

POLICY BRIEF

Built for Purpose?

Demand-Led Scenarios for Europe's AI Gigafactories

Julia Christina Hess (interface) and Dr. Felix Sieker (Bertelsmann Stiftung)

October 22, 2025

Table of Contents

1.	Executive Summary	4
2.	Glossary	6
3.	Introduction	8
4.	Training and Deploying Frontier AI Models	10
4.1.	Mapping Frontier AI Models	10
4.2.	Phases of Generative AI Training and Deployment	13
5.	User and Provider Landscape	15
5.1.	Users	16
5.2.	Providers	17
6.	Mapping Existing and Announced GPU Clusters	18
6.1.	Measuring the Size of AI/GPU Clusters	19
6.2.	Strategies to Meet Growing AI Compute Demand	21
7.	Takeaways for AIGFs – Recommendation Based on Two Scenarios	29
7.1.	Anchor Customers with high AI compute demand	30
7.2.	Multi-Client User Setup (Low-to-Moderate Demand)	30
8.	Annex 1	33
8.1.	Table US/EU AI Compute Mapping	33

In cooperation with

| BertelsmannStiftung

1. Executive Summary

The AI Continent Action Plan, published in April 2025, outlines the EU's ambition to become a 'leading AI continent,' primarily through a major expansion of compute infrastructure. Central to this effort is the creation of a network of 19 AI Factories across Europe, each equipped with up to 25,000 H100 GPU equivalents. These facilities are designed to provide small and medium-sized enterprises (SMEs), researchers, and startups with access to the computing power needed to develop and test AI systems. In addition, the EU plans to establish five GPU clusters that are four times larger than the AI factories. These large-scale facilities, known as AI Gigafactories (AIGFs), will each host at least 100,000 H100 GPU equivalents and are intended for training and deploying frontier AI models. To support this initiative, a dedicated €20 billion fund under InvestAI will cover approximately one-third of the capital expenditures for each site.

In this policy brief, we analyse the feasibility of the AIGF initiative by taking into account the private sector dynamics that can be observed around AI compute buildouts. We argue that by placing so much weight on expanding compute infrastructure, the European Commission's proposals seem to assume that inadequate compute is the primary reason Europe has yet to realise its AI potential. This supply-side focus overlooks the factor that will ultimately determine whether billions of euros deliver meaningful outcomes: demand.

Drawing on our mapping of existing datacentre buildouts and the landscape of compute providers and users, we suggest two plausible operating models for the AIGFs that factor in demand:

- (1) **Anchor customer model:** Anchor customer model: Secure one or a few anchor customers with very high compute demand, as seen in the United States and China.
- (2) **Multi-client model:** Serve a broader set of clients with low to moderate AI workloads.

In our analysis of the global user-provider landscape shaping the AI compute ecosystem, we find that leading AI labs – such as OpenAI, xAI and Google – constitute the only user group capable of generating the high AI workloads that the AIGFs aim to attract. At the moment, Europe currently hosts only one such lab, Mistral, which makes the conventional 'anchor customer' model – where a single leading AI lab ensures utilisation – highly improbable.

Instead, a multi-client model that aggregates demand from a diverse set of users, including companies, startups, and academia, is the better suited option for Europe. Individually, these users generate low-to-moderate AI workloads for commercial applications that existing AI factories cannot adequately supply. In this scenario, AIGFs would need to offer more than raw compute to remain competitive with private providers, such as neoclouds. By providing value-added services – including structured onboarding, curated software stacks, and ongoing support – AIGFs could foster dynamic AI ecosystems for SMEs, startups, and companies.

In conclusion, policymakers should focus on three priorities when reviewing upcoming AIGF proposals and their stated goals:

- **Demand quantification:** Include projected AI workloads and user commitments.
- **Realistic objectives:** Align buildouts with EU AI industry dynamics and strategy.
- **Clear value proposition:** Differentiate AIGFs from hyperscalers and neoclouds through services and ecosystem offerings.

These takeaways offer guidance for policymakers in Europe and its Member States to come up with feasible and effective solutions to strengthen Europe's AI ecosystem in the long-term.

2. Glossary

AI Factories	A European Commission initiative to upgrade EuroHPC supercomputers into AI factories by adding up to 25,000 H100 GPU equivalents. The goal is to make high-performance compute available to a wide range of users – startups, SMEs, researchers and the public sector.
AI Gigafactories (AIGFs)	A European Commission concept for very large facilities equipped with at least 100,000 H100 GPU equivalents. These sites are supposed to train and deploy very large AI models.
AI models	Mathematical and computational system trained to recognise patterns or make decisions based on input data.
FLOPs	The standard measure for compute quantifying the number of calculations a processor can perform, usually at a per-second rate.
Frontier AI models	Very large AI models with billions of parameters, often trained with massive training compute.
Graphics Processing Unit (GPU)	A specialised processor designed for handling parallel computations, originally developed for videogames and now crucial for AI training and inference.
GPU clusters	Many GPUs networked across servers (often with high-speed interconnections) to train very large models or run heavy inference at scale.

Hyperscalers

The largest cloud providers operating global, massive-scale infrastructures (e.g., AWS, Microsoft Azure, Google Cloud).

Neoclouds

A new class of cloud providers specialising in GPU capacity for AI. Many evolved from crypto-mining operators, which repurposed these facilities into GPU clusters. They offer flexible contracts and rapid deployment.

Power capacity

The electrical supply a data centre can deliver, typically measured in megawatts or gigawatts, thereby constraining how much compute you actually can run.

Token

A unit of text processed by an AI model – often a piece of a word, rather than a full word. Models read and generate text one token at a time.

3. Introduction

2025 marked a clear shift in the European Commission's approach to artificial intelligence (AI) – away from a primary focus on AI regulation and towards competitiveness and industrial capacity. As President Ursula von der Leyen said at the AI Action Summit in France, the aim is to secure Europe's 'specific place in the global race for AI'.¹ Her speech struck a newly assertive tone, underscoring a newfound self-confidence: 'Too often, I hear that Europe is late to the race – while the US and China have already gotten ahead. I disagree because the AI race is far from over. Truth is, we are only at the beginning. The frontier is constantly moving, and global leadership is still up for grabs'.²

Two months later, on 9 April 2025, the European Commission published the AI Continent Action Plan, setting out how the EU aims to become a 'global leader in Artificial Intelligence, a leading AI continent'.³ At its core is a major expansion of computing infrastructure, pursued through two tracks:

First, a network of 19 AI factories⁴ (each comprising up to 25,000 H100 GPU equivalents) across Europe will provide small and medium enterprises (SMEs), researchers and startups with access to AI compute to develop and test AI systems.⁵

Second, the European Commission proposed five AIGFs designed for training and deploying frontier AI models.⁶ Each AIGF would host at least 100,000 H100 GPU equivalents⁷ – roughly four times the capacity of an AI factory. A dedicated €20 billion fund under InvestAI – a €200 billion initiative announced by von der Leyen at the AI Action Summit – aims to cover about one-third of each site's capital expenditures.

1 European Commission (2025). Speech by President von der Leyen at the Artificial Intelligence Action Summit. https://ec.europa.eu/commission/presscorner/detail/en/speech_25_471.

2 European Commission (2025). Speech by President von der Leyen at the Artificial Intelligence Action Summit. https://ec.europa.eu/commission/presscorner/detail/en/speech_25_471.

3 European Commission (2025). The AI Continent Action Plan. <https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan>.

4 Jakob Steinschaden (2025). 6 neue AI Factories in Tschechien, Litauen, den Niederlanden, Rumänien, Spanien und Polen. <https://www.trendingtopics.eu/6-neue-ai-factories/>.

5 European Commission (2025). AI Factories. <https://digital-strategy.ec.europa.eu/en/policies/ai-factories>.

6 European Commission (2025). The AI Continent Action Plan. <https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan>.

7 Commission (2025). Call for expression of interest in AI gigafactories (AIGFs). <https://www.eurohpc-ju.europa.eu/document/AIGIGAFACTORIESCONSULTATION.pdf>.

On paper, all these initiatives appear to be a coherent response to Europe's past struggles in the global AI race. However, the proposed solution mainly focuses on the underlying assumption that the lack of compute capacity – or insufficient supply – is the main reason why Europe has not been able to fully realise its AI potential. This perspective misses the other side – demand, or more precisely, the interconnected network of users and providers that will determine whether the AIGFs actually are utilised and whether their establishment ultimately proves to be a meaningful investment with regard to the envisioned objective. In this policy brief, we focus only on AIGFs and argue that AIGFs' goal – to train and deploy frontier AI models – can be achieved only if sufficient demand is factored in from the beginning.

The politically coined AIGFs are basically what the industry calls 'GPU clusters'.⁸ Their business model generally is rooted in two distinct strategies: (1) securing an anchor customer, often a frontier AI lab, to ensure stable utilisation, and (2) offering on-demand capacity to a diverse set of clients with a variety of compute needs.

The current definition of and vision for AIGFs are aligned closely with the first model – focusing on providing large compute capacities for users with demand for high AI workloads to train and deploy frontier AI models. While this strategy works in the United States (US) – where most frontier AI labs are based and, as we demonstrate in the empirical part of this policy brief, drive AI data centre buildouts – Europe's demand is far less certain because there are too few frontier labs to meet demand. Thus, policymakers must be realistic about the feasibility of the envisioned goal – training and deploying frontier AI models. Given that Europe currently has only one leading AI lab, Mistral, the likelihood of pursuing Scenario One, i.e., building AIGFs to train and deploy frontier AI models by securing an anchor customer with exceptionally high demand, appears limited. We argue that Scenario Two – aggregating demand by engaging a diverse set of stakeholders – is far more feasible in the European context, taking into account the EU AI ecosystem's characteristics and the debate around expanding AI adoption across industry segments. Pursuing this path would require adjusting AIGFs' objectives to (1) articulate a clear value proposition that differentiates them from other compute providers, particularly neoclouds, and (2) evidence sufficient aggregated demand to avoid idle capacity.

8 There is no single agreed-upon definition of large-scale AI compute infrastructure: Terms such as 'AI data centre', 'AI supercomputer', 'AI cluster', 'mega cluster' and 'GPU cluster' are used interchangeably. In this paper, we used the latter, as it makes a direct connection with what lies at the heart of such a cluster: graphics processing units (GPUs). While traditional high-performance computing (HPC) mainly operates on central processing units (CPUs), which are designed for versatility and handle many different tasks sequentially, GPUs are built for parallel processing, allowing them to perform thousands of calculations at once. This makes GPUs particularly effective for handling the enormous computational demand for training and deploying AI models.

4. Training and Deploying Frontier AI Models

The European Commission's AI Continent Action Plan places particular emphasis on frontier AI models, stating that the proposed AIGFs are supposed to train and deploy frontier AI models with billions of parameters.⁹¹⁰ This focus reflects the commission's view that Europe currently is lagging behind in frontier AI development. To put these goals and observations into context, the following section briefly maps AI labs that have brought such models to market and describes Europe's position within this landscape.

4.1. Mapping Frontier AI Models

Only a small number of AI labs are developing today's frontier models, and measuring their performance is difficult. One practical approximation is to examine training compute. This does not imply causality, but in practice, larger models trained with more compute often achieve better performance. However, measuring AI models' performance this way is difficult due to a lack of publicly available data on training compute. One of the best estimations comes from Epoch AI, which estimates training compute for leading models and defines 'large models' here as those trained with more than 10^{25} floating-point operations (FLOP).¹¹ FLOPs are the standard measure for compute, quantifying the number of calculations a processor can perform, usually at a per-second rate. It is a key performance indicator for AI processors. The first model trained at this scale was GPT-4, released by OpenAI in March 2023. By October 2025, 36 AI models had crossed the 10^{25} -FLOP threshold (see Chart 1).¹²

9 European Commission (2025). The AI Continent Action Plan. <https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan>.

10 The European Commission's AI Continent Action Plan states that AI gigafactories' objective is to 'develop and train' frontier AI models. In the subsequent call for expressions of interest, the Commission added that its purpose also includes deploying very large AI models. Accordingly, we define AIGFs' goal as to train and deploy frontier AI models.

11 Epoch AI (2025). Over 30 AI models have been trained at the scale of GPT-4. <https://epoch.ai/data-insights/models-over-1e25-flop>.

12 EpochAI listed 33 models as of June 2025. We added three more models released since June 2025 - Grok-4, GPT-5, and GLM-4.6 - based on desk research. Training compute for Grok-4 and GPT-5 certainly exceeds 10^{25} FLOPs; for GLM-4.6, the evidence is suggestive but not conclusive.

Chart 1: Global Landscape of Frontier AI Models by Developer and Region**1 Global Landscape of Frontier AI Models by Developer and Region**

American: 28 models (78%) Asian: 6 models (17%) European: 2 models (6%) Total: 36 Models

OpenAI
9 models

GPT-4
5 Sora
o1
GPT-4,5
GPT-5

GPT-4 Turbo
GPT-40
o3
GPT-4.1

Google DeepMind
6 models

Gemini 1.0 Ultra
Gemini 2.0 Pro
Gemini 2.5 Pro
Gemini 1.5 Pro
Veo 2
Veo 3

Anthropic
5 models

Claude 3 Opus
Claude 3.7 Sonnet
Claude Sonnet 4

Claude 3.5 Sonnet
Claude Opus 4

xAI
3 models

Grok-2
Grok-3
Grok-4

Meta AI
2 models

Llama 3.1-405B
Llama 4 Behemoth


OTHER
11 models

Inflection-2
Inflection-2.5
Doubao-pro

GLM-4 (0116)
Nemotron-4 340B
Pangu Ultra

Mistral Large
Mistral Large 2
GLM-4.6

Aramco Metabrain AI
GLM-4-Plus

 This visualization is adapted from Epoch AI's original chart on frontier model development (2025) by Rahman et al.

Source: Epoch AI and own research

As Chart 1 indicates, leadership in frontier AI is highly concentrated: 25 of 36 models come from just five labs, with OpenAI alone responsible for nine. The United States (US; highlighted in light purple) dominates this landscape, as US-based labs trained 28 of the models. This is a reflection of deep capital pools,

great access to AI talent and strong academia–industry linkages. The talent concentration is particularly strong: According to the Global AI Talent Tracker,¹³ 57% of the world's top AI researchers worked in the US, compared with 16% in the United Kingdom (UK), France and Germany combined.

At first glance, China (highlighted in yellow), appears to be underrepresented, with only six models (GLM-4, GLM-4-Plus, GLM-4.6, Doubao-pro and Pangu Ultra), but this should not be read as a lack of technical capacity to develop models. Two factors explain the gap: First, many Chinese labs do not disclose compute details. Second – and more importantly – China faces severe compute constraints due to export controls that limit access to the most advanced chips. Despite these compute constraints, China has managed to produce very capable models by contributing significantly to the development of algorithmic innovations for more efficient training. This is made possible by the strong talent base China can draw on: In 2022,¹⁴ 47% of all top-tier AI researchers measured in the Global AI Talent Tracker originated from China, while 28% still worked in the country.

Europe (highlighted in dark blue) appears only via France's Mistral, which trained two models of that size. This underrepresentation has three main causes. First, Europe lacks sufficient AI talent and continues to lose both homegrown and internationally trained graduates, particularly to the US¹⁵. Second, data access remains a constraint: Legal frameworks in the US and China allow for large-scale training on web-crawled data at speeds and scales Europe has yet to match.^{16 17} Third – and in the view of the European Commission, the decisive factor – is compute. The background is that compute needs for training are rising sharply:

-
- 13 Macro Polo (2025). The Global AI Talent Tracker 2.0. <https://archivemacropolo.org/interactive/digital-projects/the-global-ai-talent-tracker/>. The AI Global Talent Tracker measures AI talent by sampling authors of accepted papers at the Neural Information Processing Systems (NeurIPS) conference – widely regarded as the top AI conference – along with all oral presentation authors. It then hand-codes each author's education and current affiliation using public sources.
- 14 MacroPolo (2025). The Global AI Talent Tracker 2.0. <https://archivemacropolo.org/interactive/digital-projects/the-global-ai-talent-tracker/>.
- 15 Siddhi Pal (2024). Where is Europe's AI workforce coming from? <https://www.interface-eu.org/where-is-europes-ai-workforce-coming-from-conclusion>. For further information on AI talent distribution: Siddhi Pal, Catherine Schneider & Ruggero Marino Lazzaroni (2025). Technical Tiers: A New Classification Framework for Global AI Workforce Analysis <https://www.interface-eu.org/technical-tiers-in-ai-talent>, and Siddhi Pal, Catherine Schneider & Laura Nurski (2025). Solving Europe's AI talent equation: Supply, demand and missing pieces. <https://www.interface-eu.org/solving-europes-ai-talent-equation-ai-talent-in-europe>.
- 16 Joshua Love et al. (2023) Entertainment and Media Guide to AI: Geopolitics of AI Text and data mining around the globe. <https://www.lexology.com/library/detail.aspx?g=bb8a1903-83b4-48df-9c10-c00484e30848>
- 17 Blake Brittain (2025). Anthropic wins key US ruling on AI training in authors' copyright lawsuit. https://www.reuters.com/legal/litigation/anthropic-wins-key-ruling-ai-authors-copyright-lawsuit-2025-06-24/?utm_source=chatgpt.com.

Since 2010, the FLOPs required for state-of-the-art AI models have grown by 4.4x per year,^{18 19} while training expenditures have increased by 2.4x annually, driven largely by US frontier labs. Against this backdrop, the commission conceived the AIGF initiative.

4.2. Phases of Generative AI Training and Deployment

The European Commission's stated aim with the AIGFs is to train and deploy frontier AI models. In this context, knowing that compute demand varies widely – depending on the model pipeline's stage – is important. Development and use of generative AI can be divided roughly into three phases: **pre-training, post-training and inference**.^{20 21} Not all stages are equally compute-intensive.

- **Pre-training:** During this phase, models learn general patterns from very large datasets using self-supervised learning. This happens usually only once per base model and is a very large investment because the model processes vast datasets over many training steps: Training can run for weeks on thousands of specialised chips (see '[Case Study ChatGPT training clusters](#)' [info box](#)).
- **Post-training:** The pre-trained model then is refined to improve performance for specific goals. Two main approaches are used that can range from moderate to heavy compute requirements:²²
 - **Fine-tuning:** training the model on domain-specific data (e.g., medical literature) to improve accuracy in that field. This is usually only moderately compute-intensive (smaller, targeted datasets normally are used).

18 Jaime Sevilla & Edu Roldán (2024). Training compute of frontier AI models grows by 4-5x per year. <https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year>.

19 Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej & David Owen (2025). How much does it cost to train frontier AI models? <https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models>.

20 Felix Sieker, Alek Tarkowski, Lea Gimpel & Cailean Osborne (2025). Public AI – White Paper. https://www.bertelsmann-stiftung.de/Public_AI_2025.pdf.

21 To give an example on how the phases of generative AI training and deployment look in practice: A public service portal might launch a 'citizen services copilot'. The base model – already pre-trained at scale – is licensed from a provider. The agency then fine-tunes it on local regulations and FAQs. Once the system goes live, computational resources are needed primarily to handle the thousands of daily queries from citizens interacting with the model.

22 For this policy brief, we define low-to-moderate compute demand as workloads that can be executed on a small cluster (single to lowdouble-digit modern GPUs) for hours to a few days. In contrast, heavy workloads typically need hundreds to thousands of GPUs over weeks and specialised infrastructure. See: KI Bundesverband (2025). KI-Infrastruktur für das Training großer Modelle in Deutschland. https://ki-verband.de/VER1_KI-Rechenzentren-1-1.pdf.

- **Reinforcement learning (RL):** adjusting the model's behaviour based on human or automated feedback, so that it better aligns with desired objectives and values. This can be very compute-heavy due to larger and more complex datasets being processed.²³
- **Inference:** Finally, the model is deployed in user-facing applications, such as chatbots or search engines. At this stage, additional computational resources – known as *inference compute* – are required each time the model generates an output. Inference costs are recurring and can become the dominant expense at scale. They grow mainly with the number of tokens processed per request: Longer prompts and outputs require more computations. This effect is particularly strong for 'reasoning models' that produce step-by-step explanations (often called chain-of-thought), which generate many intermediate tokens.²⁴

What does this mean for AIGFs? To train and deploy frontier models, the most compute-heavy parts are (1) pre-training, (2) post-training's reinforcement learning phases and (3) deployment (inference) of reasoning models. For pre-training, AIGFs must provide very large, well-connected GPU clusters – usually on one site. In contrast, reinforcement learning within post-training and inference of reasoning models likely can be equally compute-intensive, but can be distributed across several sites. However, this is only a snapshot of today's practice: The field is moving at an extraordinary pace, and the compute demand across these stages may look very different in the coming years.

23 In reinforcement learning, a model generates answers to given prompts, which often will be 'judged' by a separate model. This step is very compute-heavy, as today, for reinforcement learning, a substantial amount of synthetic data is created to generate prompts and a second model usually is trained and deployed to provide feedback (judge) on answers to prompts from the model being trained.

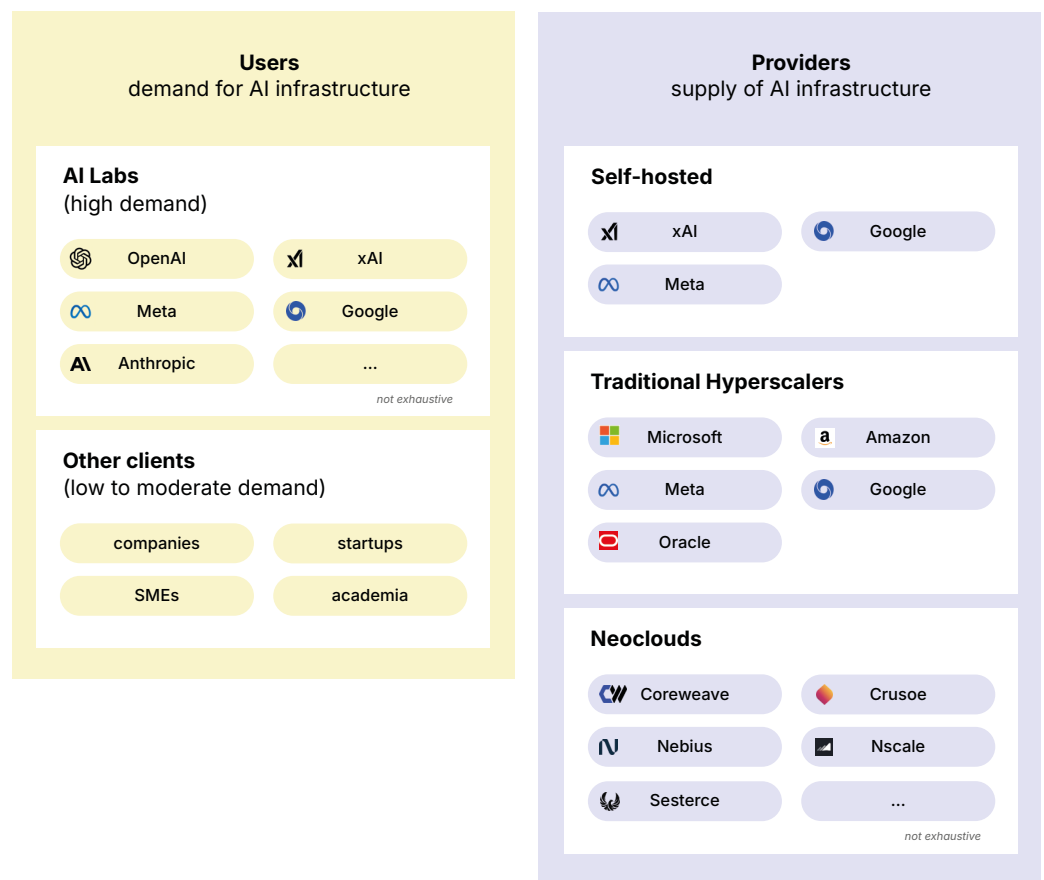
24 Felix Sieker, Alek Tarkowski, Lea Gimpel & Cailean Osborne (2025). Public AI – White Paper. https://www.bertelsmann-stiftung.de/Public_AI_2025.pdf.

5. User and Provider Landscape

In the AI Continent Action Plan, the discussion of frontier AI model development and AIGF deployment clearly establishes that the European Commission is beginning to assume an active role within a distinct network of users and providers shaping the global AI ecosystem centred around GPU clusters. The following section introduces these actors and their strategies, providing essential context for understanding how the AI industry typically designs, builds and operates AI compute infrastructure.

Chart 2: AI Compute Infrastructure Ecosystem

2 AI Compute Infrastructure Ecosystem



Over the past six years, the industry share of global AI compute has risen sharply, from 40% in 2019 to 80% in 2025.²⁵ According to Hawkins et al. (2025), 95% of this commercially available AI compute infrastructure is operated²⁶ by companies headquartered in the US or China. This shift from public to private sector ownership reflects changes in workloads that supercomputers are expected to handle. While government supercomputers are designed for a broad range of scientific tasks and to support foundational research, the demand today can be traced back to large-scale AI workloads.

5.1. Users

Users of AI compute infrastructure can be separated into two groups: first, a small number of **AI labs** that have developed large models, such as Open AI, xAI, Meta, Google, Anthropic, Mistral or DeepSeek.²⁷ They comprise a mix of newly established labs often backed by hyperscalers (e.g., Anthropic-Amazon, OpenAI-Microsoft) and labs operated by hyperscalers themselves (e.g., Google, Meta). They lead advanced AI model development and pursue multiple strategies to meet their high demand for AI compute for both training and inference.

The second **user group** entails **all other clients** with low-to-moderate demand and diverse AI workloads – such as companies, startups, SMEs and academia – in current discourse, often termed ‘industrial AI’. This group is broad and heterogeneous, spanning, for example, healthcare startups training diagnostic tools on sensitive patient data, SMEs in automotive developing predictive maintenance systems, financial firms experimenting with fraud detection or creative industries adopting generative AI for content production.^{28 29} However, their demand for compute is difficult to assess – varying widely by sector and evolving rapidly with technological progress – and AI’s disruptive potential has not been confirmed yet for many industries. This group represents the innovation ecosystem’s broad base, in which applied AI solutions are developed, commercialised and embedded across industries. These actors require affordable,

25 Konstantin F. Pilz, James Sanders, Robi Rahman & Lennart Heim (2025). Trends in AI Supercomputers. <https://arxiv.org/abs/2504.16026>.

26 Zoe Hawkins, Vili Lehdonvirta & Boxi Wu (2025). AI Compute Sovereignty: Infrastructure Control Across Territories, Cloud Providers and Accelerators. <https://dx.doi.org/10.2139/ssrn.5312977>.

27 See table in the [annex](#).

28 McKinsey & Company (2025): What is generative AI? <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai>.

29 European Commission (2025): AI in automotive: applications, opportunities and barriers. <https://digital-strategy.ec.europa.eu/en/events/ai-automotive-applications-opportunities-and-barriers>.

flexible and variable compute resources, currently through cloud-based services, AI factories and private investments in smaller-scale compute infrastructures.

5.2. Providers

Providers of AI compute infrastructure also can be separated into two groups: First, **hyperscalers** typically offer (general) cloud capacity via multi-service platforms for a variety of businesses and individual customers across multiple continents. They either build AI clusters themselves or lease them from third parties that design, own and operate the facilities. These arrangements allow hyperscalers to deploy their own IT equipment, and it is common to pre-lease space in unfinished facilities to confirm demand.³⁰ Their infrastructure does not focus merely on AI workloads, even though new buildouts increasingly are focusing on customised AI infrastructure. Some hyperscalers, such as Google and Meta, occupy a dual role in the ecosystem: as operators of AI infrastructure and as users through their own AI labs. This dual role gives them a clear advantage and flexibility towards other AI labs, enabling them to use GPU clusters for their own purposes or rent capacity to external clients.

However, **neoclouds** are a new class of cloud providers specialising in providing GPU-heavy compute capacity for AI. Many evolved from crypto mining operators and repurposed these facilities – already provisioned for high power and cooling – into GPU clusters. Unlike hyperscalers that typically pursue multi-year, region-wide buildouts, neoclouds focus on a small set of AI labs and startups with very large near-term needs using flexible contracts and rapid deployments. As evidenced by [the compute mapping table in annex 1](#) and their previous activities in crypto mining, neoclouds tend to scale capacity more quickly and ambitiously, often paired with comparatively low prices.³¹

30 Dylan Patel, Jeremie Eliahou Ontiveros & Maya Barkin (2025). Microsoft's Data Center Freeze – 1.5GW Self-Build Slowdown & Lease Cancellation Misconceptions. <https://semianalysis.com/2025/04/28/microsofts-datacenter-freeze/>.

31 Dylan Patel (2024). Inference Math, Simulation and AI Megaclusters - Stanford CS 229S - Autumn 2024. <https://www.youtube.com/watch?v=hobvps-H38o>.

6. Mapping Existing and Announced GPU Clusters

The next section will focus on mapping existing and announced GPU clusters – which display a similar infrastructure setup as described by the political term ‘AIGF’ – in the US and Europe. This analysis provides the foundation for identifying the different strategies that users may pursue to meet their demand for AI compute according to their specific needs. To set the stage, the following section introduces the key terms and metrics commonly used to describe AI compute.

There is no single agreed-upon definition of large-scale AI compute infrastructure: terms such as ‘AI data centre’, ‘AI supercomputer’,³² ‘AI cluster’, ‘mega cluster’ and ‘GPU cluster’ are used interchangeably. The latter makes a direct connection to what lies at the heart of such a cluster: graphics processing units (GPUs). While traditional high-performance computing (HPC) mainly operates on central processing units (CPUs), which are designed for versatility and handle many different tasks sequentially, GPUs are built for parallel processing, allowing them to perform thousands of calculations at once. This makes GPUs particularly effective for handling the enormous computational demand for training and deploying AI models.³³

Ambiguity of the term ‘AI Chips’

The term ‘AI chip’ is not clearly defined. It technically can mean everything from CPUs, GPUs, application-specific integrated circuits (ASICs) or field programmable gate arrays (FPGAs), to memory and power components, etc. But in the context of AI compute scaling, the term is associated mostly with GPUs. Measured by GPU shipments, Nvidia is the market leader, and the H100³⁴ accounted for more than 50% of total AI compute performance in 2024.³⁵ However, many other companies are designing AI chips for specific applications tailored towards their specific needs (ASICs). AMD is competing

32 For simplification and standardisation purposes, the term ‘AI supercomputer’ will be used throughout the paper to refer to AI computing infrastructure.

33 Felix Sieker, Alek Tarkowski, Lea Gimpel & Cailean Osborne (2025). Public AI – White Paper. https://www.bertelsmann-stiftung.de/Public_AI_2025.pdf.

34 Nvidia (2025). Nvidia H100 Tensor-Core-GPU. <https://www.nvidia.com/de-de/data-center/h100/>.

35 Konstantin F. Pilz, James Sanders, Robi Rahman & Lennart Heim (2025). Trends in AI Supercomputers. <https://arxiv.org/abs/2504.16026>.

with its Instinct series.³⁶ Google designs its own ASICs, known as tensor processing units (TPUs)³⁷, Meta relies on so-called 'Meta Training and Inference Accelerators' (MTIAs),³⁸ and Amazon on Trainium.³⁹ All of these systems also need high-bandwidth memory (HBM) to meet the intense data throughput demands of AI training and inference.⁴⁰

6.1. Measuring the Size of AI/GPU Clusters

An AI/GPU cluster's capacity typically is characterised using three metrics

- floating point operations per second (**FLOPS**)
- chip count (often in **H100 equivalents**)
- power capacity (in megawatts (**MW**) and gigawatts (**GW**))

FLOPs are the standard measure for compute quantifying the number of calculations a processor can perform, usually at a per-second rate. For example, Nvidia's H100 can deliver up to 4 quadrillion FLOPs (4,000 TeraFLOPs).⁴¹ However, in practice, comparing raw FLOPs across GPU generations is difficult, as many additional technical details must be factored in.⁴² Furthermore, it quickly aggregates to 'trillions' or 'quadrillions' of operations per second – a very abstract metric. Another metric for comparison and to normalise performance is translating GPU generations in **H100 equivalents**, factoring in their respective performance in FLOPs.^{43 44}

-
- 36 Advanced Micro Devices Inc. (2025). AMD Instinct™ MI300 Series Accelerators. <https://www.amd.com/en/products/accelerators/instinct/mi300.html>.
- 37 Google Cloud. AI-Entwicklung mit Google Cloud TPUs beschleunigen. <https://cloud.google.com/tpu?hl=de>.
- 38 Eran Tal, Nicolaas Viljoen, Joel Coburn & Roman Levenstein (2024). Our Next-generation Meta Training and Inference Accelerator. <https://ai.meta.com/blog/next-generation-meta-training-inference-accelerator-AI-MTIA/>.
- 39 Amazon Web Services (2025). AWS Trainium. <https://aws.amazon.com/de/ai/machine-learning/trainium/>.
- 40 Dylan Patel (2024). Inference Math, Simulation and AI Megaclusters - Stanford CS 229S - Autumn 2024. <https://www.youtube.com/watch?v=hobvps-H38o>.
- 41 Nvidia (2025). Nvidia H100 Tensor Core GPU. <https://www.nvidia.com/en-us/data-center/h100/>.
- 42 Philip Kiely (2025). Comparing GPUs across architectures and tiers. <https://www.baseten.co/blog/comparing-gpus-across-architectures-and-tiers>.
- 43 Konstantin F. Pilz, James Sanders, Robi Rahman & Lennart Heim (2025). Trends in AI Supercomputers. <https://arxiv.org/abs/2504.16026>
- 44 CharlesD (2024). Estimates of GPU or equivalent resources of large AI players for 2024/5. <https://www.lesswrong.com/estimates-of-gpu-or-equivalent-resources-of-large-ai-players>.

Case Study ChatGPT Training Clusters

OpenAI released ChatGPT-4 in March 2023. Estimates suggest that training the model required 20,000–24,000 Nvidia A100 GPUs over a period of approximately three months, excluding later fine-tuning stages.^{45 46} By comparison, SemiAnalysis calculated that a cluster of 100,000 Nvidia H100 GPUs – representing four to five times more hardware than the newer, more powerful generation – could complete the same training task in just four days.⁴⁷ The subsequent model, GPT-5, is reported to have been trained on 100,000 GPUs in two training runs.^{48 49 50} However, details on exact training times remain undisclosed.

Many AI supercomputer announcements also refer to power capacity (in MW and GW) as a reference point. Power capacity represents the maximum electrical load a site can support to run – and cool – chips at scale. It is introduced increasingly as the standard metric by setting the upper bound on the size of AI compute infrastructure that a facility can host, independent of the GPU type and generation used. On top of this, it allows for clear differentiation between what is available and when during a phased rollout, which is connected directly to questions about grid connection or permitting. Ideally, all three metrics are publicly available to paint the full picture.

EU Sovereignty and European AI Data Centres

The goal of achieving technical sovereignty remains hotly debated in EU policy circles. Despite claims to the contrary, no fully sovereign AI compute stack currently exists. While Europe has seen the rise of domestic cloud providers ('neoclouds'), supply of the hardware and software prevents sovereignty. This reality is already evident from the previous section, which

45 [r/singularity](https://www.reddit.com/r/singularity/jensen_huang_just_gave_us_some_numbers_for_the/) (2023). Jensen Huang just gave us some numbers for the training of GPT4 that are useful to predict GPT5. https://www.reddit.com/r/singularity/jensen_huang_just_gave_us_some_numbers_for_the/.

46 Dylan Patel (2024). Inference Math, Simulation and AI Megaclusters - Stanford CS 229S - Autumn 2024. <https://www.youtube.com/watch?v=hobvps-H38o>.

47 Dylan Patel (2024). Inference Math, Simulation and AI Megaclusters - Stanford CS 229S - Autumn 2024. <https://www.youtube.com/watch?v=hobvps-H38o>.

48 Dylan Patel (2024). Inference Math, Simulation and AI Megaclusters - Stanford CS 229S - Autumn 2024. <https://www.youtube.com/watch?v=hobvps-H38o>.

49 Dylan Patel & Daniel Nishball (2024). 100,000 H100 Clusters: Power, Network Topology, Ethernet vs. InfiniBand, Reliability, Failures, Checkpointing. <https://semianalysis.com/2024/06/17/100000-h100-clusters-power-network/>.

50 Matthew Griffin (2025). OpenAI GPT-5 is costing \$500 million per training run and still failing. <https://www.fanaticalfuturist.com/2025/05/openai-gpt-5-is-costing-500-million-per-training-run-and-still-failing/>.

outlined the companies active in developing the various types of chips needed for AI.

For example, GPUs are designed predominantly in the US. Europe currently does not play a role in this segment, and innovation in areas such as neuromorphic computing via startups, such as SpiNNcloud systems, remains in its infancy. Furthermore, Europe also does not manufacture them – wafer fabrication takes place at TSMC in Taiwan based on deposition and etching equipment from Japan (e.g., Tokyo Electron) and the US (e.g., Applied Materials), and lithography equipment from the Netherlands (e.g., ASML) to give just a few of many examples of chip production's interdependent nature. These examples touch on only a narrow section of a far larger supply chain spanning substrates, chemicals, firmware, interconnects and software. In summary, while Europe can enhance its control and resilience, a fully sovereign AI compute infrastructure remains out of reach for now, and key AI compute supply chain dependencies will endure in the long run.

6.2. Strategies to Meet Growing AI Compute Demand

The [large table that maps AI compute in the US and EU in the annex](#) presents a consolidated overview of all initiatives identified through in-depth desk research, complemented by the detailed dataset provided by EpochAI.⁵¹ Mapping AI compute capacity above 30,000 GPUs⁵² that already exists or is planned in the US and Europe highlights the massive scale of current expansion efforts. However, caution is warranted, as many of these initiatives remain at the announcement stage. The table presents publicly available information on each initiative's location, name, starting year, providers, users and capacity measured in H100 equivalents and power capacity. Public data are particularly scarce for the last two columns, as reflected by the numerous question marks.

51 Konstantin Pilz, Robi Rahman, James Sanders & Lennart Heim (2025). GPU Clusters. <https://epoch.ai/data/gpu-clusters>.

52 We chose to map GPU clusters with more than 30,000 GPUs to ensure that they do not overlap with the AI factories, which are defined as having up to 25,000 GPUs.

The following section breaks down the large table that can be found in [Annex 1](#) into subsets of initiatives grouped according to three user strategies for meeting AI compute needs: (1) AI labs building out their own infrastructure, (2) AI labs acting as anchor customers and (3) all user groups (AI labs, companies, etc.) renting cloud infrastructure.

Strategy 1: Own Infrastructure Built by AI labs

Table 1 depicts all publicly available information on GPU clusters that display AI labs building their own infrastructure.

Table 1: AI Labs Building Their Own GPU Clusters

A distinction is made between GPU clusters that are already operational or under construction (shown in purple) and those that have only been announced (not shown).

Country	Location	Name	Available Since/ From	Provider	User	H100 equivalents	Power Capacity	Links
US	Omaha, Nebraska	Papillion Campus	2021	Google	Google	?	?	Source
US	Omaha, Nebraska	Google Omaha AI Data Center	2024	Google	Google	?	150MW	Source
US	Austin, Texas	Tesla Cortex Phase 1	2024	Tesla	Tesla	50,000	130MW	Source
US	?	Meta AI Research Supercluster (RSC)	2024	Meta	Meta	2× 24,000	?	Source
US	Memphis, Tennessee	Colossus Phase 1 (xAI)	2024	xAI	xAI	100,000	150MW	Source
US	Lancaster, Ohio	Lancaster Data Center Campus	2024	Google	Google	?	?	Source
US	Council Bluffs, Iowa	Google Council Bluffs	2025	Google	Google	?	?	Source
US	Austin, Texas	Tesla Cortex Phase 2	2025	Tesla	Tesla	?	?	Source
US	Memphis, Tennessee	Colossus Phase 2 (xAI)	2025	xAI	xAI	230,000	200MW	Source

US	New Albany, Ohio	New Albany Cluster	2025	Google	Google	?	?	Source
US	Lincoln, Nebraska	Lincoln Data Center	2025	Google	Google	?	?	Source
US	Austin, Texas	Tesla Cortex Phase 3	2026	Tesla	Tesla	100,000	?	Source
US	Toledo, Ohio	Meta Prometheus	2026	Meta	Meta	?	1 GW	Source
US	Richland Parish, Louisiana	Louisiana Meta AI cluster Hyperion	2027	Meta	Meta	?	1.5GW	Source
US	Memphis, Tennessee	xAI Colossus Phase 3 (also called Colossus 2)	2029	xAI	xAI	1,000,000	1-1.5GW	Source

As evidenced in Table 1, AI labs such as Google, Meta and xAI⁵³ play pivotal roles in the AI ecosystem, each **building out their own massive compute infrastructures to meet growing training and deployment needs**. As previously mentioned, Google and Meta are both AI labs and hyperscalers.

Google has deployed several data centre clusters in Nebraska (Papillion Campus, Omaha AI Data Center) and in Ohio (New Albany Cluster, Lancaster Data Center Campus) within close proximity. These are serving Google's needs for both HPC and AI workloads. However, publicly available information does not provide many details on each site's capacity and hardware setups. Market analysts have suggested that all four clusters combined will comprise a GW-scale cluster by 2026.⁵⁴ Furthermore, Google is currently building a cluster in Columbus, Ohio.

Meta announced two new GPU clusters in the GW range in Louisiana (Prometheus by 2026 and Hyperion by 2027). Furthermore, xAI is expanding its Colossus GPU cluster in Memphis massively, with the ambitious goal to install a million H100 equivalents by 2029.⁵⁵

53 Tesla plays a special role; it is not an AI lab itself, but operates several compute-heavy business lines, including its Full-Self-Driving (FSD) System or its general-purpose robotic humanoid, Optimus.

54 Dylan Patel, Daniel Nishball & Jeremie Eliahou Ontiveros (2024). Multi-Data Center Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure. <https://semianalysis.com/multi-datacenter-training-openais/-google-s-ai-training-infrastructure>.

55 Artisan Baumeister (2025). Colossus 2 enthüllt: Warum XAI das größte Datencenter-Spektakel der Gegenwart wagt. <https://www.techzeitgeist.de/colossus-2-enthueilt-warum-xai-das-groesste-datencenter-spektakel-der-gegenwart-wagt>.

Strategy 2: AI Labs Acting as Anchor Customers

Table 2 below provides all publicly available information on GPU clusters in which AI labs act as anchor customers.

Table 2: AI Labs as Anchor Customers

A distinction is made between GPU clusters that are already operational or under construction (shown in purple) and those that have only been announced (not shown).

Country	Location	Name	Available Since/ From	Provider	User	H100 equivalents	Power Capacity	Links
US	Abilene, Texas	OpenAI Stargate Abilene Oracle OCI Supercluster Phase 1	2025	Oracle	OpenAI	?	200MW	Source
US	St. Joseph County, Indiana	Project Rainier (AWS)	2025	AWS	Anthropic	?	455MW	Source
US	Phoenix, Arizona	OpenAI/ Microsoft Goodyear	2025	Microsoft Azure	OpenAI	100,000	?	Source
Norway	Kvandal	Stargate Norway	2026	Nscale	OpenAI	100,000	230MW	Source
US	Muskogee, Oklahoma	CoreWeave Muskogee	2026	Undisclosed client	CoreWeave	?	100MW	Source
US	Mount Pleasant, Wisconsin	OpenAI/Microsoft Mt Pleasant, Wisconsin Phase 1	2026	Microsoft Azure	OpenAI	?	300MW	Source
US	Atlanta, Georgia	OpenAI/Microsoft Atlanta	2026	Microsoft Azure	OpenAI	?	324MW	Source
US	Abilene, Texas	OpenAI Stargate Abilene Oracle OCI Supercluster Phase 2	2026	Oracle	OpenAI	?	1.2GW	Source
France	Bruyères-le-Châtel, Essonne	Fluidstack France Gigawatt Campus	2026	Fluidstack	Mistral	500,000	?	Source

US	Denton, Texas	CoreWeave Cluster (OpenAI/Microsoft)	2027	CoreWeave	OpenAI	?	297MW	Source
France	Bruyères- le-Châtel, Essonne	Fluidstack France Gigawatt Campus	2028	Fluidstack	Mistral	?	1GW	Source

Table 2 indicates that AI labs often act as anchor customers, thereby confirming the use of the total or majority of new AI compute capacities due to their high compute demand for AI training and deployment. OpenAI's cooperation with hyperscalers is an interesting case study. ChatGPT's success in 2022 fuelled OpenAI's massive infrastructure buildout.⁵⁶ Initially, OpenAI partnered exclusively with Microsoft Azure, which built a dedicated AI training cluster called Goodyear in Arizona.⁵⁷ Another OpenAI/Microsoft cluster (Mount Pleasant) is planned for 2026, with a power capacity of 300MW.⁵⁸ However, OpenAI ended its exclusive partnership with Microsoft in January 2025 and started diversifying partnerships with other hyperscalers and neoclouds as a strategic customer. Since then, OpenAI has announced several new deals, such as the partnership with the neocloud provider Nscale in Norway (Stargate Norway),⁵⁹ the neocloud CoreWeave in Denton (CoreWeave cluster)⁶⁰ and the Oracle/Crusoe Abilene deal (Stargate Abilene Oracle OCI Supercluster Phase 1/Phase 2).⁶¹ Anthropic is pursuing a similar approach, aiming for a large GPU cluster with a power capacity of 455MW by the end of 2025, provided by the hyperscaler AWS.⁶² The European AI lab Mistral also announced a strategic partnership recently with neocloud provider Fluidstack, ultimately aiming for a GW-scale GPU cluster in France by 2028.⁶³

-
- 56 Dylan Patel, Jeremie Eliahou Ontiveros & Maya Barkin (2025). Microsoft's Data Center Freeze – 1.5GW Self-Build Slowdown & Lease Cancellation Misconceptions. <https://semianalysis.com/microsofts-datacenter-freeze/>.
- 57 Karen Hao (2024). AI Is Taking Water From the Desert. <https://www.theatlantic.com/ai-water-climate-microsoft/>.
- 58 Dylan Patel, Daniel Nishball & Jeremie Eliahou Ontiveros (2024). Multi-Data Center Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure. <https://semianalysis.com/multi-datacenter-training-openais/>.
- 59 Quickchannel (2025). Announcement: Stargate Norway. <https://qcnl.tv/p/ibfrFnPcQraFxbZSpXzVQ>.
- 60 CoreWave (2025). CoreWeave Announces Agreement With OpenAI to Deliver AI Infrastructure. <https://www.coreweave.com/news/coreweave-announces-agreement-with-openai-to-deliver-ai-infrastructure>.
- 61 Dylan Patel, Jeremie Eliahou Ontiveros & Maya Barkin (2025). Microsoft's Data Center Freeze – 1.5GW Self-Build Slowdown & Lease Cancellation Misconceptions. <https://semianalysis.com/2025/04/28/microsofts-datacenter-freeze/>.
- 62 Jeremie Eliahou Ontiveros, Dylan Patel, AJ Kourabi & Myron Xie (2025). Amazon's AI Resurgence: AWS & Anthropic's Multi-Gigawatt Trainium Expansion. <https://semianalysis.com/amazons-ai-resurgence-aws-anthropics-multi-gigawatt-trainium-expansion/>.
- 63 Fluidstack (2025). Fluidstack to Build 1 GW AI Supercomputer in France. <https://www.fluidstack.io/fluidstack-to-build-1-gw-ai-supercomputer-in-france>.

Strategy 3: On Demand

Table 3 below depicts publicly available information on GPU clusters that display on-demand strategies.

Table 3: On Demand

A distinction is made between GPU clusters that are already operational or under construction (shown in purple) and those that have only been announced (not shown).

Country	Location	Name	Available Since /From	Provider	User	H100 equivalents	Power Capacity	Links
US	Mountain View, California	Lambda Labs Cluster	2023	Lambda	multi-tenant	?	21MW	Source
US	Lawrence, Kansas	Lawrence Livermore NL EI Capitan Phase 2	2024	US Department of Energy	multi-tenant	?	35MW	Source
Finland	Mäntsälä	Nebius Finland	2025	Nebius	multi-tenant	60,000	75MW	Source
US	?	Oracle OCI Supercluster	2025	Oracle	multi-tenant	?	?	Source
US	?	Together AI cluster	2025	Hypertec	multi-tenant	90,955	?	Source
France	Paris	Nebius Paris cluster	2024	Nebius	multi-tenant	?	?	Source
US	?	Oracle OCI Supercluster B200s	2025	Oracle	multi-tenant	?	?	Source
US	?	Project Ceiba (AWS)	2025	AWS	multi-tenant	53,000	?	Source
US	Ellendale, Delaware	Applied Digital CoreWeave Ellendale Phase 1	2025	CoreWeave	multi-tenant	?	100MW	Source
US	Kansas City, Missouri	Nebius Kansas City Cluster	2025	Nebius	multi-tenant	35,000	40MV	Source
France	Valence Romans Agglo	Sesterce Valence	2026	Sesterce	multi-tenant	40,000	?	Source

UK	Loughton, Essex	Nscale Loughton	2026	Nscale	multi-tenant	45,000	90MW	Source
US	Ellendale, Delaware	Applied Digital CoreWeave Ellen dale Phase 2	2026	CoreWeave	multi-tenant	?	250MW	Source
US	Vineland, New Jersey	Nebius New Jersey	2026	Nebius	multi-tenant	?	300MW	Source
France	Southern France	Sesterce Southern France 250MW	2027	Sesterce	multi-tenant	200,000	250MW	Source
US	Ellendale, Delaware	Applied Digital Ellendale Phase 3	2027	CoreWeave	multi-tenant	?	400MW	Source
France	Grande Est	Sesterce Grand Est France A	2028	Sesterce	multi-tenant	250,000	300MW	Source
France	Grande Est	Sesterce Grand Est France B	2028	Sesterce	multi-tenant	250,000	300MW	Source
EU	Colocation	Hypertec Cluster	2029	Hypertec	multi-tenant	100,000	2GW	Source
France	Grande Est	Sesterce Grand Est France A	2030	Sesterce	multi-tenant	500,000	600MW	Source
France	Grande Est	Sesterce Grand Est France B	2030	Sesterce	multi-tenant	500,000	600MW	Source

Table 3 highlights the current popularity of the third strategy pursued by all user groups: to rent cloud infrastructure from hyperscalers – such as Microsoft, Amazon (AWS) or Oracle – or from neocloud providers such as CoreWeave, Nebius or Nscale. This works for all types of AI workloads (small to high, training or inference) and provides a high flexibility level. While this can be categorised as an additional strategy for AI labs, the second group of users described above – those with low-to-moderate demand – typically rents AI compute capacity in the cloud. Depending on workload profile, development phase (training stages or inference) and budget, they choose between a hyperscaler or neocloud. Neoclouds have been announcing massive new AI infrastructure deals in particular. Examples include

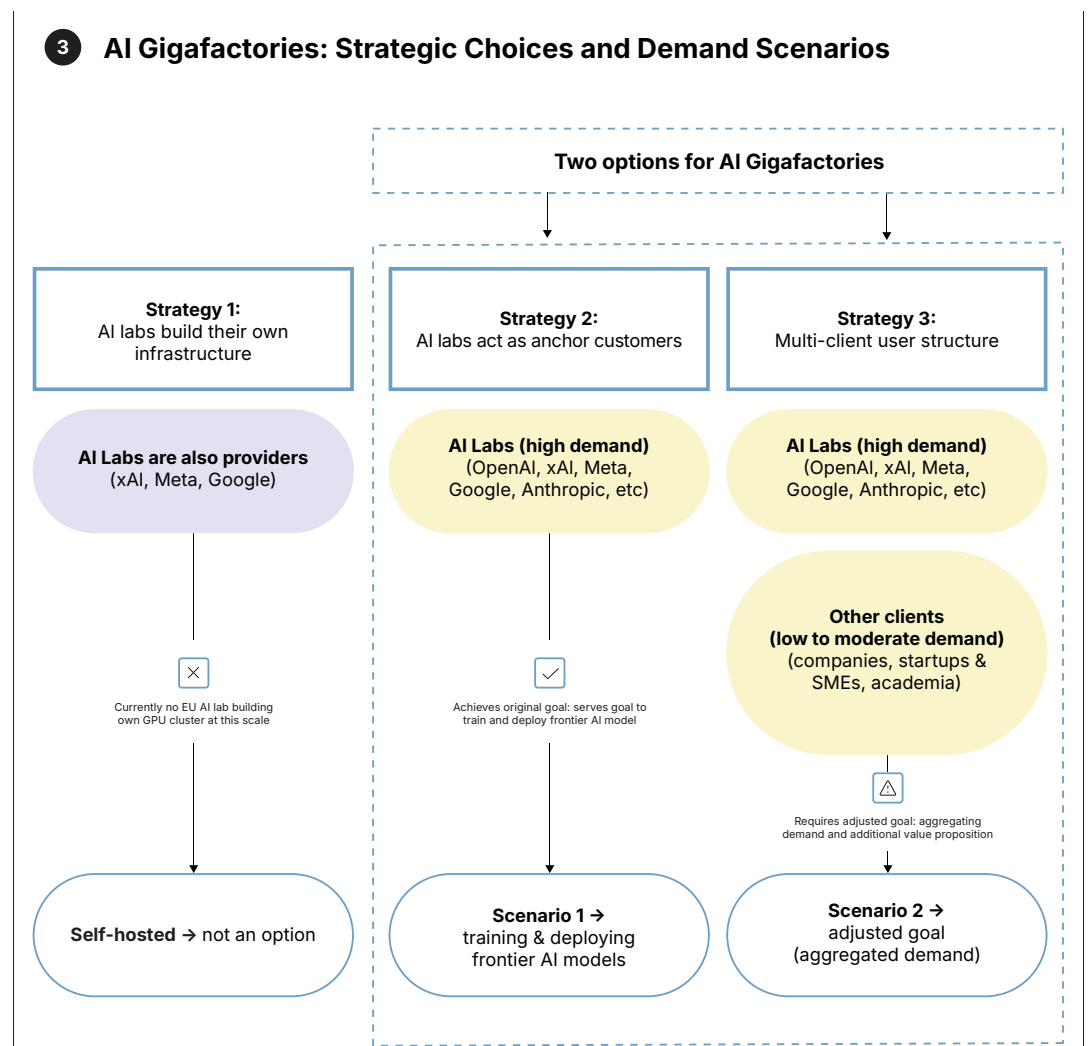
the deal between Applied Digital and CoreWeave in Ellendale,^{64 65} and the two large-scale AI infrastructure investments by Sesterce in France.⁶⁶

-
- 64 Applied Digital (2025). Applied Digital Announces 250MW AI Data Center Lease With CoreWeave in North Dakota. <https://ir.applieddigital.com/applied-digital-announces-250mw-ai-data-center-lease-with>.
- 65 Jeremie Eliahou Ontiveros, Dylan Patel & Daniel Nishball (2025). How Oracle Is Winning the AI Compute Market. <https://semianalysis.com/how-oracle-is-winning-the-ai-compute-market/>.
- 66 Georgia Butler (2025). Sesterce invests €450m in AI data center in Valence, France. <https://www.datacenterdynamics.com/sesterce-invests-450m-in-ai-data-center-in-valence-france/?ref=frenchtechjournal.com>.

7. Takeaways for AIGFs – Recommendation Based on Two Scenarios

The analysis indicates that different stakeholder groups – including AI labs, companies, startups, SMEs and researchers – pursue distinct approaches to meet their demand for AI compute. Furthermore, users *and* providers' strategies are linked closely. We are talking about billions in hardware investments for large-scale GPU clusters alone, so providers must avoid idle capacity at all costs. Thus, hyperscalers or neoclouds generally follow one of two strategies: securing an anchor customer, often a frontier AI lab to ensure stable utilisation, or offering on-demand capacity for a diverse set of clients with a variety of compute needs.

Chart 3: AI Gigafactories: Strategic Choices and Demand Scenarios



7.1. Anchor Customers with high AI compute demand

To achieve AIGFs' stated goal to train and deploy frontier AI models, securing an anchor customer or few strategic customers with high AI compute demand is key.

In the US, recent AI cluster buildouts have been driven either by labs operating their own infrastructure (e.g., xAI) or labs acting as anchor customers (e.g., OpenAI). Although we did not analyse Chinese labs' strategies, this pattern notably also can be observed in China: Some labs (e.g., DeepSeek) train on self-built clusters.⁶⁷ For example, Huawei's Pangu trains on inhouse designed chips (Ascend NPUs).⁶⁸ Other companies with AI labs – such as ByteDance, a major GPU customer of Oracle⁶⁹ – reportedly train models on that capacity. In Europe, Mistral – in teaming up with Fluidstack – already has applied this strategy.⁷⁰ Accordingly, when evaluating gigafactory bids, the European Commission should place explicit weight on the named customer, their demand projections' robustness and their commitments' reliability. However, because Europe currently has only one leading AI lab – Mistral – the prospects of securing European anchor customers with high demand are very limited in the near future.

7.2. Multi-Client User Setup (Low-to-Moderate Demand)

If securing a strategic customer proves unattainable, a different outcome could be to aggregate sufficient demand from clients with low-to-moderate AI workloads. **Importantly, in this multi-client scenario, AIGFs likely no longer would be built to train and deploy frontier AI models.** Instead, they would cater to a diverse set of users, effectively 'competing' with AI factories – the 13 smaller-scale facilities established all over Europe – in terms of client portfolios, and with neoclouds in terms of scale.

67 Dylan Patel, AJ Kourabi, Doug O'Laughlin & Reyk Knuhtsen (2025). DeepSeek Debates: Chinese Leadership On Cost, True Training Cost, Closed Model Margin Impacts. <https://semianalysis.com/2025/01/31/deepseek-debates/>.

68 ArXiv In-depth Analysis (2025). Pangu Ultra: Can Dense LLMs Compete at Scale? Huawei's 135B Parameter Bet on Ascend NPUs. <https://blog.gopenai.com/pangu-ultra-can-dense-llms-compete-at-scale-huaweis-135b-parameter-bet-on-ascend-npus>.

69 Jeremie Eliahou Ontiveros, Dylan Patel & Daniel Nishball (2025). How Oracle Is Winning the AI Compute Market. <https://semianalysis.com/how-oracle-is-winning-the-ai-compute-market/>.

70 Fluidstack (2025). Fluidstack to Build 1 GW AI Supercomputer in France. <https://www.fluidstack.io/fluidstack-to-build-1-gw-ai-supercomputer-in-france>.

Competing with neoclouds is risky, as they are often cheaper due to economies of scale and readily available access. As evidenced in our [table 3](#), neoclouds are betting on extraordinarily rising demand for AI compute, leading to very large data-centre expansion in GPU cluster plans. If AI adoption slows, and demand does not grow as projected, excess capacity could lead to price competition that might leave AIGFs struggling to compete. Moreover, many leading neoclouds – such as Nebius or Nscale – are European companies, thereby reducing AIGFs' sovereignty advantage that might otherwise encourage companies to switch to AIGFs.

Therefore, AIGFs should be positioned between AI factories and neoclouds, and compete on services and offerings, not simply on compute prices/GPU-hour prices: Serving a heterogeneous set of users will require more than raw compute – it demands a broader range of services very similar to the definition of 'AI factories' as 'dynamic ecosystems that foster innovation, collaboration and development in the field of AI'.⁷¹ In this scenario, AIGF should adopt elements from the AI factory model and act as a one-stop shop – a single storefront that bundles the ingredients needed to train and deploy all kinds of AI applications, such as structured onboarding, maturity diagnostics, access to compute resources, curated software stacks and ongoing support. Simultaneously, they could focus on larger, high-priority commercial needs that current AI factories either cannot meet or are too slow to provide.

Practical options include dedicating compute to Public AI models developed by European AI labs or pursuing sectorial approaches that aggregate enough stable demand – e.g., automotive in Germany or health and media in Spain. In this way, AIGFs could become more than infrastructure and foster an AI ecosystem that links SMEs, start-ups and academia.

In conclusion, we are proposing that policymakers should keep three things in mind when drafting criteria for AIGFs, as well as reviewing AIGF proposals:

(1) Demand quantification: Demand estimates must be part of each submission. Reliable public data are scarce, so it is even more important that submissions include demand estimates of potential AI workloads being served and evidence of commitments, e.g., in the form of letters of intent.

71 European Commission (2025). The AI Continent Action Plan. <https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan>.

(2) A realistic goal: Strengthening the European AI ecosystem effectively requires a clear strategy and feasible objective behind the buildout of AI compute, one that is connected to industry dynamics in play.

(3) A clear value proposition: AIGFs' success is tied strongly to a clear positioning to and differentiation from other types of providers – namely hyperscalers and neoclouds.

8. Annex 1

8.1. Table US/EU AI Compute Mapping

A distinction is made between GPU clusters that are already operational or under construction (shown in purple) and those that have only been announced (not shown).

Country	Location	Name	Available Since/ From	Provider	User	H100 equivalents	Power Capacity	Sources
US	Omaha, Nebraska	Papillion Campus	2021	Google	Google	?	?	Source
								Source
US	Omaha, Nebraska	Google Omaha AI Datacenter	2024	Google	Google	?	150MW	Source
US	Mountain View, California	Lambda Labs Cluster	2023	Lambda	multi-tenant	?	21MW	Source
US	Lawrence, Kansas	Lawrence Livermore NL EI Capitan Phase 2	2024	US Department of Energy	multi-tenant	?	35MW	Source
US	Austin, Texas	Tesla Cortex Phase 1	2024	Tesla	Tesla	50 000	130MW	Source
US	Undisclosed Location	Meta AI Research Supercluster (RSC)	2024	Meta	Meta	2× 24 000	?	Source
France	Paris	Nebius Paris cluster	2024	Nebius	multi-tenant	?	?	Source
US	Memphis, Tennessee	Colossus Phase 1 (xAI)	2024	xAI	xAI	100 000	150MW	Source
US	Council Bluffs, Iowa	Google Council Bluffs	2025	Google	Google	?	?	Source
US	Lancaster, Ohio	Lancaster Data Center Campus	2025	Google	Google	?	?	Source

US	Abilene, Texas	OpenAI Stargate Abilene Oracle OCI Supercluster Phase 1	2025	Oracle	OpenAI	?	200MW	Source
								Source
								Source
US	?	Oracle OCI Supercluster B200s	2025	Oracle	multi-tenant	?	?	Source
Finland	Mäntsälä	Nebius Finland	2025	Nebius	multi-tenant	60 000	75MW	Source
US	?	Oracle OCI Supercluster	2025	Oracle	multi-tenant	?	?	Source
US	?	together AI cluster	2025	Hypertec	multi-tenant	90 955	?	Source
US	Kansas City, Missouri	Nebius Kansas City Cluster	2025	Nebius	multi-tenant	35 000	40MW	Source
								Source
								Source
US	Austin, Texas	Tesla Cortex Phase 2	2025	Tesla	Tesla	?	?	Source
US	?	Project Ceiba (AWS)	2025	AWS	multi-tenant	53 000	?	Source
US	Phoenix, Arizona	OpenAI/Microsoft Goodyear	2025	Microsoft Azure	OpenAI	100 000	?	Source
US	Ellendale, Delaware	Applied Digital CoreWeave Ellendale Phase 1	2025	CoreWeave	multi-tenant	?	100MW	Source
US	St. Joseph County, Indiana	Project Rainier (AWS)	2025	AWS	Anthropic	?	455MW	Source
US	Memphis, Tennessee	Colossus Phase 2 (xAI)	2025	xAI	xAI	230 000	200MW	Source
US	New Albany, Ohio	New Albany Cluster	2025	Google	Google	?	?	Source
US	Lincoln, Nebraska	Lincoln Datacenter Campus	2025	Google	Google	?	?	Source

US	Austin, Texas	Tesla Cortex Phase 3	2026	Tesla	Tesla	100 000	?	Source
								Source
								Source
Norway	Kvandal	Stargate Norway	2026	Nscale	OpenAI	100 000	230MW	Source
								Source
US	Muskogee, Oklahoma	CoreWeave Muskogee	2026	CoreWeave	undisclosed client	?	100MW	Source
								Source
France	Valence Romans Agglo	Sesterce Valence	2026	Sesterce	multi-tenant	40 000	?	Source
UK	Loughton, Essex	Nscale Loughton	2026	Nscale	multi-tenant	45,000	90MW	Source
US	Ellendale, Delaware	Applied Digital CoreWeave Ellendale Phase 2	2026	CoreWeave	multi-tenant	?	250MW	Source
US	Mount Pleasant, Wisconsin	OpenAI/Microsoft Mt Pleasant, Wisconsin Phase 1	2026	Microsoft Azure	OpenAI	?	300MW	Source
								Source
US	Vineland, New Jersey	Nebius New Jersey	2026	Nebius	multi-tenant	?	300MW	Source
US	Atlanta, Georgia	OpenAI/Microsoft Atlanta	2026	Microsoft Azure	OpenAI	?	324MW	Source
US	Abilene, Texas	OpenAI Stargate Abilene Oracle OCI Supercluster Phase 2	2026	Oracle	OpenAI	?	1.2GW	Source
France	Bruyères-le-Châtel, Essonne	Fluidstack France Gigawatt Campus	2026	Fluidstack	Mistral	500 000	?	Source
								Source
US	Toledo, Ohio	Meta Prometheus	2026	Meta	Meta	?	1 GW	Source

US	Memphis, Tennessee	xAI Colossus Phase 3 (also called Colossus 2)	2026	xAI	xAI	1 000 000	1-1.5GW	Source
France	Southern France	Sesterce Southern France 250MW	2027	Sesterce	multi-tenant	200 000	250MW	Source
US	Denton, Texas	CoreWeave Cluster (OpenAI/Microsoft)	2027	CoreWeave	OpenAI	?	297MW	Source
US	Ellendale, Delaware	Applied Digital Ellendale Phase 3	2027	CoreWeave	multi-tenant	?	400MW	Source
US	Richland Parish, Louisiana	Louisiana Meta AI cluster Hyperion	2027	Meta	Meta	?	1.5GW	Source
								Source
France	Grande Est	Sesterce Grand Est France A	2028	Sesterce	multi-tenant	250 000	300MW	Source
France	Grande Est	Sesterce Grand Est France B	2028	Sesterce	multi-tenant	250 000	300MW	Source
France	Bruyères-le-Châtel, Essonne	Fluidstack France Gigawatt Campus	2028	Fluidstack	Mistral	?	1GW	Source
								Source
EU	Colocation	Hypertec Cluster	2029	Hypertec	multi-tenant	?	2GW	Source
								Source
								Source
France	Grande Est	Sesterce Grand Est France A	2030	Sesterce	multi-tenant	500 000	600MW	Source
France	Grande Est	Sesterce Grand Est France B	2030	Sesterce	multi-tenant	500 000	600MW	Source

Acknowledgements

We would like to thank Jan-Peter Kleinhans, Arno Amabile, Albert Cañigüeral Bagó, Nicolas Flores-Herr, Jenia Jitsev, Christian Temath, Dr. Ekaterina Prytkova, Simone Vannuccini, Sarah Budai, Nicole Lemke, Catherine Schneider and Maria Nowicka for their constructive feedback and support during the research and writing process; Maximilian Gottwald and Jack Walmsley for their help in research and text edits; Alina Siebert for designing the charts and the publication layout; Luisa Seeling for her support in editing the text and Sebastian Rieger and Iana Pervazova for helping us spread the word about this publication.

Authors

Julia Christina Hess

Senior Policy Researcher Global Chip Dynamics

jhess@interface-eu.org

+49 30 81 45 03 78 80

Felix Sieker

Project Manager at Bertelsmann Stiftung

felix.sieker@bertelsmann-stiftung.de

+49 30 275788-156

Imprint

interface – Tech analysis and policy ideas for Europe

W www.interface-eu.org

E info@interface-eu.org

T +49 (0) 30 81 45 03 78 80

F +49 (0) 30 81 45 03 78 97

interface – Tech analysis and policy ideas for Europe e.V.

c/o Publix

Hermannstraße 90

D-12051 Berlin

Layout & Infographics

Alina Siebert

Design by Make Studio

www.make.studio

Code

Convoy Interactive

This paper is published under Creative Commons License (CC BY-SA). This allows for copying, publishing, citing and translating the contents of the paper, as long as the Stiftung Neue Verantwortung is named and all resulting publications are also published under the license "CC BY-SA". Please refer to creativecommons.org/licenses/by-sa/4.0/ for further information on the license and its terms and conditions.