to run inference on open weight models using their own custom accelerators. While these companies are currently smaller players, they have begun to secure notable deals with leading Gen AI developers, such as Mistral<sup>53</sup> and Meta,<sup>54</sup> to manage the inference of their models.<sup>55</sup>

### **2.1.2 Gen AI developers and deployers frequently use multiple viable alternatives for compute**

- (47) Not only is the number of providers increasing, but based on available evidence, Gen AI developers and deployers appear to make use of several options to procure and access compute – and appear to make extensive use of on-prem solutions. This suggests that there could be the potential for healthy competition between different types of providers.
- (48) A recent global study by IDC reported that around half of respondents rely primarily on on-prem solutions (covering on-prem, virtual cloud and colocation) for both Gen AI development (training and fine-tuning) and Gen AI deployment (inference), see Figure 4.<sup>56</sup>

---

<sup>52</sup> AI accelerators are specifically optimised for AI workloads while GPUs are general accelerators used also for other tasks such as video rendering.

<sup>53</sup> Reuters (2025). AI chip firm Cerebras partners with France's Mistral, claims speed record, Available [here](#). (Accessed: 9 June 2025).

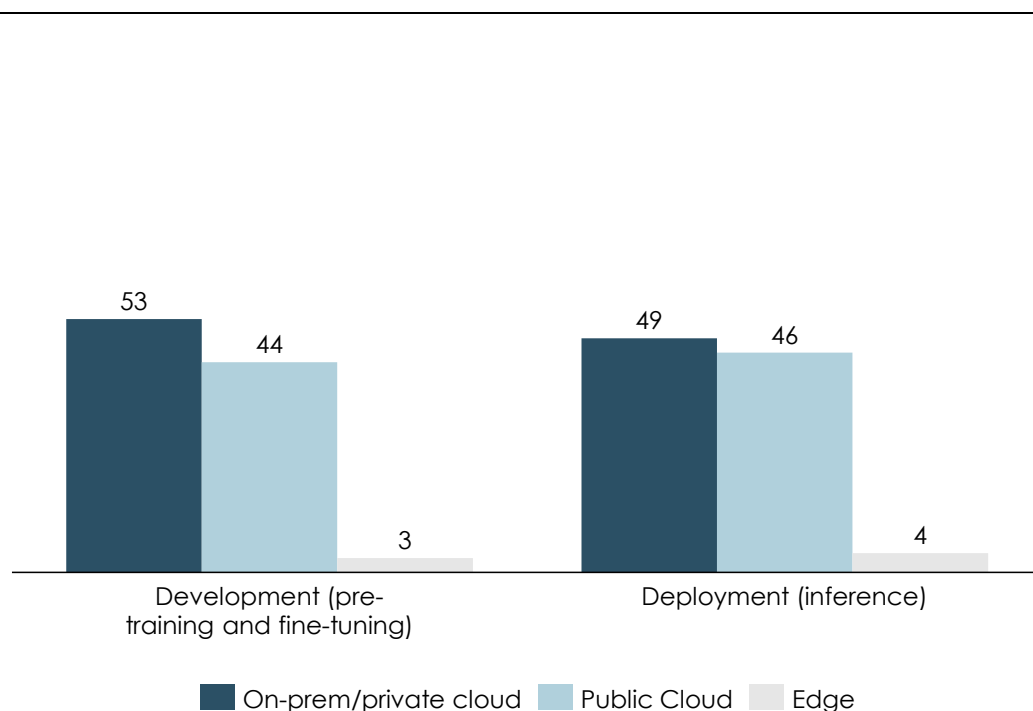
<sup>54</sup> Business Wire (2025). Meta Collaborates with Cerebras to Drive Fast Inference for Developers in New Llama API, Available [here](#). (Accessed: 9 June 2025).

<sup>55</sup> They also report faster inference performance compared to larger providers such as AWS, Microsoft Azure, and Google Cloud. Artificial analysis (2025). LLM API Providers Leaderboard, Available [here](#). (Accessed: 17 April 2025).

<sup>56</sup> We note there is no information about the distribution of respondents across firm size, industry and focus area (development vs deployment). The study also does not specify whether this includes only services related to pure access to compute or other services such as data storage.

**Figure 4**  
**IDC study shows that on-prem is the most common source of compute for development and deployment workloads**

Share of respondents (per cent)



Note: Respondents were asked "Where does your organisation primarily develop and deploy AI models?" There was a total of 411 responses. IDC defines on-premises as private cloud, traditional (non-cloud) deployments and colocation. Development is defined as training and fine-tuning/customising AI models, while deployment is defined as inferencing.

Source: Copenhagen Economics based on IDC (2024). On-Premises AI Infrastructure Balances Innovation and Security p.11, Available [here](#). (Accessed: 10 June 2025)

- (49) While the choice between cloud-based and on-premises compute solutions may reflect a broader trade-off between operating expenses ("opex") and capital expenses ("capex"),<sup>57</sup> this evidence shows that on-prem is the most common choice for several types of workloads throughout the value chain by at least a large share of developers and deployers. The propensity to consider on-prem as a viable option for Gen AI workloads is consistent with other data points. For example, according to a study prepared by the AI Infrastructure Alliance, 40 per cent of large firms developing or deploying AI (including Gen AI) were planning to purchase more on-premises capacity in 2024.<sup>58</sup>

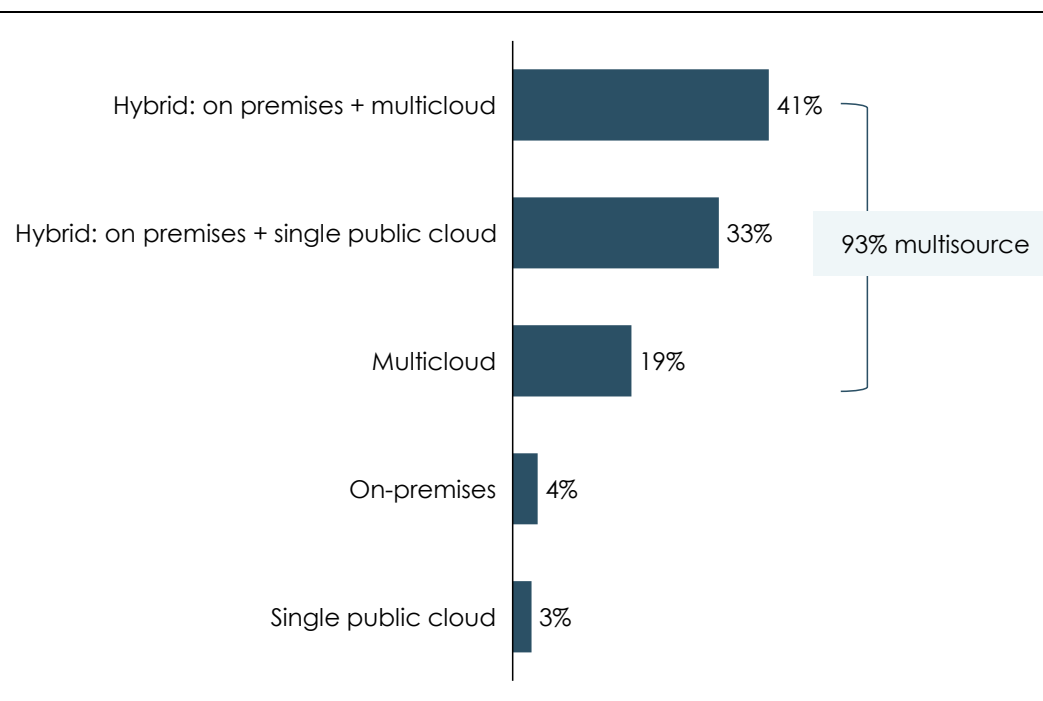
<sup>57</sup> (Public) cloud services allow Gen AI developers and deployers to scale usage flexibly and pay only for what they use, which can be particularly attractive for a firm that faces uncertainty about the long-term commercial impact of Gen AI. In contrast, on-premises infrastructure typically involves higher upfront investment but may offer cost advantages over time for stable, predictable workloads. As firms gain greater clarity on the business value and demand patterns of Gen AI applications, some may shift toward capital heavy strategies, such as on-premises or hybrid solutions, to optimise long-term cost-efficiency.

- (50) In addition, according to Google Cloud’s “2025 State of AI infrastructure” study, 93 per cent of firms<sup>59</sup> surveyed use multiple sources of compute for their Gen AI workloads, relying on on-prem combined with cloud, or on multiple cloud providers, see Figure 5.<sup>60</sup>

**Figure 5**

**Google Cloud study: 93 per cent of firms multisource compute for Gen AI**

“Which cloud infrastructure approach does your organisation primarily use for Gen AI workloads?” Share of respondents



Source: Copenhagen Economics based on Google Cloud (2025). State of AI infrastructure p.43, Available here. (Accessed: 10 June 2025)

- (51) While this Google study does not shed direct light on the substitutability of different compute solutions for specific tasks, widespread multi-sourcing suggests that Gen AI developers and deployers are able to evaluate and consider multiple options as relevant for their compute needs, and that there can be competitive pressure within and across different solutions.

<sup>58</sup> The study conducted a survey on firms with more than 500 employees (55 per cent with more than 10,000 employees). AI Infrastructure Alliance (2024). The State of AI infrastructure report 2024. p.10, Available here. (Accessed: 20 May 2025).

<sup>59</sup> We note the Google study uses the broader term organisations.

<sup>60</sup> While there is no information on any difference in the choice between training and inference workloads, we understand that the survey covered both Gen AI developers, deployers and users (which for the purpose of this study we consider as part of deployers).

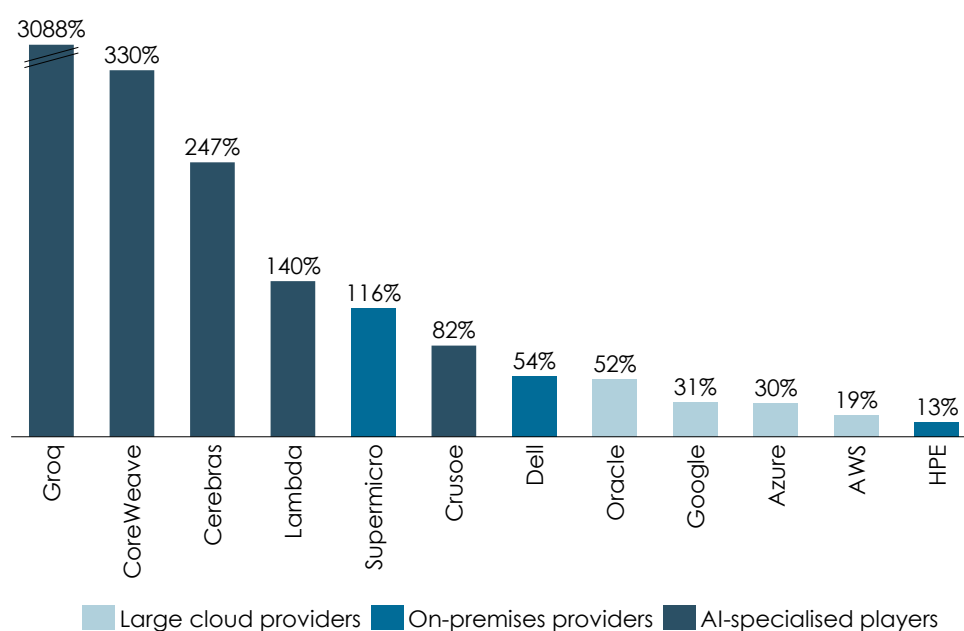
- (52) The tendency to source compute from multiple cloud providers is also reflected in the choices of many leading Gen AI developers. For instance, OpenAI now has contracts with Microsoft Azure<sup>61</sup>, Oracle<sup>62</sup>, CoreWeave<sup>63</sup> and Google Cloud<sup>64</sup>; Mistral sources from Microsoft Azure<sup>65</sup>, CoreWeave<sup>66</sup>, Cerebras<sup>67</sup>, and Google Cloud,<sup>68</sup> whilst also building its own infrastructure in partnership with NVIDIA<sup>69</sup>. Cohere has used Google Cloud<sup>70</sup> and Oracle<sup>71</sup>, and Anthropic has partnered with both AWS<sup>72</sup> and Google Cloud.<sup>73</sup>
- (53) More generally, many leading Gen AI developers that released notable models in 2024<sup>74</sup> tended to source compute from multiple alternatives, and there is a wide variety in the compute providers serving these firms.<sup>75</sup>
- (54) This is consistent with trends observed in cloud services more broadly. A global study conducted by Flexera in 2025 on cloud services also found high rates of multi-cloud and hybrid solutions used by firms. According to the study, 86 per cent of firms multi-source and 70 per cent have a hybrid cloud strategy, using at least one private cloud and one public cloud source.<sup>76</sup>

- 
- <sup>61</sup> OpenAI (2023). OpenAI and Microsoft extend partnership, Available [here](#). (Accessed: 9 June 2025)
- <sup>62</sup> Oracle (2024). OpenAI Selects Oracle Cloud Infrastructure to Extend Microsoft Azure AI Platform, Available [here](#). (Accessed: 9 June 2025).
- <sup>63</sup> OpenAI extended its portfolio of compute suppliers further via a USD 12 billion contract with CoreWeave in 2025. CoreWeave (2025). CoreWeave Announces Agreement with OpenAI to Deliver AI Infrastructure, Available [here](#). (Accessed: 9 June 2025).
- <sup>64</sup> Reuters (2025). Exclusive: OpenAI taps Google in unprecedented cloud deal despite AI rivalry, sources say, Available [here](#). (Accessed: 4 July 2025).
- <sup>65</sup> Microsoft Azure (2024). Microsoft and Mistral AI announce new partnership to accelerate AI innovation and introduce Mistral Large first on Azure, Available [here](#). (Accessed: 9 June 2025).
- <sup>66</sup> CoreWeave (2024). Mistral AI and CoreWeave Demonstrate Partnership at NVIDIA GTC, Mistral AI Hackathon, Available [here](#). (Accessed: 9 June 2025).
- <sup>67</sup> Reuters (2025). AI chip firm Cerebras partners with France's Mistral, claims speed record, Available [here](#). (Accessed: 9 June 2025).
- <sup>68</sup> Reuters (2023). Google Cloud partners with Mistral AI on generative language models, Available [here](#). (Accessed: 9 June 2025).
- <sup>69</sup> NVIDIA (2025). France Bolsters National AI Strategy With NVIDIA Infrastructure, Available [here](#). (Accessed: 17 June 2025).
- <sup>70</sup> Google cloud (2022). How Cohere is accelerating language model training with Google Cloud TPUs, Available [here](#). (Accessed: 9 June 2025).
- <sup>71</sup> Oracle (2024). Cohere and Oracle partnership brings generative AI solutions to enterprises, Available [here](#). (Accessed: 9 June 2025).
- <sup>72</sup> Anthropic (2024). Powering the next generation of AI development with AWS, Available [here](#). (Accessed: 9 June 2025).
- <sup>73</sup> Anthropic (2023). Anthropic Partners with Google Cloud, Available [here](#). (Accessed: 9 June 2025)
- <sup>74</sup> Epoch.ai defines a notable model as one that meets any of the following criteria: (i) state-of-the-art improvement on a recognized benchmark, (ii) highly cited (over 1000 citations), (iii) historical relevance, (iv) significant use. We note that this count may include AI models more generally rather than only Gen AI models.
- <sup>75</sup> We identify 24 firms that released a notable model according to the Epoch.ai database. After excluding integrated cloud providers (AWS, Microsoft Azure and Google Cloud) and Chinese AI developers, we identified compute sources or cloud partnerships for Anthropic (Available [here](#) and [here](#)), Apple (Available [here](#)), Cohere for AI (Available [here](#), [here](#) and [here](#)), Databricks (Available [here](#) and [here](#)), EvolutionaryScale (Available [here](#)), Inflection AI (Available [here](#) and [here](#)), LG AI Research (Available [here](#)), Meta AI (Available [here](#), [here](#) and [here](#)), Mistral AI (Available [here](#), [here](#) and [here](#)), NVIDIA (Available [here](#), [here](#) and [here](#)), OpenAI (Available [here](#), [here](#), [here](#) and [here](#)), Reka AI (Available [here](#) and [here](#)), Saudi Aramco (Available [here](#)), Suno (Available [here](#)), Writer (Available [here](#), [here](#) and [here](#)) and xAI (Available [here](#)). (All accessed: 18 June 2025).
- <sup>76</sup> Flexera (2025). State of the Cloud Report, Available [here](#). (Accessed: 28 April 2025).

## 2.2 PROVIDERS OF COMPUTE FOR GEN AI ARE GROWING ACROSS THE BOARD

- (55) Growing demand for Gen AI services is driving increased revenues across almost all cloud providers. The three largest cloud providers in Europe (AWS, Microsoft Azure, and Google Cloud) have all reported increased revenues from their cloud businesses, likely influenced by the growing demand for Gen AI services. In parallel, AI-specialised players are also reporting substantial revenue growth. While it is difficult to make direct comparisons between players due to substantial differences in baseline revenue and breadth of service portfolio, many AI-specialised players, cloud providers and on-premises providers are seeing revenue growth rates exceeding those of the three largest cloud providers, see Figure 6.

**Figure 6**  
**Several AI-specialised players are experiencing very high growth**  
Revenue growth from 2023 to 2024



Note: The figures for AWS and Google Cloud are their year-on-year total revenue growth from 2023 to 2024. For Microsoft Azure the figure is their year-on-year revenue growth of Azure and other Microsoft cloud services for the fiscal year ended June 30<sup>th</sup>, 2024. The figure for Oracle reflects year-on-year growth of cloud infrastructure services revenue in three months ended November 30<sup>th</sup>. For Dell, HPE and Supermicro, the figures show the year-on-year revenue growth of their server segments in their 2024 fiscal years. The figures for CoreWeave, Lambda, Crusoe, Cerebras, and Groq are revenue estimates as of November 2024.

Source: Copenhagen Economics based on firms' financial statements, public announcements and industry articles.<sup>77</sup>

<sup>77</sup> Amazon (2025). 2024 Annual Report, Available [here](#); Alphabet Inc. (2025). 2024 Annual Report, Available [here](#); Microsoft (2025). 2024 Annual report, Available [here](#); Oracle (2024). Oracle Announces Fiscal 2025 Second Quarter Financial Results, Available [here](#); Dell Technologies Inc. (2025). Consolidated Statements of Income and Related Financial Highlights, Available [here](#); HPE (2024). Condensed Consolidated Statements of Earnings, Available [here](#); Supermicro (2025). 2024 Annual Report, Available [here](#); Pitchbook (2024). Emerging Space Brief: AI Neoclouds, Available [here](#). (All accessed: 11 June 2025).

- (56) SemiAnalysis identifies AI-specialised players as the main incremental driver of increased demand for GPUs, with their share expected to grow to over a third of total demand.<sup>78</sup> Unlike large cloud providers, such as AWS, Microsoft Azure, and Google Cloud, these players typically offer a narrower range of services, but may offer several advantages to Gen AI developers and deployers, such as: lean, high-performance infrastructure; fast access to accelerators (GPUs); low-latency, high-bandwidth clusters optimised for Gen AI workloads;<sup>79</sup> and close partnerships with upstream providers such as NVIDIA, enabling them to be early adopters of new chips.<sup>80</sup> Their transparency on chip availability and pricing also makes them appealing to Gen AI startups, research labs, and firms with dedicated AI needs.<sup>81</sup>
- (57) Independent benchmarks, such as the GPU Cloud ClusterMax rating system recently developed by SemiAnalysis, suggest that AI-specialised providers can rival or even exceed the service quality of larger cloud providers.<sup>82</sup> Their success is further highlighted by key contracts with leading Gen AI developers, such as CoreWeave partnerships with OpenAI, Mistral, and others, see Box 3.

### Box 3: Case study

#### CoreWeave: an example of an AI-specialised player



CoreWeave is an American AI cloud provider specialising in Gen AI infrastructure. Founded in 2017, the firm originally acquired Nvidia GPUs for use in cryptocurrency mining but pivoted to providing cloud infrastructure for Gen AI following the 2018 crypto market downturn.<sup>1</sup> This strategic shift positioned CoreWeave at the forefront of Gen AI infrastructure.

Since then, CoreWeave has amassed one of the largest pools of high-performance GPUs in the world, with a fleet exceeding 250,000 units.<sup>2</sup> This large-scale, combined with innovative and reliable service, led industry analyst firm SemiAnalysis to rate them as the top GPU cloud provider on the market.<sup>3</sup> CoreWeave has attracted large Gen AI developers such as OpenAI, Mistral, Cohere, and StabilityAI as clients.<sup>4</sup> CoreWeave provides these firms with large-scale GPU clusters for training their Gen AI models, and for Cohere they are constructing a large data centre in Canada.<sup>5</sup> CoreWeave also maintains an extensive partnership with Nvidia, a major investor, supplier, and customer.<sup>2</sup> In April 2025, CoreWeave became one of the first cloud providers to make generally available Nvidia's new generation cutting-edge Grace Blackwell, with Cohere and Mistral among the first customers.<sup>6</sup>

Sources: <sup>1</sup>Wired (2024). How a Scrappy Cryptominer Transformed Into the Multibillion-Dollar Backbone of the AI Boom, Available [here](#). <sup>2</sup>Financial Times (2025). CoreWeave seeks new \$1.5bn debt deal after downsized IPO, Available [here](#). <sup>3</sup>SemiAnalysis (2025). The GPU Cloud ClusterMAX™ Rating System, Available [here](#). <sup>4</sup>CoreWeave. The AI hyperscaler for GPU Cloud Computing, Available [here](#). <sup>5</sup>DCD (2024). AI startup Cohere and CoreWeave plan multibillion-dollar data center in Canada, Available [here](#). <sup>6</sup>NVIDIA (2025). Thousands of NVIDIA Grace Blackwell GPUs Now Live at CoreWeave, Available [here](#). (All accessed: 17 June 2025)

<sup>78</sup> SemiAnalysis (2024). AI Neocloud Anatomy and Playbook, Available [here](#). (Accessed: 20 May 2025)

<sup>79</sup> See for example: SemiAnalysis (2024). AI Neocloud Anatomy and Playbook, Available [here](#). (Accessed: 20 May 2025).

<sup>80</sup> This includes CoreWeave, Nebius, Lambda and Crusoe. Financial Times (2025). NVIDIA seeks to build its business beyond Big Tech, Available [here](#). (Accessed: 20 May 2025).

<sup>81</sup> See for example: NextDC (2025). Neoclouds vs Hyperscalers: The Rise of AI-First Infrastructure (What It Means for You), Available [here](#). (Accessed: 20 May 2025).

<sup>82</sup> SemiAnalysis (2025). The GPU Cloud ClusterMAX™ Rating System, Available [here](#). (Accessed: 20 May 2025).

- (58) Investors also appear to view AI-specialised players as competitive and with growth potential. According to PitchBook, “Neocloud startups [...] have exploded in popularity. VC funding has reached USD 3.7 billion across 50 deals in 2024, up from USD 1.0 billion invested into 39 deals in 2023,”<sup>83</sup> As a result, some AI-specialised providers of compute are now reaching valuations of multiple billions of dollars, see **Table 2**.

**Table 2**  
**Recent valuations of some leading AI-specialised cloud providers**

PROVIDER	VALUATION (BILLION USD)
CoreWeave	76.1 <sup>1</sup> (June 2025)
Lambda	2.5 <sup>2</sup> (Mar 2025)
Crusoe	2.8 <sup>3</sup> (Dec 2024)
Nebius	12.1 <sup>4</sup> (June 2025)

Note: The figures are either market capitalisation or a valuation based on their latest funding round.

Sources: <sup>1</sup>Yahoo Finance. CoreWeave, Inc, Available [here](#); <sup>2</sup>Crunchbase (2025). NVIDIA Continues Torrid AI Startup Investment Pace, Outstripping Microsoft and Google, Available; <sup>3</sup>Crusoe (2024). Crusoe Closes \$600M in Series D Round at \$2.8 Billion Valuation to Power AI, Available [here](#); <sup>4</sup>Yahoo Finance. Nebius Group, Available [here](#). (All accessed: 17 June 2025).

- (59) Other cloud and on-premises providers, which now offer cloud services for Gen AI and/or systems for on-premises solutions, are also exhibiting steady growth:
- Oracle’s revenue growth of 52 per cent in cloud infrastructure services<sup>84</sup> appears to be mostly driven by its competitive offer of cloud compute for pre-training. Oracle is ranked among the top providers in SemiAnalysis’s GPU Cloud ClusterMax rating system (behind only CoreWeave).<sup>85</sup> Oracle’s advancements in AI infrastructure have attracted notable Gen AI developers as clients, including OpenAI and Cohere.<sup>86</sup>
  - On-premises providers, such as Dell, HPE and Supermicro have all experienced double-digit growth in revenues in their server and infrastructure divisions driven by Gen AI demand. They produce servers with built-in GPUs, such as NVIDIA’s recently released GB200 NVL72 AI server racks.<sup>87 88</sup>

<sup>83</sup> Pitchbook (2025). NVIDIA-backed Lambda raises 480 million as AI neocloud funding surges, Available [here](#). (Accessed: 20 May 2025).

<sup>84</sup> Infrastructure services here are interpreted as Infrastructure-as-a-Service (IaaS) cloud services.

<sup>85</sup> SemiAnalysis (2025). The GPU Cloud ClusterMAX™ Rating System, Available [here](#). (Accessed: 20 May 2025).

<sup>86</sup> Financial Times (2024). The \$200bn man: Larry Ellison’s wealth rebounds as Oracle joins AI boom, Available [here](#). (Accessed: 9 June 2025).

<sup>87</sup> These racks enable 72 of NVIDIA’s latest GPUs to act as one and are one of the most powerful out-of-box compute options with embedded accelerators currently on the market. Dell, HPE, and Supermicro sell these systems with their own cooling systems included.

<sup>88</sup> These products include, among others, the Dell AI Factory enterprise-grade AI servers, see Dell Technologies Inc. (2024). Dell Technologies Expands Dell AI Factory with NVIDIA to Turbocharge AI Adoption, Available [here](#). (Accessed: 02 July 2025); HPE Private Cloud AI services, see HPE (2024). Hewlett Packard Enterprise and NVIDIA announce ‘NVIDIA AI Computing by HPE’ to accelerate generative AI industrial revolution, Available [here](#). (Accessed: 02 July 2025); and Supermicro rack-scale AI compute solutions, see PR Newswire (2025). Supermicro Unveils Industry’s Broadest Enterprise AI Solution Portfolio for NVIDIA Blackwell Architecture to Accelerate AI Factory Deployments in European Market, Available [here](#). (Accessed: 02 July 2025).

- (60) The observed revenue growth across a diverse set of compute providers, with some specialised and on-premises providers growing faster than the largest cloud providers, suggests that new entrants can be competitive with existing players. These outcomes are not consistent with concerns that developers and deployers are reliant on just a small number of firms for access to compute.

## 2.3 SUBSTANTIAL INVESTMENT IN COMPUTE CAPACITY

- (61) Firms and public organisations are investing heavily to expand compute capacity for Gen AI workloads.<sup>89</sup> This ongoing rise in investment, both private and public, suggests a fast-moving sector that is undergoing significant transformation.
- (62) Research by Dell’Oro Group forecasts that global investments in data centres will reach USD 1 trillion annually by 2029, with about half directed towards servers optimised for AI training and domain-specific workloads.<sup>90</sup>
- (63) Currently, the three largest cloud providers, AWS, Google Cloud, Microsoft Azure, continue to make significant investments globally. Specifically, Microsoft recently announced its plan to invest USD 80 billion in 2025 “to build out AI-enabled datacentres to train AI models and deploy AI and cloud-based applications around the world”<sup>91</sup> while AWS’s expected capital expenditure for 2025 is USD 100 billion, of which the “vast majority” will go towards making AI capabilities available to AWS customers;<sup>92</sup> and Google has planned USD 75 billion of capital expenditure for 2025 to accelerate its AI ambitions.<sup>93</sup> These growing investments indicate the growing importance for cloud providers to position themselves in the evolving Gen AI landscape, and are suggestive of intensifying competition.<sup>94</sup>

---

<sup>89</sup> We assume most of this is targeted towards Gen AI specifically.

<sup>90</sup> Dell’Oro Group (2025). Data Center capex to surpass 1 trillion by 2029, Available [here](#). (Accessed: 20 May 2025).

<sup>91</sup> Microsoft (2025). The Golden Opportunity for American AI, Available [here](#). (Accessed: 10 June 2025)

<sup>92</sup> CNBC (2025). Amazon plans to spend \$100 billion this year to capture ‘once in a lifetime opportunity’ in AI, Available [here](#). (Accessed: 10 June 2025).

<sup>93</sup> Although it is not clear whether those investments are also partially related to the development of Google’s own Gen AI models and services rather than their cloud services for third party Gen AI developers and deployers. See: Alphabet (2025). Alphabet Announces Fourth Quarter and Fiscal Year 2024 Results, Available [here](#). (Accessed: 10 June 2025).

<sup>94</sup> Cerre (2025). “However, a more dynamic analysis suggests that the largest three firms might be under more competitive pressure than market share figures would assume. [...] Together with the extraordinarily high level of investments being made by the hyperscalers, with cloud computing providers investing about \$250 billion in global investment for AI in 2025 alone, suggests that competition between the largest cloud computing companies is intensifying” Cerre (2025). A Competition policy for cloud and AI, Page 16, Available [here](#). (Accessed: 2 June 2025).

- (64) Dell’Oro research expects that other players will also increase their capital expenditures significantly. Several initiatives, both private and public, are contributing to this trend. In Europe, the European Commission’s AI Continental Action Plan, launched in April 2025, sets out to mobilise EUR 200 billion (via a mix of public and private sources) to develop up to five AI “gigafactories” across strategic sectors such as healthcare and science.<sup>95</sup> <sup>96</sup> The plan also includes a dedicated EUR 20 billion European AI fund and builds on a prior EUR 10 billion programme supporting 13 AI factories across 17 member states between 2021 and 2027.<sup>97</sup> In the United States, the Stargate Project—a collaboration between SoftBank, OpenAI, Oracle, and MGX—aims to mobilise over USD 500 billion in AI-related infrastructure.<sup>98</sup>
- (65) Complementing EU-level efforts are a number of national and public-private initiatives. One recent example is NVIDIA’s announced initiative to expand AI infrastructure across Europe. The firm is working with governments, local cloud providers, telecom operators, and startups in countries such as France, Germany, Italy, Spain, and the UK. As part of this effort, NVIDIA plans to deploy more than 3,000 exaflops of computing capacity.<sup>99</sup> The initiative also includes the creation of AI factories and technology centres, including the world’s first industrial AI cloud dedicated to European manufacturers, located in Germany.<sup>100</sup> Finally, a non-exhaustive list of recent major announcements suggests more than EUR 48 bn in investments in compute infrastructure for AI (including Gen AI) in the next 5 years in Europe.<sup>101</sup>

---

<sup>95</sup> European Commission (2025). AI Action Plan, Available [here](#). (Accessed: 20 May 2025). These initiatives are not limited to pure infrastructure investments but cover other elements such as research and skills development.

<sup>96</sup> These gigafactories, large-scale facilities equipped with over 100,000 advanced accelerators (at least four times more powerful than AI factories), will be modelled on the “CERN-for-AI” public-private collaborative framework and will prioritize open access for startups and scale-ups. See: European Commission (2025). EU launches InvestAI initiative to mobilise €200 billion of investment in artificial intelligence, Available [here](#). (Accessed: 20 May 2025).

<sup>97</sup> European Commission (2025). AI Factories, Available [here](#). (Accessed: 20 May 2025).

<sup>98</sup> OpenAI (2025). Announcing The Stargate Project, Available [here](#).

<sup>99</sup> As a reference, the current Europe’s fastest supercomputer, JUPITER, has a computing power of nearly 800 petaflops and is on track to soon achieve 1 exaflop, making it Europe’s first exascale supercomputer. See: EuroHPC (2025). EuroHPC Supercomputers Put Europe at the Forefront of Global Supercomputing, Available [here](#).

<sup>100</sup> NVIDIA Newsroom (2025). Europe builds AI Infrastructure with NVIDIA to fuel region’s next industrial transformation, Available [here](#). (Accessed: 16 June 2025).

<sup>101</sup> DCD (2024). CoreWeave to invest £1bn in UK data centers, Available [here](#); CoreWeave (2024). CoreWeave's European Expansion, Available [here](#); Nebius (2024). Nebius to invest more than USD 1 billion to build AI infrastructure in Europe, Available [here](#); <sup>9</sup>UK Government (2025). Prime Minister sets out blueprint to turbocharge AI, Available [here](#); Iliad Group (2025). The Iliad group is investing €3 billion in AI, Available [here](#); DCD (2025). Brookfield and Data4 to spend €20bn on AI infrastructure in France, Available [here](#); Fortune (2025). Oracle bets big on U.K. AI boom with \$5 billion cloud investment, Available [here](#); ECB (2025). Euro foreign exchange reference rates, Available [here](#). (All accessed: 16 May 2025).

- (66) While these investment plans point to a substantial increase in global data centre capacity, it remains uncertain whether supply will fully keep pace with rapidly growing demand for Gen AI workloads. Other factors such as power or permitting constraints may limit the expansion in certain regions. Conversely, there are also some indications that the pace of AI-related expansion may be slowing down.<sup>102</sup> According to Reuters, Microsoft recently cancelled data centre lease agreements with two private operators, amid oversupply concerns.<sup>103</sup>

## 2.4 EVIDENCE SUGGESTS DECLINING PRICES AND AN EXPANDING SERVICE OFFER

- (67) While some Gen AI firms expect an increase in overall cloud spending in 2025, this appears to be reflective of growing demand and higher usage volumes rather than increasing prices.<sup>104</sup> A detailed analysis of pricing is beyond the scope of this study – however, available data suggests that the price of compute for Gen AI has been declining in recent years, which is consistent with increasing competitive pressure.<sup>105</sup> In parallel, evidence suggests that cloud providers are continuously expanding their service offer to meet Gen AI developers and deployers needs.
- (68) First, analysis by Silicon Data on the prices for GPU access of an unspecified leading cloud provider, indicates that the inflation-adjusted price per FP32 FLOP (a standard measure of GPU processing performance)<sup>106</sup> has decreased by approximately 74 per cent since 2019, see Figure 7.

---

<sup>102</sup> See, for instance, CNBC (2025). AI data center boom isn't going bust but the 'pause' is trending at big tech companies, Available [here](#) (Accessed: 16 October 2025).

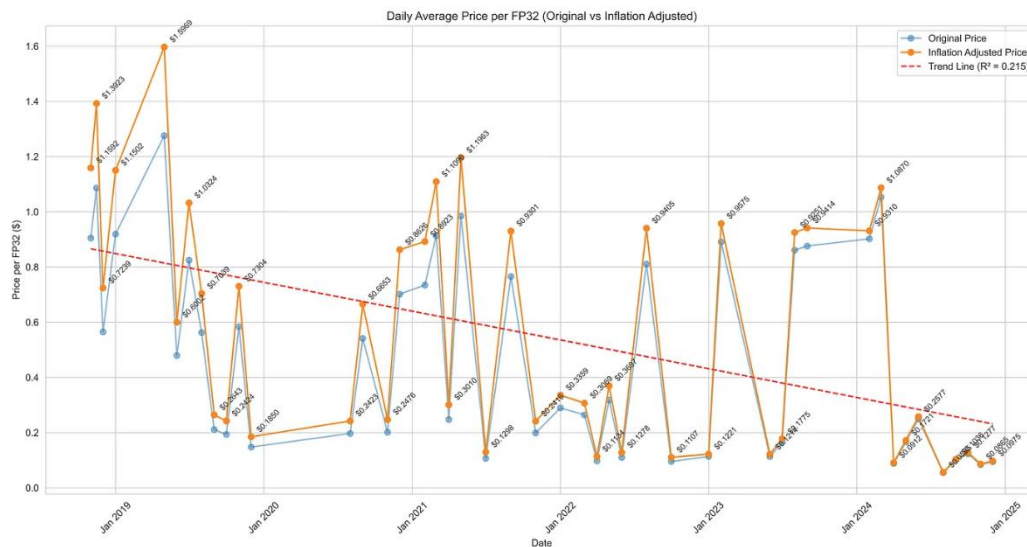
<sup>103</sup> Reuters (2025). Microsoft data center leases slowing, analysts say, raising investor attention, Available [here](#). (Accessed: 20 May 2025).

<sup>104</sup> S&P Global Market Intelligence reported that nearly 70 per cent of cloud users expect their firm's public cloud spending (not limited to Gen AI workloads) to increase in 2025. However, it appears that this trend is explained mostly by expected developments in volumes and composition of demand rather than prices. Only 34 per cent of respondents cited higher cloud provider prices among the primary factors driving increased spending, while in most cases respondents reported greater demand for advanced AI services (55 per cent), new IT initiatives (49 per cent), and workload migration (47 per cent). From a competition perspective, this distinction is relevant. In the presence of a supply constraint, rising demand would typically be expected to lead to rising unit prices. Instead, there is high-level evidence that unit prices have been declining recently, even as usage expands. See S&P Global (2025). Cloud Price Quarterly: New services, regions and prices to kick off the year, Available [here](#). (Accessed: 9 June 2025).

<sup>105</sup> The trend of declining prices may be consistent with broader patterns observed in standard cloud computing in earlier periods. For example, AWS's list prices have been found to decline at an average annual rate of 7% between 2009 and 2016 (Byrne, D et al. (2021). Available [here](#)), and by a total of 74% between 2010 and 2019 (Coyle, D. & Ngyuen, D. (2019). Available [here](#)).

<sup>106</sup> FLOP stands for floating point operations per second. As FLOPs capture the amount of computational power delivered, this metric can be seen as a proxy for quality-adjusted prices of compute.

**Figure 7**  
**Prices for compute have been declining**  
Daily average price for a computing operation of fixed size (FP32 FLOP)



Note: The cost of performing one 32-bit floating point operation (FP32 FLOP) is a standard measure of compute performance in Gen AI and can be seen as a proxy for quality-adjusted prices of compute.

Source: Li, C (2025). The Evolution of GPU Pricing: A Deep Dive into Cost per FP32 FLOP for Hyperscalers, Available [here](#). (Accessed: 18 June 2025)

- (69) Silicon Data suggests that prices stabilising at lower levels after 2023 may indicate “greater market competition, with [large cloud providers] and new entrants driving down cloud GPU rental costs” together with better cost optimisation and economies of scale in the “cloud GPU market” and more efficient hardware and software.<sup>107</sup> Some evidence suggests that price competition may have intensified recently due to competitive pressure from AI-specialised players. According to Forbes, AI-specialised players have been “massively undercutting giants like AWS and Oracle on price” – although they also often offer a slimmer service.<sup>108</sup> Relatedly, other industry commentators have pointed to a decline in the rental prices of the popular NVIDIA H100 GPUs offered by providers of GPU as a service from 2023 highs of USD 8 per hour to less than USD 2 per hour in 2024.<sup>109</sup> Furthermore, the rental prices of older processors, which remain viable for inference, are likely to keep decreasing as chip manufacturers continue to produce newer and more powerful hardware.<sup>110</sup>

<sup>107</sup> Li, C (2025). The evolution of GPU pricing: a deep dive into cost per FP32 FLOP for hyperscalers, Available [here](#). (Accessed: 18 June 2025). However, the relative importance of these factors cannot be disentangled based on the available data.

<sup>108</sup> Forbes (2024). Investors’ \$20 Billion Bet On The ‘NeoClouds’ Driving The AI Arms Race, Available [here](#). (Accessed: 21 June 2025).

<sup>109</sup> These figures refer to the prices set by firms offering GPU rental, such as CoreWeave. We note that while the source refers to US prices, European cloud providers of GPUs such as Genesis and Datacrunch also offer NVIDIA H100s at just above USD 2 per hour. See: CNBC (2025). The concern with CoreWeave’s 250,000 NVIDIA chips ahead of its IPO, Available [here](#). (Accessed: 10 June 2025).

<sup>110</sup> Cheah, E (2024). \$2 H100s: How the GPU Rental Bubble Burst, Available [here](#). (Accessed: 10 June 2025)

- (70) Second, according to S&P Global Market Intelligence’s Cloud Price Index, the prices of compute for Gen AI from the three largest cloud providers (AWS, Microsoft Azure and Google Cloud), appear to be declining, with several price cuts identified.<sup>111</sup> For example, AWS has just recently announced price cuts up to 45 per cent for virtual machines with NVIDIA GPUs.<sup>112</sup>
- (71) Third, in many cases, Gen AI developers and deployers<sup>113</sup> do not report that the cost of compute is a main barrier to developing and deploying Gen AI solutions. According to ClearML’s “the State of AI Infrastructure at Scale 2024” report, less than 10 per cent of respondents ranked “Financial costs of compute” as their main concern.<sup>114</sup> Similarly, according to the Google Cloud study, “cost” (although not specific to compute and not decomposed into price and volume effects) ranked 4<sup>th</sup> out of 11 challenges for “Gen AI adoption”.<sup>115</sup>
- (72) In parallel with price decreases, evidence suggests that firms are expanding their service offer, with nearly all of the largest cloud providers having introduced or expanded their service offer to Gen AI developers and deployers over the past five years. This is consistent with a constantly evolving market consistent with intensifying competition. Specifically, seven out of eight of the largest cloud providers at the global level according to Statista<sup>116</sup> (AWS, Microsoft Azure, Google Cloud, Oracle, IBM, Alibaba and Tencent) now offer, among others: (i) direct access to accelerators, (ii) tools to support developers in training and fine-tuning, and (iii) an AI platform to access existing off-the-shelf models (e.g. GPUs), see Box 4.<sup>117</sup>

---

<sup>111</sup> S&P Global (2025). Hyperscaler GenAI gets cheaper as pricing finds its level, Available [here](#). (Accessed: 10 June 2025).

<sup>112</sup> AWS News Blog (2025). Announcing up to 45% price reduction for Amazon EC2 NVIDIA GPU-accelerated instances, Available [here](#). (Accessed: 18 June 2025).

<sup>113</sup> This includes also Gen AI users.

<sup>114</sup> We interpret this finding to apply to Gen AI specifically. We note that in another question of the same survey “Compute limitations (availability, compute costs)” was reported as a top challenge for “scaling AI”. However, this could reflect a concern that total costs would run up when volume increases – rather than the price per unit being unreasonably high.

<sup>115</sup> ClearML (2024). The State of AI infrastructure at scale 2024, p.9, Available [here](#). (Accessed: June 10 2025). We note that cost could be interpreted as “total cost” which would be a function of cost per unit and volumes. It remains unclear how respondents interpreted the question when answering.

<sup>116</sup> Statista (2025). Amazon and Microsoft Stay Ahead in Global Cloud Market, Available [here](#). (Accessed: 9 June 2025). Salesforce rounds out the top eight, but they do not have an offer to support Gen AI development or deployment.

<sup>117</sup> While not a traditional cloud provider, NVIDIA also provides software services NeMo and AI foundry through which users can develop AI agents and applications, and an inferencing service in the form of NVIDIA NIM. See: NVIDIA (2025). Generative AI runs on NVIDIA, Available [here](#). (Accessed: 9 June 2025).

**Box 4**

**Examples of Gen AI  
related services  
offered to  
developers and  
deployers**



AWS offers a managed platform for developing and deploying Gen AI models via AWS SageMaker, access to existing Gen AI models for inference via AWS Bedrock, and compute for training via Elastic compute cloud services.



Microsoft Azure allows cloud user to develop custom Gen AI applications and agents in the Azure AI Foundry platform, based off an extensive list of Gen AI models in Azure's AI Model Catalogue. Recent updates to the service have focused on improving support for multi-agent solutions and better risk identification. More direct compute access is available via Azure virtual machines.



Google Cloud

Google Cloud has developed its Vertex AI platform offering a full suite of Gen AI development services from building custom agents with no code to training users' own models. Google's Model Garden contains a selection of Gen AI models that users can access and fine-tune. Compute power is available via Google compute engine which also features innovative scheduling via their Dynamic Workload Scheduler.



IBM's WatsonX platform enables users to develop and deploy Gen AI services into their applications. IBM also provides pre-built Gen AI agents that customers can adopt into their services via the WatsonX Orchestrate platform. On-demand access to Nvidia GPUs and Intel's Gaudi accelerators is possible through IBM cloud. Recently, IBM has announced new features that improve AI risk evaluation and management on WatsonX.



Oracle Cloud infrastructure's Generative AI is a platform that allows users to adopt LLMs for various applications either as they are, or with custom fine-tuning. Oracle also offers compute for pre-training, fine-tuning and inference via the OCI Supercluster service, which due to recent engineering innovations now provides unprecedented scale and performance.



Alibaba Cloud Platform for AI (PAI) is a one-stop machine learning platform that enables users to develop and deploy Gen AI models using several built-in optimisation tools such as PAI Blade. Alibaba Cloud also offers GPU cloud services.



Tencent Cloud offers an all-in-one Gen AI model development and deployment service via their TI Platform. Users can benefit from several services such as multi-instance scheduling and auto-tuning of parameters. Tencent also provides GPU rental services.

Source: Copenhagen Economics based on cloud providers' publicly available information.<sup>118</sup>

- (73) Further supporting the evolution in service offer, data from S&P Global Market Intelligence's Cloud Price Index shows that, in 2024 alone, AWS, Microsoft Azure, and Google Cloud collectively added approximately six thousand new Stock Keeping Units (SKUs) – that is, individual service features or configurations – across their AI platforms for hosting and accessing Gen AI models, respectively Amazon Bedrock, Azure OpenAI, and Google Vertex.<sup>119</sup> According to S&P Global Market Intelligence, these additions outnumbered price changes by a factor of 20 to 1, highlighting a strong focus on expanding and refining service portfolios without corresponding price increases.<sup>120</sup>

---

<sup>118</sup> AWS. Generative AI, Available [here](#); Microsoft Azure. Azure AI solutions, Available [here](#); Google Cloud. AI and machine learning products, Available [here](#); Google Cloud (2023). Dynamic Workload Scheduler, Available [here](#); IBM. Explore our AI solutions, Available [here](#); IBM (2025). IBM enhances the capabilities of watsonx.governance, Available [here](#); Oracle. Artificial Intelligence, Available [here](#); Oracle (2024). Now Generally Available: The Largest, Fastest AI Supercomputer in the Cloud, Available [here](#); Alibaba Cloud (2024). An Introduction to Alibaba Cloud Platform for AI, Available [here](#); Alibaba Cloud (2023). Introduction to Alibaba Cloud GPU Service, Available [here](#); Tencent Cloud. Tencent Cloud TI Platform, Available [here](#); Tencent Cloud. Cloud GPU service, Available [here](#). (All accessed: 17 June 2025).

<sup>119</sup> S&P Global (2025). Hyperscaler GenAI gets cheaper as pricing finds its level, Available [here](#). (Accessed: 10 June 2025). As also recognised by S&P, this method has some limitations – for instance, a SKU added in one month and removed the next would count as two separate changes, despite cancelling each other out, and some SKU changes may reflect corrections rather than substantive updates. However, it still provides a useful proxy for identifying where and how providers are expanding their service offerings.

<sup>120</sup> We note however that these additions are not specific to direct access to compute for development of new Gen AI models but rather seem to be related to deployment (inference).

### 3 VARIOUS EXPECTED DEVELOPMENTS MAY FURTHER STRENGTHEN COMPETITION

#### Key findings

- Demand for Gen AI workloads is expected to shift down the Gen AI value chain towards fine-tuning and inference, with 80 per cent of workloads expected to be used for inference tasks by 2028, up from 40 per cent in 2023. This type of less computationally intensive workloads can potentially be provided by a wider range of compute providers, increasing contestability and competitive pressure.
- Models are becoming more efficient: smaller models are becoming more capable, and inference costs have fallen between 9x and 900x per year reducing reliance on providers of large number of accelerators.
- On-device solutions for inference workloads are expected to become increasingly important reducing reliance on cloud or on-prem.

(74) In the previous chapter, we assessed current market outcomes and the extent to which they are consistent with well-functioning competition in the provision of compute for Gen AI. This chapter focuses on expected developments that may affect the supply of and demand for compute for Gen AI going forward. While growth in Gen AI services is likely to lead to increased demand for compute overall, several trends – including shifts in workload patterns and advances in model efficiency – mean that the compute needs of individual Gen AI developers and deployers could in future potentially be served more easily by an increasing number of compute providers.

(75) Specifically, in this chapter we will cover:

- The expected shift towards less computationally intensive workloads (Section 3.1);
- Ongoing and expected innovations in model efficiency (Section 3.2).

#### 3.1 INDUSTRY TRENDS SUGGEST DEMAND IS SHIFTING TOWARDS LESS COMPUTATIONALLY INTENSIVE WORKLOADS

(76) In the early stages of Gen AI development, activities were mostly concentrated on pre-training of large Gen AI models<sup>121</sup>, which often requires substantial upfront investments to access large amount of compute from high-end infrastructure. However, there is growing industry consensus that this is changing with implications for access to compute.

<sup>121</sup> “From the start of the current generation of LLMs in 2017 until mid-2024, leading LLMs invested heavily in the AI model pre-training phase that consumes huge volumes of training data (or tokens) and computing power to estimate billions of internal parameters. LLMs exhibited a scaling law”. See: Bruegel (2025). How DeepSeek has changed artificial intelligence and what it means for Europe, Available [here](#). (Accessed: 20 May 2025).

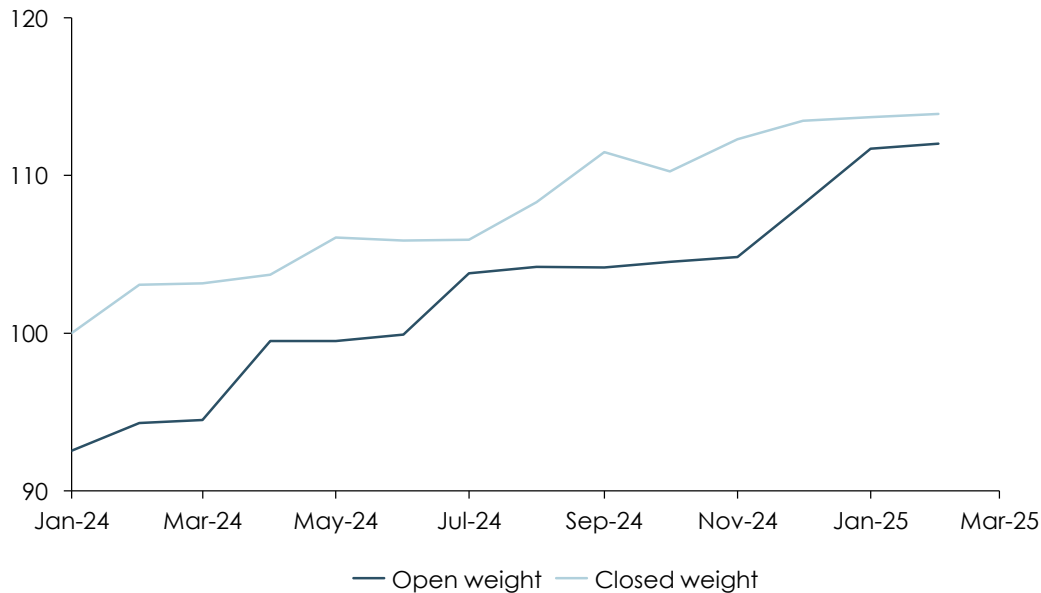
- (77) First, a growing share of Gen AI development activities are now concentrated in fine-tuning of existing models. S&P's Global Market Intelligence listed "*models [becoming] more specialised and domain-specific*" as one of the main "2025 Trends in Data, AI and Analytics". According to S&P Global Market Intelligence's AI & Machine Learning Infrastructure 2024 survey, 67 per cent of firms using Gen AI technologies are fine-tuning an existing Gen AI model.<sup>122</sup> At the same time, the growing adoption of post-training techniques such as model distillation – where a large pre-trained model (the "teacher") is compressed into a smaller, more efficient version (the "student") while preserving much of its performance (see following section) – enables models to build on top of existing ones, allowing Gen AI development with progressively lower computational requirements.
- (78) The availability of competitive open weight (which can be downloaded, adapted, and deployed independently of the distribution platforms) models means that Gen AI developers and deployers can build on top of state-of-the-art models, which further supports the shift towards fine-tuning. Although closed weight models continue to lead on some benchmarks, open weight models appear to be only a few months behind, and the performance gap appears to be narrowing, see Figure 8.

---

<sup>122</sup> According to S&P's Global Market Intelligence AI & Machine Learning, Infrastructure 2024 survey, "67% of organisations using Gen AI technologies are fine-tuning a pretrained foundation model, reflecting a pressing need to improve output accuracy and relevance for their users. A shift toward model specialisation presents opportunities for a wide range of technology vendors – beyond the well-funded companies with frontier models – to add value and differentiate themselves in an intensely crowded AI landscape." See: S&P Global Market Intelligence (2025). 2025 Trends in Data, AI & Analytics, Available [here](#). (Accessed: 9 June 2025).

**Figure 8**  
**Open weight models have comparable performance to closed weight models**

Chatbot Arena score of top performing model (indexed to January 2024 closed weight = 100)



Note: The Chatbot Arena ranking developed by LMSYS uses Elo ratings based on crowd-sourced pairwise comparisons: users are shown two anonymized model responses to the same prompt and vote for the better one. This system aggregates thousands of votes to produce a dynamic leaderboard that reflects relative model performance in human preference.<sup>123</sup>

Source: Copenhagen Economics based on HAI (2025). The 2025 AI Index Report, Available [here](#). (Accessed: 10 June 2025)

- (79) Second, inference is also growing in importance as Gen AI moves from experimentation to deployment at scale. More complex applications and more pervasive adoption of Gen AI is expected to drive demand for cost-effective compute and solutions optimised for inference tasks.<sup>124</sup> Gartner has projected that, by 2028, more than 80 per cent of accelerators deployed in data centres will be used to execute Gen AI inference workloads, up from 40 per cent in 2023.<sup>125</sup>

<sup>123</sup> LMArena (2025). Chatbot Arena, Available [here](#). (Accessed: 9 June 2025).

<sup>124</sup> S&P Global Market Intelligence (2025). 2025 Trends in Data, AI & Analytics, Available [here](#). (Accessed: 20 May 2025).

<sup>125</sup> Gartner (2024). Forecast Analysis: Semiconductors and Electronics, worldwide, Available [here](#). (Accessed: 20 May 2025). While the report refers to AI inference workloads in general, we interpret this as Gen AI inferencing. Similar sentiment was reported by e.g. Morgan Stanley which estimated that more than 75 per cent of power and computational demand for data centres in the US will be for inference in the coming years, though they warned of “significant uncertainty over exactly how the transition will play out. See Financial Times (2025). How ‘inference’ is driving competition to NVIDIA’s AI chip dominance, Available [here](#). (Accessed: 10 June 2025).

- (80) Accordingly, industry commentators expect that future investment will be increasingly directed toward this type of workloads. Microsoft recently announced it is “refocusing on inference,” prioritising the optimisation of lower-cost, high-efficiency infrastructure over continued expansion of high-cost training capabilities.<sup>126</sup>
- (81) The shift towards fine-tuning and inference has potential implications for competition in the provision of compute:
- Fine-tuning of existing models typically requires less sophisticated compute infrastructure than pre-training. According to European think tank Bruegel, “*the approach to model training based on training-data collection, pre-training and fine-tuning by a single firm is being replaced by a more horizontal networked model,*” which can reduce the compute requirements for individual downstream players.<sup>127</sup> This may expand the range of firms able to serve Gen AI developers and deployers and make it easier for smaller compute providers to compete.
  - Inference workloads can be distributed across different infrastructure types and often prioritise low latency and proximity to end-users. This reduces the reliance on dense clusters of high-end accelerators and creates more opportunities for smaller providers or local deployment solutions. Inference-specific hardware (accelerators), which tends to be more cost-effective and subject to greater supplier competition, further broadens the pool of potential providers.<sup>128</sup>

## 3.2 INNOVATION IN COMPUTE USAGE

- (82) Similar to the potential shift towards less computationally intensive workloads, several innovations in compute usage may also reduce reliance on large-scale supply of cutting-edge compute infrastructure.

### 3.2.1 Smaller models are becoming more capable and inference costs are falling

- (83) Recent developments show a trend towards smaller, more efficient models that require less compute, both in training and in inference, to achieve comparable performance levels, thereby broadening the range of actors who can develop and deploy Gen AI.<sup>129</sup>
- (84) One of the most visible signs of this trend is the emergence of Small Language Models (SLMs), which are designed to have fewer parameters than traditional Large Language Models (LLMs) while remaining effective for specific tasks.

---

<sup>126</sup> Mustafa Suleyman (CEO, Microsoft AI) has explained that, while pre-training of large models remains important, the computational intensity of training is flattening, and firms are finding higher ROI in inference and downstream AI services. He pointed to a deceleration in the return on pure scale increases during training (more tokens and parameters) and emphasized a focus on optimization, customization, and deployment. See: Business Insider (2025). Microsoft is taking its foot off the AI accelerator. What does that mean? Available [here](#). (Accessed: 20 May 2025).

<sup>127</sup> Bruegel (2025). How DeepSeek has changed artificial intelligence and what it means for Europe, Available [here](#). (Accessed: 20 May 2025).

<sup>128</sup> Financial Times (2025). How ‘inference’ is driving competition to NVIDIA’s AI chip dominance, Available [here](#). (Accessed: 10 June 2025).

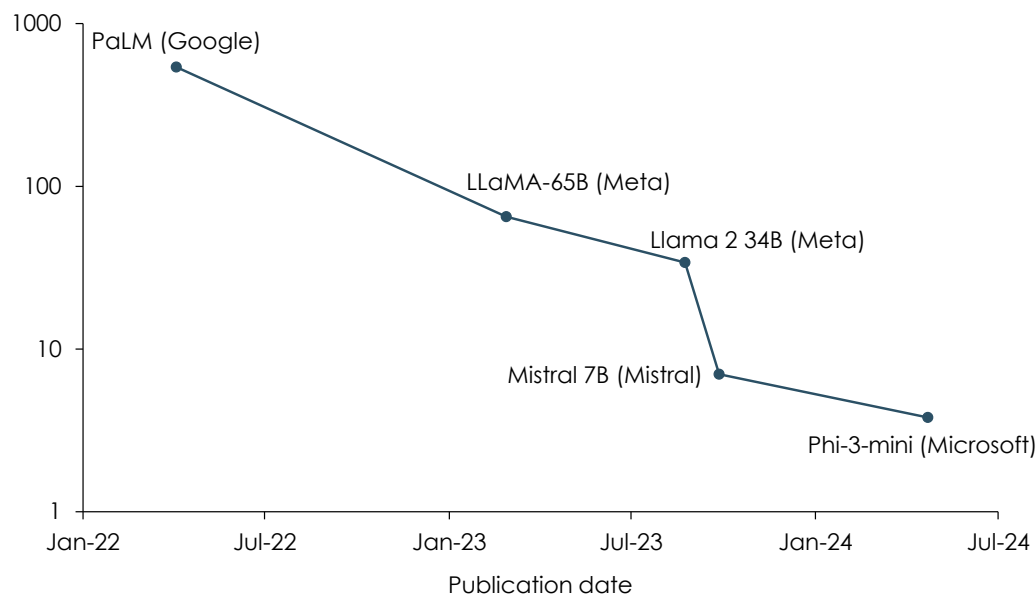
<sup>129</sup> Epoch AI (2024). Frontier language models have become much smaller, Available [here](#). (Accessed: 9 June 2025).

- (85) The performance of small models continues to improve significantly over time. In 2024, models with less than 4 billion parameters (Microsoft Phi-3-mini) could achieve the same performance as a 540 billion parameter model (Google PaLM) just two years prior, see Figure 9.

**Figure 9**

**High performance is being reached by increasingly smaller models over time**

Number of parameters in smallest model to reach MMLU score above 60 per cent (billions, log scale)



Note: The Massive Multitask Language Understanding (MMLU) test is a widely used language model benchmark designed to evaluate a model's real-world knowledge and problem solving.

Source: Copenhagen Economics based on HAI (2025). The 2025 AI Index Report, Available [here](#). (Accessed: 10 June 2025)

- (86) Small language models are less demanding to train<sup>130</sup> and more practical to deploy in resource-constrained settings, such as edge devices or other settings with reduced availability of compute (see Section 3.2.3). Notably, several leading AI developers have introduced their own SLMs, including Google's Gemma, Microsoft's Phi, Meta's Llama 3 8B, and Mistral's 7B model. These SLMs are often tailored for fine-tuning and task-specific inference, making them more accessible to a broader set of developers and deployers.

<sup>130</sup> For example, in the LLaMA2 family, the smallest model (7b) took about 184k GPU hours to train, while the largest took 1.72m, representing a roughly 9x difference. See Table 2 in Touvron et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. Available [here](#). (Accessed: 5 August 2025).

- (87) Additionally, inference efficiency in general (not limited to small models) has seen substantial gains, with the cost of generating model outputs at a given performance level dropping sharply over the past couple of years. According to analysis by Epoch.ai, inference costs have declined dramatically, with reported annual reductions ranging from 9x to 900x, depending on the model and use case. For instance, the cost of achieving the level of performance associated with GPT-4 on a set of PhD-level science questions has fallen by approximately 40-fold per year.<sup>131</sup>

### 3.2.2 Model efficiency is driven by continued innovation in model design and hardware

- (88) Gains in computational efficiency are primarily driven by advances in (i) model architecture and training techniques and (ii) the underlying hardware used to support Gen AI workloads.
- (89) First, Gen AI developers are experimenting and adopting several advanced model architecture and training techniques to achieve efficiency, reducing compute needs. Notable examples include:
- Model distillation: where a smaller “student” model learns to replicate the behaviour of a larger, pre-trained “teacher” model. This method reduces the size and compute requirement of the resulting model while preserving much of its performance.<sup>132</sup> Leading Gen AI developers e.g. Google, Anthropic, OpenAI, Meta and DeepSeek are introducing distillation into their larger models.
  - Mixture-of-Experts (MoE): where only a small subset of the model’s parameters (experts) are activated for any given input. During training, this means that only a limited portion of the model is updated at each step, significantly reducing compute and memory requirements compared to fully dense models.<sup>133</sup>
- (90) A relevant example of this trend is DeepSeek which released an open weight LLM (DeepSeek V3) as well as a reasoning model<sup>134</sup> (DeepSeek R1) and claimed that they were trained using extremely limited compute resources compared to other state-of-the-art Gen AI models. Despite this, DeepSeek’s models achieved competitive performance across a range of benchmarks. The model’s efficiency was enabled by a combination of architectural innovations and training optimisations. We note that DeepSeek was neither affiliated with, nor supported by, any major cloud providers when released its LLM and reasoning models, showing that new market entrants can emerge from varied backgrounds. DeepSeek’s approach demonstrates how smart optimisation strategies can reduce reliance on a large-scale supply of cutting-edge hardware, see Box 5.

---

<sup>131</sup> The range in inference cost decline reflects differences in the performance benchmarks used. The slowest decline (9x) corresponds to the cost of achieving GPT-3.5 Turbo (March 2023 -level performance on general knowledge tasks. The fastest decline (900x) applies to achieving GPT-4o (May 2024) -level performance on Ph.D.-level science questions. See: Epoch AI (2025). LLM inference prices have fallen rapidly but unequally across tasks, Available [here](#). (Accessed: 9 June 2025).

<sup>132</sup> BuiltIn (2025). What is model distillation? Available [here](#). (Accessed: 9 June 2025).

<sup>133</sup> Similarly, at inference time, sparse activation allows the model to perform fewer computations while maintaining high overall capacity, resulting in reduced latency and lower resource usage.

<sup>134</sup> A Gen AI reasoning model is designed to perform multi-step logical or analytical tasks by interpreting input, drawing inferences, and generating structured or contextually coherent outputs based on learned patterns.

**Box 5: Case Study****DeepSeek trained  
frontier models with  
limited compute**

Founded in 2023, DeepSeek is a Chinese technology firm that has quickly emerged as a notable player in the field of AI. Only being in existence for just over a year and half, it gained widespread public attention in January 2025, following the release of its DeepSeek-R1 model. This model introduced a more efficient methodology for developing AI systems, which distinguished the firm from its competitors.<sup>1</sup>

The earlier DeepSeek-V3 was released in December 2024 as an open weight model<sup>2</sup> and is comparable with the leading models from large AI firms, including Google's Gemini.<sup>3</sup> Despite China facing export restrictions on advanced semiconductors, which limit access to cutting-edge Nvidia chips, DeepSeek has successfully trained its models using alternative hardware solutions.<sup>4</sup>

In terms of technical claims, DeepSeek has stated that training its V3 model — which comprises 671 billion parameters — required only 2,048 GPUs and 5.6 million dollars, which is particularly notable when compared to Meta's Llama 3 model, which has 405 billion parameters but reportedly required eleven times the compute to train.<sup>5</sup> Furthermore, the R1 model was trained using the V3 model through a technique called distillation, which reduced the overall computational requirements for its training. When released, R1 matched the performance levels of OpenAI's o1 model with an overall GPU cost of just 1 million dollars after the pre-training.<sup>6</sup>

According to industry experts, DeepSeek employed a series of techniques to achieve computational efficiency and performance:

- **Multi-Head Latent Attention (MLA)**<sup>7</sup>: In simplified terms, this method compresses input data more effectively, reducing memory usage and speeding up processing times. As a result, it contributes to improved model performance with fewer computational resources.
- **Mixture of Experts (MoE)**<sup>8</sup>: This machine learning approach divides tasks among specialised expert networks, each responsible for handling a specific aspect of the problem, thereby increasing overall efficiency and specialisation within the model.
- **Multi-Token Prediction**<sup>9</sup>: Unlike generating predictions token-by-token, this technique allows the model to predict entire sequences or sentences at once. This significantly accelerates the data generation and processing pipeline.

Sources: <sup>1</sup>Forbes (2025). Here's how much AI firm DeepSeek and its founder are worth, Available [here](#); <sup>2</sup>DeepSeek (2024). Introducing DeepSeek-V3, Available [here](#); <sup>3&5</sup>Unite.AI (2024). How DeepSeek Cracked the Cost Barrier with \$5.6M, Available [here](#); <sup>4</sup>The Science Survey (2025). A Deep-Dive Into DeepSeek, Available [here](#); <sup>6</sup>Epoch AI (2025). What went into training DeepSeek-R1, Available [here](#); <sup>7-9</sup>Wang, C & Kantarcioglu, M (2025). A Review of DeepSeek Models' Key Innovative Techniques, Available [here](#). (All accessed: 17 June 2025)

- (91) Second, continued progress in hardware development, and innovation by semiconductor startups and existing manufacturers upstream, may also strengthen competition in the provision of compute, by increasing the number of players and/or reducing entry barriers. As discussed in section 2.4, the price of compute has been declining in recent years, with performance per dollar improving around 30 per cent each year.<sup>135</sup> This helps to lower the infrastructure costs associated with developing and deploying Gen AI models.

### 3.2.3 On-device is emerging as a viable alternative for some inference workloads

- (92) SLMs described above can be deployed even on personal devices such as smartphone, laptops and desktop working stations, or wearables. According to Google Cloud’s “2025 State of AI infrastructure” study, 73 per cent of respondents stated that deployment of Gen AI models at the “edge” (i.e. close to users, for example on devices) is “very important” or “extremely important”.
- (93) Hardware and chip manufacturers are developing personal devices (including laptops and desktop workstations) with meaningful AI compute capabilities which can be used to run AI models locally and efficiently. For example, Intel produces AI laptops with built-in GPU and Neural Processing Units (NPUs)<sup>136</sup> which can support various on-device AI assistants. Microsoft has included a Copilot key on some of its new Windows 11 laptops which are capable of local inferencing powered by Qualcomm or Intel GPUs. Additionally, Dell has released desktop workstations with NVIDIA GPUs.
- (94) This reduces the need for dedicated servers and virtual environments, creates new alternatives for sourcing compute, particularly for inference workloads. This shift may increase competition in the provision of compute with wider accessibility to less specialised compute infrastructure and lower latency.
- (95) According to Gartner’s research, in 2025, worldwide Gen AI spending is expected to be driven by spending on AI-enabled devices. Moreover, Gartner indicates “*The market’s growth trajectory is heavily influenced by the increasing prevalence of AI-enabled devices, which are expected to comprise almost the entire consumer device market by 2028*”.<sup>137</sup>

---

<sup>135</sup> Li, C (2025). The evolution of GPU pricing: a deep dive into cost per FP32 FLOP for hyperscalers, Available [here](#). (Accessed: 18 June 2025). Similar conclusions are found in Epoch AI (2024). Performance per dollar improves around 30% each year, Available [here](#). (Accessed: 11 June 2025). To measure the performance improvement per dollar of leading hardware over time, the nominal release prices of GPUs, including server hardware costs, are adjusted for inflation using the industry producer price index. These inflation-adjusted prices are then used to calculate the rate of improvement in performance per dollar, normalised to the release date of the NVIDIA H100 GPU (October 2022).

<sup>136</sup> Neural Processing Units (NPUs) are accelerators designed to efficiently execute AI and machine learning tasks, particularly neural network computations, with high speed and low power consumption in constraint environments such as edge computing.

<sup>137</sup> Business Wire (2025). Gartner Forecasts Worldwide GenAI Spending to Reach \$644 Billion in 2025, Available [here](#). (Accessed: 20 May 2025). Most of Gen AI spending in 2025 is expected to go toward hardware — including servers, smartphones, and PCs — as AI functionalities become increasingly embedded in physical devices, accounting for around 80% of total Gen AI-related expenditure.

## 4 CONCLUDING REMARKS

- (96) Compute is a critical input for the development and deployment of Gen AI. Against this backdrop, competition authorities, such as the European Commission and the UK's CMA, have expressed concerns that concentration in the capacity of compute, particularly by large integrated providers, could hinder Gen AI adoption. Moreover, the perceived insufficient availability of compute for Gen AI has led some authorities and policymakers to suggest that it might be appropriate to regulate the supply of cloud services, which is one important source of access to compute for Gen AI.<sup>138</sup>
- (97) As covered in this report, available public evidence suggests that providers of compute for Gen AI operate in an increasingly competitive environment. A wide range of players, including large established cloud providers, AI-specialised providers, and on-premises providers, appear to be actively investing and competing to offer compute for Gen AI through various solutions. Entry and growth by new players, high levels of innovation and indications of price decreases in compute for Gen AI suggest that competitive pressure is growing.
- (98) The nature of Gen AI demand is also shifting. As pre-training techniques become more efficient and the mix of Gen AI workloads increasingly shifts from pre-training to less computationally intensive fine-tuning and inference, it may be that a broader set of compute providers can effectively serve Gen AI developers and deployers. In addition, existing and upcoming regulatory frameworks and public initiatives, such as the EU Data Act and the investments plans in the AI Continental Action Plan, aim at reducing barriers to entry and expansion in cloud services, particularly around switching providers, and expanding overall available compute capacity. These measures are still being implemented and have yet to materialise in market outcomes but could affect competition for the provision of Gen AI-relevant compute.
- (99) Given these developments, immediate regulatory intervention in the compute market may be premature. Regulators should continue to monitor market outcomes – especially to ensure sufficient access for smaller Gen AI developers and deployers – while avoiding intervention without clear evidence of market failures, which could potentially discourage innovation or harm dynamic efficiency in the long term.

---

<sup>138</sup> European Commission recently launched a public consultation for an impact assessment on a Cloud and AI development Act which is considering, among the policy options, measures to “address the computational capacity deficit”. See: European Commission (2025). Call for evidence for an impact assessment. AI Continent – new cloud and AI development act. Page 2. Available [here](#). (Accessed: 20 May 2025).

## REFERENCES

Abendroth Dias, K., et al. (2025). Generative AI Outlook Report - Exploring the Intersection of Technology, Society and Policy, Publications Office of the European Union, Luxembourg. Available [here](#).

AI Business (2024). Musk xAI Ditches Oracle Cloud to Build Massive GPU Cluster for Grok 3, Available here. (Accessed: 19 June 2025)

AI Infrastructure Alliance, 2024. The State of AI infrastructure report 2024, Available here. (Accessed: 20 May 2025)

Alibaba Cloud (2023). Introduction to Alibaba Cloud GPU Service, Available here. (Accessed: 17 June 2025)

Alibaba Cloud (2024). An Introduction to Alibaba Cloud Platform for AI, Available here. (Accessed: 17 June 2025)

Alphabet (2025). Alphabet Announces Fourth Quarter and Fiscal Year 2024 Results, Available [here](#). (Accessed: 10 June 2025)

Alphabet Inc. (2025). 2024 Annual Report, Available here. (Accessed: 11 June 2025)

Amazon (2025). 2024 Annual Report, Available here. (Accessed: 11 June 2025)

Amazon SageMaker AI. Fine-tune a Model, Available here. (Accessed: 9 June 2025)

AMD (2024) AMD Delivers Leadership AI Performance with AMD Instinct MI325X Accelerators, Available here. (Accessed: 19 June 2025)

Anthropic (2023). Anthropic Partners with Google Cloud, Available here. (Accessed: 9 June 2025)

Anthropic (2023). Anthropic partners with Google Cloud, Available here. (Accessed: 18 June 2025)

Anthropic (2024). Anthropic and Amazon's Trainium partnership, Available here. (Accessed: 18 June 2025)

Anthropic (2024). Powering the next generation of AI development with AWS, Available here. (Accessed: 9 June 2025)

Apple (2025). Apple reveals M3 Ultra taking Apple silicon to a new extreme, Available here. (Accessed: 19 June 2025)

Artificial analysis (2025). LLM API Providers Leaderboard, Available here. (Accessed 17 April 2025)

Arxiv (2024). NVLM: Open Frontier-Class Multimodal LLMs, Available here. (Accessed: 19 June 2025)

AWS (2023). Amazon EC2 Inf2 Instances for Low-Cost, High-Performance Generative AI Inference are Now Generally Available, Available here. (Accessed: 19 June 2025)

AWS News Blog (2025). Announcing up to 45% price reduction for Amazon EC2 NVIDIA GPU-accelerated instances, Available here. (Accessed: 18 June 2025)

AWS. Generative AI, Available here. (Accessed: 17 June 2025)

AWS. Writer and AWS bring generative AI to largest ranked companies in the world, Available here. (Accessed: 19 June 2025)

BioRxiv (2024). Simulating 500 million years of evolution with a language model, Available here. (Accessed: 18 June 2025)

Blaize (2022). Edge AI and Vision Alliance awards Blaize, Inc., Best Edge AI Processor in annual Product of the Year ceremony, Available here. (Accessed: 19 June 2025)

Bruegel, 2025. How DeepSeek has changed artificial intelligence and what it means for Europe, Available here. (Accessed: 20 May 2025)

BuiltIn (2025). What is model distillation? Available here. (Accessed: 9 June 2025)

Business Insider, 2025. Microsoft is taking its foot off the AI accelerator. What does that mean? Available here. (Accessed: 20 May 2025)

Business Wire (2025). Gartner Forecasts Worldwide GenAI Spending to Reach \$644 Billion in 2025, Available here. (Accessed: 20 May 2025)

Business Wire (2025). Meta Collaborates with Cerebras to Drive Fast Inference for Developers in New Llama API, Available here. (Accessed: 9 June 2025)

Byrne, Corrado, Sichel (2021). The Rise of Cloud Computing: Minding Your Ps, Qs and Ks. National Institute Economic Review.

Cerebras (2024). Cerebras Systems Unveils World's Fastest AI Chip with Whopping 4 Trillion Transistors, Available here. (Accessed: 19 June 2025)

Cheah, E (2024). \$2 H100s: How the GPU Rental Bubble Burst, Available here. (Accessed: 10 June 2025)

ClearML (2024). The State of AI infrastructure at scale 2024, Available here. (Accessed: June 10 2025)

CMA (2023). AI Foundation Models: Initial Report, Available here. (Accessed: 20 May 2025)

- CMA (2024). [Draft] Markets Guidance, Available here. (Accessed: 20 May 2025)
- CMA (2024). AI Foundation Models, Available here. (Accessed: 20 May 2025)
- CNBC (2025). AI data center boom isn't going bust but the 'pause' is trending at big tech companies, Available [here](#) (Accessed: 16 October 2025)
- CNBC (2025). Amazon plans to spend \$100 billion this year to capture 'once in a lifetime opportunity' in AI, Available [here](#). (Accessed: 10 June 2025)
- CNBC (2025). The concern with CoreWeave's 250,000 NVIDIA chips ahead of its IPO, Available here. (Accessed: 10 June 2025)
- Cohere. Introducing Aya, Available here. (Accessed: 18 June 2025)
- CoreWeave (2024). CoreWeave's European Expansion, Available here. (Accessed: 16 May 2025)
- CoreWeave (2024). Mistral AI and CoreWeave Demonstrate Partnership at NVIDIA GTC, Mistral AI Hackathon, Available here. (Accessed: 9 June 2025)
- CoreWeave (2024). Mistral AI and CoreWeave Demonstrate Partnership at NVIDIA GTC, Mistral AI Hackathon, Available here. (Accessed: 19 June 2025)
- CoreWeave (2025). CoreWeave Announces Agreement with OpenAI to Deliver AI Infrastructure, Available here. (Accessed: 9 June 2025)
- CoreWeave (2025). CoreWeave Announces Agreement with OpenAI to Deliver AI Infrastructure, Available here. (Accessed: 19 June 2025)
- Coyle & Ngyuen (2019). Cloud computing, cross-border data flows and new challenges for measurement in economics. NBER Chapters, in: Measuring and Accounting for Innovation in the Twenty-First Century, pages 519-551, National Bureau of Economic Research.
- Crunchbase (2025). NVIDIA Continues Torrid AI Startup Investment Pace, Outstripping Microsoft And Google, Available here. (Accessed: 17 June 2025)
- Crusoe (2024). Crusoe Closes \$600M in Series D Round at \$2.8 Billion Valuation to Power AI, Available here. (Accessed: 17 June 2025)
- Databricks (2024). Introducing DBRX: A New State-of-the-Art Open LLM, Available here. (Accessed: 18 June 2025)
- DCD (2024). CoreWeave to invest £1bn in UK data centers, Available here. (Accessed: 16 May 2025)
- DCD (2025). Brookfield and Data4 to spend €20bn on AI infrastructure in France, Available here. (Accessed: 16 May 2025)

- DeepSeek (2024). Introducing DeepSeek-V3, Available [here](#). (Accessed: 17 June 2025)
- Dell Technologies Inc. (2025). Consolidated Statements of Income and Related Financial Highlights, Available here. (Accessed: 11 June 2025)
- Dell Technologies Inc. (2024). Dell Technologies Expands Dell AI Factory with NVIDIA to Turbocharge AI Adoption, Available [here](#). (Accessed: 02 July 2025)
- Dell'Oro Group (2025). Data Center capex to surpass 1 trillion by 2029, Available here. (Accessed: 20 May 2025)
- d-Matrix (2024). Microsoft-backed startup d-Matrix launches first AI chip, Available here. (Accessed: 19 June 2025)
- ECB (2025). Euro foreign exchange reference rates, Available here. (Accessed: 16 May 2025)
- Eetimes (2024). Recogni's 25-W Chip for AVs Processes High-Res Camera Streams, Available here. (Accessed: 19 June 2025) AI Infrastructure Alliance, 2024. The State of AI infrastructure report 2024, Available here. (Accessed: 20 May 2025)
- EnchargeAI. Technology, Available here. (Accessed: 19 June 2025)
- Engineering at Meta. (2024). Building Meta's GenAI infrastructure, Available here. (Accessed: 19 June 2025)
- Epoch AI (2024). Frontier language models have become much smaller, Available here. (Accessed: 9 June 2025)
- Epoch AI (2025). LLM inference prices have fallen rapidly but unequally across tasks, Available here. (Accessed: 9 June 2025)
- Epoch AI (2025). Performance per dollar improves around 30% each year, Available here. (Accessed: 20 May 2025)
- Epoch AI (2025). What went into training DeepSeek-R1, Available [here](#). (Accessed: 17 June 2025)
- Eseperanto.ai (2024). Esperanto Technologies and NEC Cooperate on Initiative to Advance Next Generation RISC-V Chips and Software Solutions for HPC, Available here. (Accessed: 19 June 2025)
- Etched (2024). Etched is Making the Biggest Bet in AI, Available here. (Accessed: 19 June 2025)
- European Commission (2024). Competition policy brief, p.7, Available here. (Accessed: 20 May 2025)
- European Commission (2025). AI Action Plan, Available here. (Accessed: 20 May 2025)
- European Commission (2025). Call for evidence for an impact assessment. AI Continent – new cloud and AI development act. Available here. (Accessed: 20 May 2025)

- European Commission (2025). AI Factories, Available here. (Accessed: 20 May 2025)
- European Commission (2025). EU launches InvestAI initiative to mobilise €200 billion of investment in artificial intelligence, Available here. (Accessed: 20 May 2025)
- European Parliamentary Research Service (2025). Briefing: AI factories, Available here. (Accessed: 9 June 2025)
- Eurostat (2025). Use of artificial intelligence in enterprises, Available here. (Accessed: 20 May 2025)
- Financial Times (2024). The \$200bn man: Larry Ellison's wealth rebounds as Oracle joins AI boom, Available here. (Accessed: 9 June 2025)
- Financial Times (2025). How 'inference' is driving competition to NVIDIA's AI chip dominance, Available here. (Accessed: 10 June 2025)
- Financial Times (2025). NVIDIA seeks to build its business beyond Big Tech, Available here. (Accessed: 20 May 2025)
- Flexera (2025). State of the Cloud Report, Available here. (Accessed: 28 April 2025)
- Forbes (2025). Here's how much AI firm DeepSeek and its founder are worth, Available [here](#). (Accessed: 17 June 2025)
- Forbes (2025). Meet The Tiny Startup Building Stargate, OpenAI's \$500 Billion Data Center Moonshot, Available here. (Accessed: 9 June 2025)
- Fortune (2025). Oracle bets big on U.K. AI boom with \$5 billion cloud investment, Available here. (Accessed: 16 May 2025)
- Gartner (2024). Forecast Analysis: Semiconductors and Electronics, worldwide, Available here. (Accessed: 20 May 2025).
- Goldman Sachs (2023). Generative AI could raise global GDP by 7%, Available here. (Accessed: 20 May 2025)
- Goodfellow, I; Bengio, Y; Courville, A (2016). Deep Learning, MIT press.
- Google (2025). Ironwood: The first Google TPU for the age of inference, Available here. (Accessed: 19 June 2025)
- Google Cloud (2022). How Cohere is accelerating language model training with Google Cloud TPUs, Available here. (Accessed: 9 June 2025)
- Google Cloud (2022). How Cohere is Accelerating LLM training with Google Cloud TPUs, Available here. (Accessed: 18 June 2025)
- Google Cloud (2023). Dynamic Workload Scheduler, Available here. (Accessed: 17 June 2025)

Google Cloud (2024). How Writer leverages infrastructure powered by Google Cloud and NVIDIA to scale up custom AI models. Youtube video, Available here. (Accessed: 19 June 2025)

Google Cloud (2025). State of AI infrastructure, Available here. (Accessed: 10 June 2025)

Google Cloud. AI and machine learning products, Available here. (Accessed: 17 June 2025)

Graphcore (2020). Introducing the Colossus™ MK2 GC200 IPU, Available here. (Accessed: 19 June 2025)

Groq. What is a Language Processing Unit?, Available here. (Accessed: 19 June 2025)

Groupe d'études géopolitiques (2025). Financing Infrastructure for a competitive European AI, Available here. (Accessed: 20 May 2025)

HAI (2025). The 2025 AI Index Report, Available here. (Accessed: 10 June 2025)

HPE (2024). Condensed Consolidated Statements of Earnings, Available here. (Accessed: 11 June 2025)

HPE (2024). Hewlett Packard Enterprise and NVIDIA announce 'NVIDIA AI Computing by HPE' to accelerate generative AI industrial revolution, Available [here](#). (Accessed: 02 July 2025)

IBM (2023). What is AI Inferencing? Available here. (Accessed: 9 June 2025)

IBM (2024). New Telum II Processor and IBM Spyre Accelerator, Available here. (Accessed: 19 June 2025)

IBM (2024). What is fine-tuning? Available here. (Accessed: 9 June 2025)

IBM (2025). IBM enhances the capabilities of watsonx.governance, Available here. (Accessed: 17 June 2025)

IBM. Explore our AI solutions, Available here. (Accessed: 17 June 2025)

IDC (2024). On-Premises AI Infrastructure Balances Innovation and Security, Available here. (Accessed: 10 June 2025)

Iliad Group (2025). The Iliad group is investing €3 billion in AI, Available here. (Accessed: 16 May 2025)

Intel (2025) Intel Gaudi 3 Expands Availability to Drive AI Innovation at Scale, Available here. (Accessed: 19 June 2025)

IOT Analytics (2024). Who is winning the cloud AI race? Microsoft vs. AWS vs. Google, Available here. (Accessed: 16 June 2025)

JLL (2025). 2025 Global Data Center Outlook, Available here. (Accessed: 10 June 2025)

- LG Corporation (2024). LG released new version of generative AI, EX-AONE 3.5, Available here. (Accessed: 18 June 2025)
- Li, C (2025). The Evolution of GPU Pricing: A Deep Dive into Cost per FP32 FLOP for Hyperscalers, Available here. (Accessed: 18 June 2025)
- Lightmatter (2025). A New Kind of Computer, Available here. (Accessed: 19 June 2025)
- LMarena (2025). Chatbot Arena, Available here. (Accessed: 9 June 2025)
- McKinsey (2024). AI power: Expanding data center capacity to meet growing demand, Available here. (Accessed: 20 May 2025)
- Medium (2025). The evolution of GPU pricing: a deep dive into cost per FP32 FLOP for hyperscalers, Available here. (Accessed: 18 June 2025)
- Meta Blog (2024). Introducing Meta Llama 3: The most capable openly available LLM to date, Available here. (Accessed: 19 June 2025)
- Microsoft (2024). Microsoft and Mistral AI announce new partnership to accelerate AI innovation and introduce Mistral Large first on Azure, Available here. (Accessed: 19 June 2025)
- Microsoft (2025). 2024 Annual report, Available here. (Accessed: 11 June 2025)
- Microsoft (2025). The Golden Opportunity for American AI, Available [here](#). (Accessed: 10 June 2025)
- Microsoft Azure (2024). Azure Maia for the era of AI: From silicon to software to systems, Available here. (Accessed: 19 June 2025)
- Microsoft Azure (2024). Microsoft and Mistral AI announce new partnership to accelerate AI innovation and introduce Mistral Large first on Azure, Available here. (Accessed: 9 June 2025)
- Microsoft Azure. Azure AI solutions, Available here. (Accessed: 17 June 2025)
- Microsoft Community Hub (2024). Differences between Pre-Training and Supervised Fine-Tuning (SFT), Available here. (Accessed: 9 June 2025)
- Microsoft Tech Community (2024) Elevating AI with Databricks on Azure: Introducing the Latest Large Language Models, Available here. (Accessed: 18 June 2025)
- Mistral (2025). Mistral Compute, Available here. (Accessed: 17 June 2025)
- Mythic. ME1076 M.2 A+E Key Card, Available here. (Accessed: 19 June 2025)
- Nebius (2024). Nebius to invest more than USD 1 billion to build AI infrastructure in Europe, Available here. (Accessed: 16 May 2025)

NextDC (2025). Neoclouds vs Hyperscalers: The Rise of AI-First Infrastructure (What It Means for You), Available here. (Accessed: 20 May 2025)

NVIDIA (2021). What Is Accelerated Computing? Available here. (Accessed: 9 June 2025)

NVIDIA (2024) NVLM: Open Frontier-Class Multimodal LLMs, Available here. (Accessed: 19 June 2025)

NVIDIA (2024). Nemotron-4 340B Technical Report, Available here. (Accessed: 19 June 2025)

NVIDIA (2025). France Bolsters National AI Strategy With NVIDIA Infrastructure, Available here. (Accessed: 17 June 2025)

NVIDIA (2025). NVIDIA Partner Network, Available here. (Accessed: 9 June 2025)

NVIDIA (2025). Thousands of NVIDIA Grace Blackwell GPUs Now Live at CoreWeave, Propelling Development for AI Pioneers, Available here. (Accessed: 19 June 2025)

NVIDIA. Generative AI runs on NVIDIA, Available here. (Accessed: 9 June 2025)

OECD (2023). Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research, OECD Publishing, Paris, Available here. (Accessed: 20 May 2025)

OpenAI (2023). OpenAI and Microsoft extend partnership, Available here. (Accessed: 9 June 2025)

OpenAI (2025). Announcing The Stargate Project, Available here. (Accessed: 16 May 2025)

OpenAI. Fine-tuning, Available here. (Accessed: 9 June 2025)

Oracle (2024). AI Innovators Flock to Oracle to Address Some of the World's Most Pressing Challenges, Available here. (Accessed: 19 June 2025)

Oracle (2024). Cohere and Oracle partnership brings generative AI solutions to enterprises, Available here. (Accessed: 9 June 2025)

Oracle (2024). Cohere and Oracle partnership brings generative AI solutions to enterprises, Available here. (Accessed: 18 June 2025)

Oracle (2024). Now Generally Available: The Largest, Fastest AI Supercomputer in the Cloud, Available here. (Accessed: 17 June 2025)

Oracle (2024). OpenAI Selects Oracle Cloud Infrastructure to Extend Microsoft Azure AI Platform, Available here. (Accessed: 9 June 2025)

Oracle (2024). Oracle and Reka Collaborate to Advance AI Innovation, Available here. (Accessed: 19 June 2025)

- Oracle (2024). Oracle Announces Fiscal 2025 Second Quarter Financial Results, Available here. (Accessed: 11 June 2025)
- Oracle. Artificial Intelligence, Available here. (Accessed: 17 June 2025)
- Pitchbook (2024). Emerging Space Brief: AI Neoclouds, Available here. (Accessed: 11 June 2025)
- Pitchbook (2025). NVIDIA-backed Lambda raises 480 million as AI neo-cloud funding surges, Available here. (Accessed: 20 May 2025)
- PR Newswire (2025). Supermicro Unveils Industry's Broadest Enterprise AI Solution Portfolio for NVIDIA Blackwell Architecture to Accelerate AI Factory Deployments in European Market, Available [here](#). (Accessed: 02 July 2025)
- Qualcomm (2023). Introducing Qualcomm Cloud AI 100 Ultra, Available here. (Accessed: 19 June 2025)
- Rebellions (2024). Korean AI Chipmaker Rebellions Unveils Mass-Production-Ready 'ATOM' at ISSCC 2024, setting new standards for AI acceleration, Available here. (Accessed: 19 June 2025)
- Reka AI (2024). Announcing the Latest Addition to Our Leading Multi-modal Models – Reka Core, Available here. (Accessed: 19 June 2025)
- Reuters (2023). Google Cloud partners with Mistral AI on generative language models, Available here. (Accessed: 9 June 2025)
- Reuters (2024). Startup Untether launches AI chips, Available here. (Accessed: 19 June 2025)
- Reuters (2025). AI chip firm Cerebras partners with France's Mistral, claims speed record, Available here. (Accessed: 9 June 2025)
- Reuters (2025). Exclusive: Huawei readies new AI chip for mass shipment as China seeks Nvidia alternatives, sources say, Available here. (Accessed: 19 June 2025)
- Reuters (2025). Exclusive: Meta begins testing its first in-house AI training chip, Available here. (Accessed: 19 June 2025)
- Reuters (2025). Exclusive: OpenAI taps Google in unprecedented cloud deal despite AI rivalry, sources say, Available here. (Accessed: 13 June 2025)
- Reuters (2025). Microsoft data center leases slowing, analysts say, raising investor attention, Available here. (Accessed: 20 May 2025)
- S&P Global (2025). Cloud Price Quarterly: New services, regions and prices to kick off the year, Available here. (Accessed: 9 June 2025)
- S&P Global (2025). Hyperscaler GenAI gets cheaper as pricing finds its level, Available here. (Accessed: 10 June 2025)
- S&P Global Market Intelligence (2025). 2025 Trends in Data, AI & Analytics, Available here. (Accessed: 20 May 2025)

Sambanova (2023). SambaNova Unveils New AI Chip, the SN40L, Powering its Full Stack AI Platform, Available here. (Accessed: 19 June 2025)

Samsung (2022). Samsung Introduces Game Changing Exynos 2200 Processor With Xclipse GPU Powered by AMD RDNA 2 Architecture, Available here. (Accessed: 19 June 2025)

Saudi Aramco (2025). How AI helps Aramco turn old data into new opportunity, Available here. (Accessed: 19 June 2025)

SemiAnalysis (2024). AI Neocloud Anatomy and Playbook, Available here. (Accessed: 20 May 2025)

SemiAnalysis (2025). The GPU Cloud ClusterMAX™ Rating System, Available here. (Accessed: 20 May 2025)

Sima.ai (2025). SiMa.ai Announces MLSoC Modalix System-on-Module to Accelerate Multi-modal and Gen AI Applications at the Edge, Available here. (Accessed: 19 June 2025)

Stanford CRFM (2024). Writer: Palmyra X, Available here. (Accessed: 19 June 2025)

Statista (2025). Amazon and Microsoft Stay Ahead in Global Cloud Market, Available here. (Accessed: 9 June 2025)

Supermicro (2025). 2024 Annual Report, Available here. (Accessed: 11 June 2025)

TechCrunch (2024). After raising \$1.3B, Inflection is eaten alive by its biggest investor, Microsoft, Available here. (Accessed: 18 June 2025)

TechSpot (2024). Apple opted to use Google TPUs over NVIDIA GPUs for its AI training, Available here. (Accessed: 18 June 2025)

Tencent Cloud. Cloud GPU service, Available here. (Accessed: 17 June 2025)

Tencent Cloud. Tencent Cloud TI Platform, Available here. (Accessed: 17 June 2025)

Tenstorrent (2025). Tenstorrent Launches Blackhole™ Developer Products at Tenstorrent Dev Day, Available here. (Accessed: 19 June 2025)

The Science Survey (2025). A Deep-Dive Into DeepSeek, Available [here](#). (Accessed: 17 June 2025)

The Verge (2024). OpenAI to use Oracle's chips for more AI compute, Available here. (Accessed: 19 June 2025)

Tom's Hardware (2022). Chinese Biren's New GPUs Have 77 Billion Transistors, 2 PFLOPS of AI Performance, Available here. (Accessed: 19 June 2025)

Tom's Hardware (2025). China's Cambricon posts first profit as demand for this Nvidia rival's AI processors explodes, Available here. (Accessed: 19 June 2025)

- Touvron et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. Available [here](#). (Accessed: 5 August 2025)
- UK Government (2025). Prime Minister sets out blueprint to turbocharge AI, Available here. (Accessed: 16 May 2025)
- Unite.AI (2024). How DeepSeek Cracked the Cost Barrier with \$5.6M, Available [here](#). (Accessed: 17 June 2025)
- Unite.AI (2024). Inflection 2.5 – The powerhouse LLM Rivaling GPT-4 and Gemini, Available here. (Accessed: 18 June 2025)
- Wang, C & Kantarcioglu, M (2025). A Review of DeepSeek Models' Key Innovative Techniques, Available [here](#). (Accessed: 17 June 2025)
- Yahoo Finance. CoreWeave, Inc, Available here. (Accessed: 17 June 2025)
- Yahoo Finance. Nebius Group, Available here. (Accessed: 17 June 2025)