AI Sovereignty and National Power
Supplementary Materials

Micro Parameters: Accelerators, Servers, & AI Server Cabinets

Measures of Compute: FLOP, petaFLOPS, exaFLOPS, or zettaFLOPS and /s-days

Compute power is expressed both as a quantity and a performance, or in system dynamics terms as a stock and a flow. A value and a rate of change. The standard quantity measure of compute power is a 'floating point operations' (FLOP). A quadrillion (1.0E+15) FLOP are known a s a petaFLOP which historically was a more useful unit of measure for training and data needs. The standard unit of performance, or rate of computations, was then a petaFLOP per second, expressed in this paper as petaFLOPS. This is the ability of an accelerator to conduct one petaFLOP of compute for one second. As there are 86,400 seconds in a day a one petaFLOPS-day is equivalent to 86,400 petaFLOP. As compute rates have rapidly grown however, larger denominations are more useful, especially for modeling national levels of compute. For quantity measures there are 1,000 petaFLOP is an exaFLOP and a 1,000 exaFLOP is a zettaFLOP. These larger units have the same notations for performance measures (e.g. exaFLOPS, zettaFLOPS/day).

Generators of Compute: Accelerators

To generate these levels of compute are specialized accelerators which "allow the server to more quickly process large quantities of calculations in parallel" that may be based on GPU, ASIC, or TPU chips[1, p. 16]. We select the Nvidia DGX H100 8-GPU accelerator as a baseline for this work given its prevalence in literature and our datasets [1, p. 22], [2], [3]. As hardware changes rapidly, even as we write this paper H100 is approaching the end of life. Yet it still challenges legacy data center infrastructure in ways that illuminate achieving AI Sovereignty and National Power in Agentic AI may not be realized simply by repurposing decades old data center technology[4].

Compute for AI Training

In our conceptual analogy Training AI is likened to the design and prototyping of new aircraft frames and capabilities. At a high level AI training involves long-duration is "intensive, long-duration computation with relatively predictable power draws, typically performed in dedicated facilities with specialized hardware [1, p. 22]."

We measure Training AI as the required compute to train a model (FLOP), the size of the data set being trained on (FLOP), and the time it takes to train in hours or days to train. This ignores additional fine tuning, alignment, and other considerations but provides a useful rough order, especially as the difficulty of training each iterative

generation of model increase exponentially meaning the long-tail of other adjustments are captured within exponential growth of the three measures we identify.

Compute for AI Implementation

Once trained, an AI model can serve inferences or be 'implemented' at scale, be likened in our conceptual analogy to the use of the aircraft frames in daily operations to effect an instrument of national power. As serving inference is based on prompts by users and varies in size based on the complexity of tasks, implementation "presents more variable loads across a heterogeneous hardware mix, making it more challenging to model but increasingly important as deployment scales[1, p. 22]." Although implementation used to be cheap in terms of compute for GenAI, the evolution to Agentic AI with its longer context lengths and far more complex operations is approaching the dense high intensity requirements of Compute for AI Training[5, p. 23].

Physical Units of Compute: Tiles, Cabinets & Data Center Space

Tiles

In a data center the physical space available for computing power is allocated to 'tiles'. Each tile is 2'x2' of raised floor mounted above the actual floor to allow cabling and cooling to access the computer hardware above[6].

Cabinet

On these tiles are located 'cabinets which is the physical housing for servers. Convention dictates the measurement of a server cabinet is on the interior of the housing and relate to the size of available for placing standardized rack units. U stands for a space capable of housing a shelf of 1.75" of vertical space, so a 42U server rack contains 73.5" of usable computing space on the inside, and is nearly 7' on the outside.

Cabinet Capacity for AI Servers

A standard cabinet described above, as a best practice, can hold four AI Servers each mounted on individual racks configured with Nvidia DGX H100, each consisting of 8 GPUs providing a total of 32 GPU's[4]. As each Nvidia GGX H100 can provide 4 petaFLOPS of compute, each rack of 8 GPUs can generate 32 petaFLOPS and the cabinet can generate 128 petaFLOPS[4]. The rest of the space inside the cabinet, and the tile footprint is taken up by extensive cooling, network controls, and even white space around the cabinets to allow proper heat dispersion[4], [7].

Tile Usage per AI Cabinet

For all needs inside and outside a computer cabinet Nvidia recommends 32" x 48" of space[7], or the use of a 2x2 tile configuration.

Determining Data Center SqFt from Tiling & Occupied

Although exact percentages vary, heuristic rules of thumb are that 64% of available tiles are occupied by server cabinets with the remaining going vacant for future expansion, space, heat or suboptimal distance for clustering. All the tiles, occupied and unoccupied are known as the “white space” of a data center. White space is on average 65% of a data center’s total space, with the remaining taken up by “gray space” of networking, cabling, cooling and other infrastructure outside the rack components [8].

AI Cabinet Annual Electricity & Growth Unit: kW

AI Server electricity needs vary by their use in training or implementation and how those uses may require mode-shifting between rated, maximum, operational and idle [1, p. 17]. Although this is commonly called ‘power draw’, we use ‘electricity draw’ instead so as not to confuse the reader with our use of ‘power’ in the context of National Power and compute power. Annual electricity needs “The average power draw over an entire year, considering all operating modes and power levels[1, p. 17].” Current power per rack needs for AI Servers are between 30kw-100kw[5, p. 2] up to 250kw[9, p. 4] per rack. The logarithmic plot in Figure shows a behavior mode in the average data center rack load and peak performance rack power requirements historically from 2000-2024 and future forecasts through 2034[9, p. 5]. For AI Training Data Centers the power density is more likely to be 100 kW or higher per rack while AI Inference Data Centers use power densities on the lower range of 15-30 kW [5, p. 4]. With four AI server racks per cabinets, these estimates are all multiplied x4 to achieve the total AI Cabinet need in the model. In both the current and future forecast, the power draw of AI servers used in inference delivery and training are held constant at 40% and 80% operational up time respectively[1, p. 27],

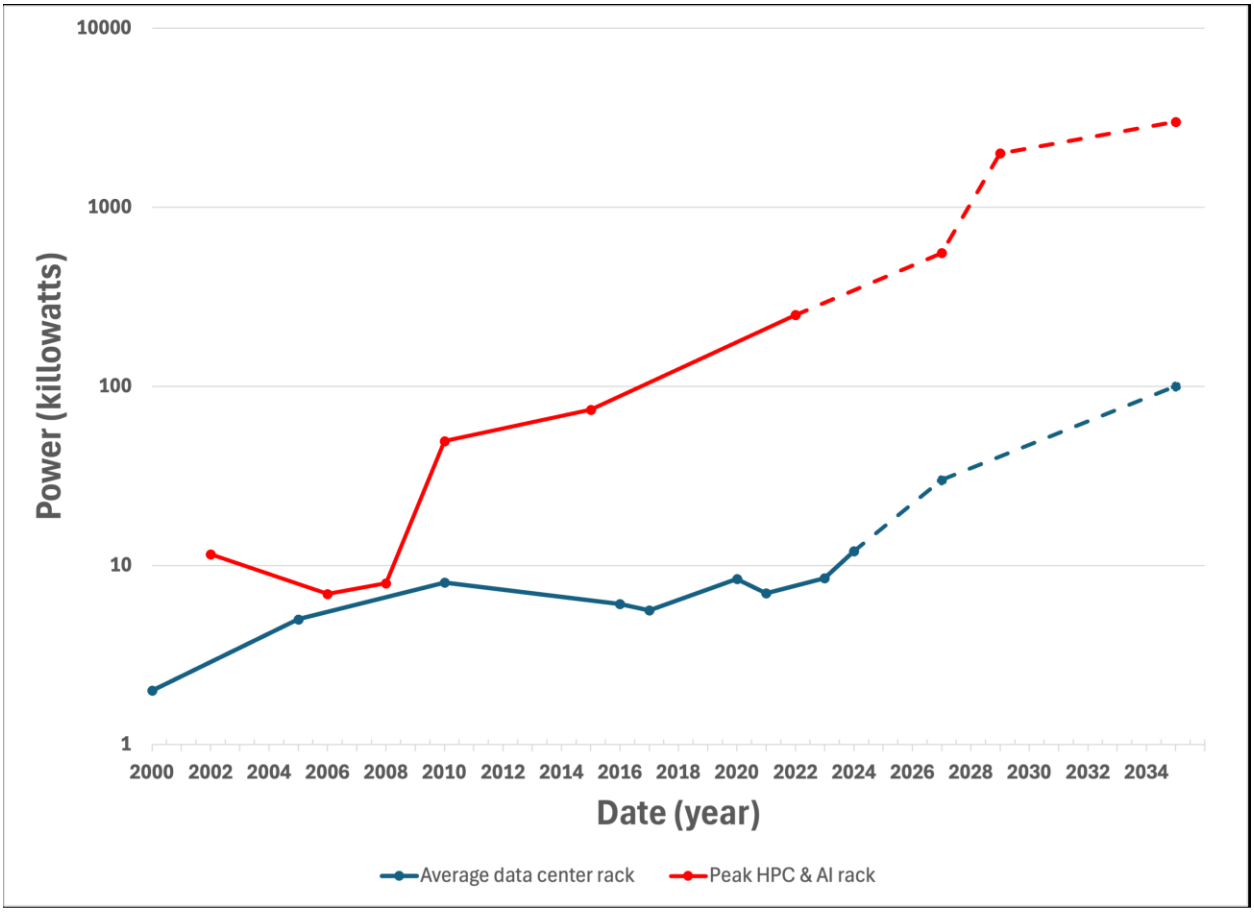


Figure bv: Historical and Forecasted Averaged and Peak Power Loads per Rack

AI Cabinet Water Needs

Water needs of AI Cabinets is directly related to the power density per rack. Although there are numerous non-water cooling strategies that can be employed at rack power densities below 100kw above that mark liquid cooling is a requirement[9, p. 2], [5, p. 10]. Direct to Chip (D2C) cooling can support 75-175 kw per rack while liquid immersion cooling can support 150-225 kW racks[5, p. 14].

AI Cabinet Accelerators

Using the Nvidia guidance of no more than four DGX H100 servers deployed in a cabinet there would be 32 accelerators per cabinet. [4]

.

Meso Parameters: Data Centers

Power Usage Efficiency (PUE) Unit: kWh/kWh

Measures data-center electricity efficiency by taking "total facility power split by IT hardware[10, p. 1846]." While a PUE of 1.5 means that 1.5MW would have to reach the data center for 1MW to be used by the computer components[5, p. 9]. Though PUE is not the only measure, it is a simplified way of measuring all loses of power to other sources ranging from administrative functions to transformer losses, which can account for up to 10% of all usage [9, p. 7]. The average annual PUE for data centers was 1.57 PUE in 2021[10]. The specific average PUE on AI Specialized data centers is 1.14 [1, p. 47]

Emissions Unit: kw

Emissions are calculated based on the kWh of electricity at the source with a factor of 0.35 kg/kWh of CO_2 equivalent[1, p. 57].

Water Usage Efficiency (WUE) Unit: liters/kWh

Similar to PUE, water usage efficiency (WUE) measures water usage at data centers. Direct or onsite WUE measures the water usage at the data center itself, used in cooling, while source WUE includes water usage in the generation of power[1, p. 39]. The average annual WUE is the annual site water usage used at the center divided by the total IT equipment energy with units of Liters per KWh (L/kWh)[11, p. 11]. Note this measure only accounts for water used in the cooling of servers, and does not include water that may be required for upstream generation of power, which is outside the boundary of our model. The average direct WUE on the four types of data centers we model are Small (.32), Midsize & Colo (.67), Hyperscale (.32), and AI Specialized (.61) [1, p. 47].

# of Data Centers by Type:

There is no standard terminology or taxonomy of data centers and different labels include small, midsize, colo, internet, and hyperscale with some existing hyperscale data centers drawing 5-100 MW of electricity [12, p. 31]. However, because of our choice to model the 1st Generation of National Power at the level of Agentic AI, which requires far more compute to train and implement than previous versions, we can focus on data centers designed for concentrating AI accelerators with corresponding electricity density and liquid cooling requirement. For our purposes, we track AI Specialized Data Centers which may have the MW capacity of a hyperscale, but is distinguished by its concentration of AI accelerators and corresponding high electricity density per rack requiring water cooling in addition to electricity[12, p. 36], a feature which requires design in the architecture of the data center itself and is difficult to retrofit.

We used a dataset of open source derived characteristics of modern US and Chinese AI 'frontier' data centers to provide rough estimates of AI specialized data centers. The

Epoch AI report lists nineteen projects, which when entries with blank values for H100 equivalent compute, power, and capital cost are removed leaves 11 AI specialized data center listed below. In Table 1 we report the Epoch AI findings of current H100 accelerator equivalents, the direct power needs and known total capital costs. Project handle is a combination of sponsor, location, and project name.

Table 1: Epoch AI OpenSource Findings on Frontier AI Data Center Projects

Handle	Current H100 equivalents	Current direct power (MW)	Current total capital cost (2025 USD billions)
Google Omaha Nebraska	109,880	189	6.05
Google Pryor Oklahoma	62,851	195	6.21
xAI Colossus 2 Memphis Tennessee	280,838	279	8.90
OpenAI-Oracle Stargate Abilene Texas	254,674	295	8.00
Amazon Canton Mississippi	214,348	341	10.89
Alibaba Zhangbei Zhangjiakou Hebei	168,128	474	15.12
xAI Colossus 1 Memphis Tennessee	275,796	498	12.88
Microsoft Fairwater Fayetteville Georgia	436,569	506	13.84
Google New Albany Ohio	235,426	543	17.34
Meta Prometheus New Albany Ohio	475,704	614	19.60
Anthropic-Amazon Project Rainier New Carlisle Indiana	471,566	751	23.97

Using this open source data and the researched parameters described above the authors derived additional data center values useful for our simulation. In Table 2 we use the H100 equivalents provided by Epoch AI to derive total zettaFLOPS of compute, the source power required using PUE and the direct water cooling needs of these centers.

Table 2: Author Estimated Values on Frontier AI Data Center Projects for Compute, Power, and Water

Project Handle	Current H100 equivalents	Est. Total Data Center Compute Generated at 4 petaFLOPS per H100	Est. Total Data Center Compute Generated in zettaFLOPS	Est. Source Power with PUE of 1.14 (MW)	Est. Water with WUE of .61 (Liters)
Google Omaha Nebraska	109,880	439,520	0.44	215.46	115,290
Google Pryor Oklahoma	62,851	251,404	0.25	222.3	118,950
xAI Colossus 2 Memphis Tennessee	280,838	1,123,352	1.12	318.06	170,190
OpenAI-Oracle Stargate Abilene Texas	254,674	1,018,696	1.02	336.3	179,950
Amazon Canton Mississippi	214,348	857,392	0.86	388.74	208,010
Alibaba Zhangbei Zhangjiakou Hebei	168,128	672,512	0.67	540.36	289,140
xAI Colossus 1 Memphis Tennessee	275,796	1,103,184	1.10	567.72	303,780

Microsoft Fairwater Fayetteville Georgia	436,569	1,746,276	1.75	576.84	308,660
Google New Albany Ohio	235,426	941,704	0.94	619.02	331,230
Meta Prometheus New Albany Ohio	475,704	1,902,816	1.90	699.96	374,540
Anthropic-Amazon Project Rainier New Carlisle Indiana	471,566	1,886,264	1.89	856.14	458,110

In Table 3 we use the H100 equivalents provided by Epoch AI to derive total AI Server Cabinets, required tiles and from that calculate the overall SqFt and km^2 footprint of the data centers.

Table 3: Author Estimated Values on Frontier AI Data Center Projects for Space

Project Handle	Current H100 equivalents	Est. Cabinets	Est. Tiles	Est Total SqFt of Data Center	Est Total km^2 of Data Center space
Google Omaha Nebraska	109,880	3,434	13,735	136,260	0.01
Google Pryor Oklahoma	62,851	1,964	7,856	77,940	0.01
xAI Colossus 2 Memphis Tennessee	280,838	8,776	35,105	348,261	0.03
OpenAI-Oracle Stargate Abilene Texas	254,674	7,959	31,834	315,816	0.03
Amazon Canton Mississippi	214,348	6,698	26,794	265,809	0.02
Alibaba Zhangbei Zhangjiakou Hebei	168,128	5,254	21,016	208,492	0.02
xAI Colossus 1 Memphis Tennessee	275,796	8,619	34,475	342,009	0.03
Microsoft Fairwater Fayetteville Georgia	436,569	13,643	54,571	541,380	0.05
Google New Albany Ohio	235,426	7,357	29,428	291,947	0.03
Meta Prometheus New Albany Ohio	475,704	14,866	59,463	589,911	0.05
Anthropic-Amazon Project Rainier New Carlisle Indiana	471,566	14,736	58,946	584,779	0.05

To gain SqFt of Data Space we estimated the number of AI Server Cabinets necessary to serve the H100 equivalents. Given the standard tile requirements and standard tile size from AI Server Cabinets we derived the white space these server cabinets would occupy. White space tiling occupied with servers usually only accounts of 64% of the data center white space, allowing us to calculate a total white space footprint, and then using a known ratio calculate the remaining gray space of the center resulting in the total data center SQFT.

Due to the enormous weight of servers and cooling most data centers have traditionally been a single story. However some data centers are experimenting with two, three and even more floors in height to optimize space and reduce latency across the clusters. The data center space estimates then may not represent the total footprint on the ground, but rather the interior space required.

Macro Parameters: National Power in Agentic AI and AI Sovereignty

We locate these as macro calculations at the level of National Power versus a meso (data center) or micro (AI Server Cabinet or even Racks) because training and implementation have different operational compute factors, and the allocation between them in our model is determined at the national level.

To validate our calculations are notionally useful for the problem we are addressing we reverse engineered the training hours of notable AI models comparing to reported results. To do this we filtered Epoch AI's dataset of Notable AI models to those published 2024 or later, that had listed training FLOP, training time in hours, accelerator type, and quantity of accelerators[3]. When comparing our calculated training in hours to the reported and scaling it our estimates ranged from .69 to 1.76 of reported with a mean of 1.05 and a median of .96. (See Supplementary for calculation sheet.)

Operational Compute

The operational compute power takes the operational uptime power draw from above sources, 40% for inference implementation and 80% for training, and multiplies that by the allocation for inference or training at the national level. So 1 petaFLOPs-day of inference compute capability would result in .4 petaFLOPS-day in operational compute.

Delivered Compute

Delivered compute is operational compute further constrained by the so-called 'compute-memory gap'. In addition to compute power generated by accelerators AI Servers require memory to store the information being processed. And the growth of compute power in FLOP is exceeding the growth rate in the speed of memory, resulting in the so-called 'compute-memory gap.' Though beyond the scope of this paper the difference between theoretical FLOP compute and delivered FLOP compute may range between 1%-30%. For this paper we assume a parameter of 10%. In the operational compute example 1 petaFLOPs-day of inference compute capability would result in .4 petaFLOPS-day in operational compute that results in 0.4 petaFLOPS-day in delivered compute.

AI Sovereignty

AI Sovereignty is the same five measures of National Power in Agentic AI divided by the amount of those values originating within the sovereign control of the nation. This normalizes all AI Sovereignty measures to 0-1 where a 1 means that 100% of a nations National Power in Agentic AI comes from sovereign sources.

% of Sovereign zettaFLOPS

zettaFLOPS of Agentic AI Implementation at each Generation

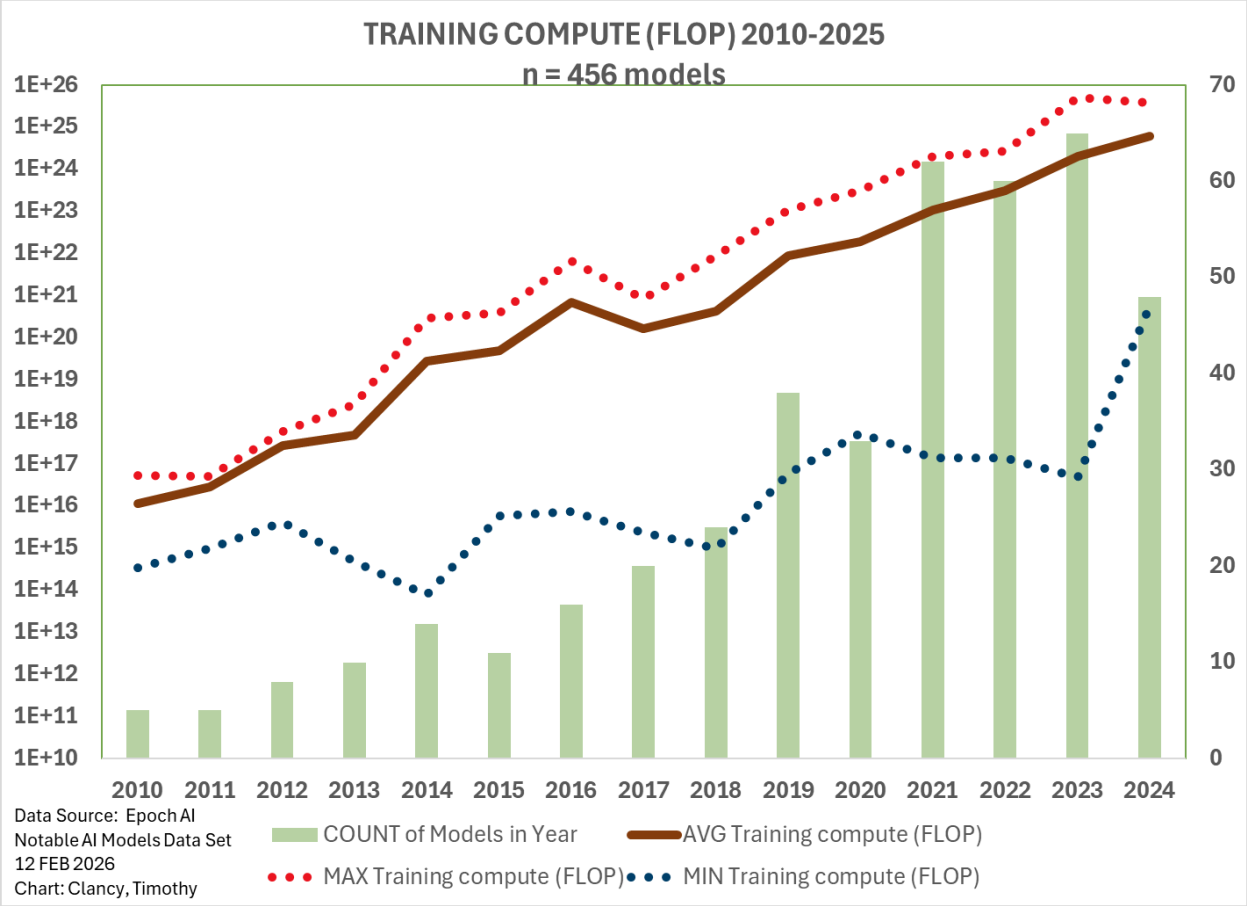
Training vs. Inference Implementation Split

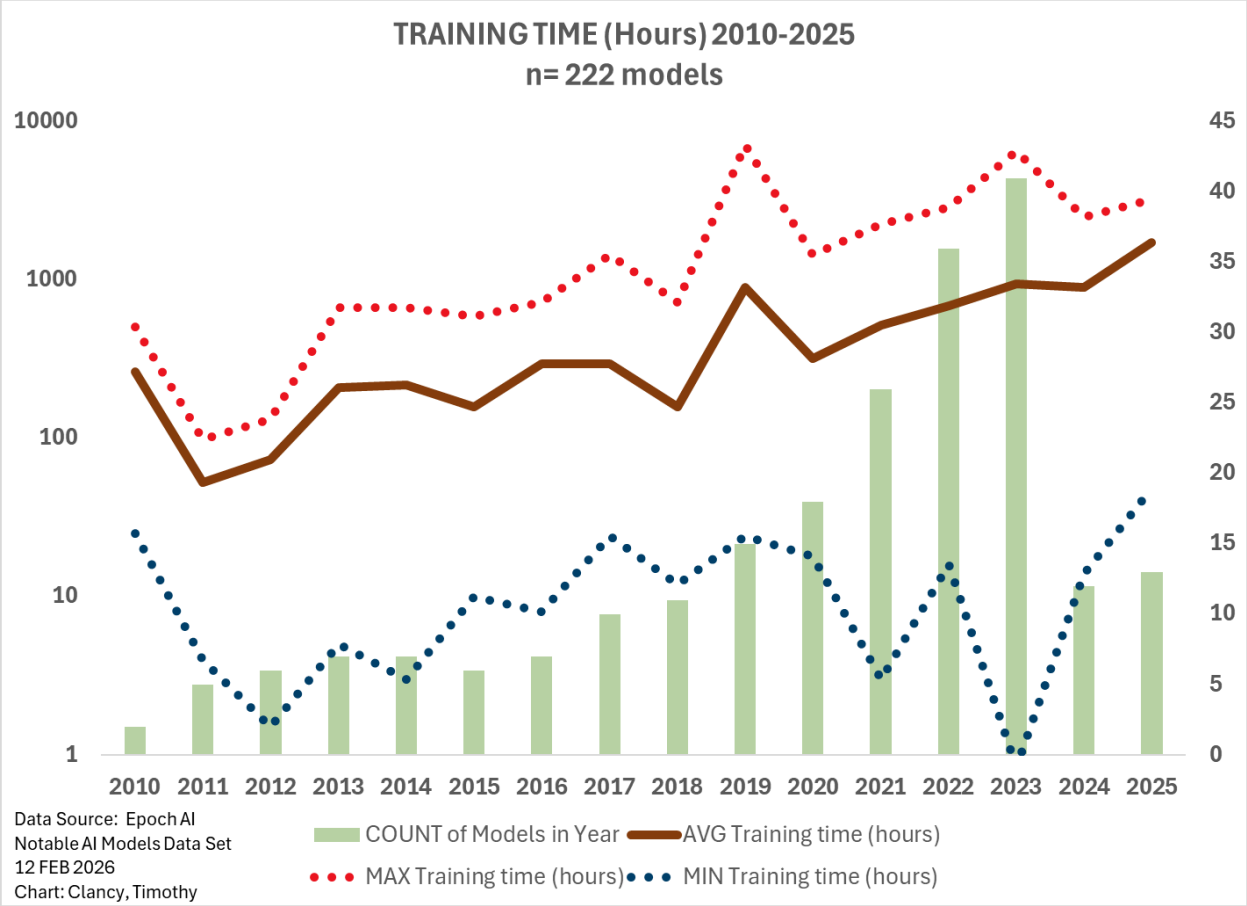
The allocation of national compute split between training and inference implementation. Table XX below indicates various historical measures of this allocation and future forecasts out to YYYY

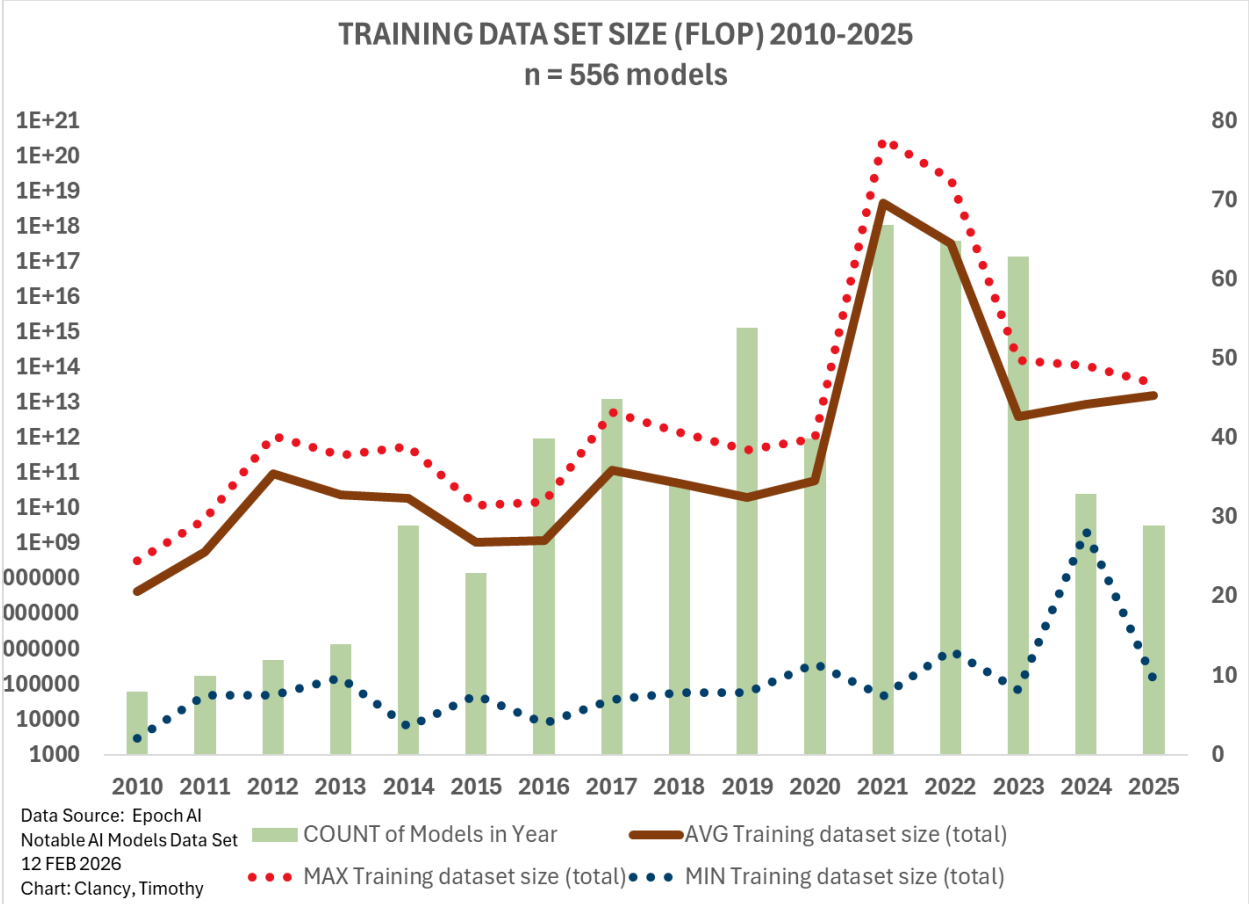
Although AI Training requires nearly 70% of AI Data Center capacity this is expected to shift over the next 5-10 years as frontier foundation Agentic AI models are trained, and the workload to shift to AI Inference[13],[14].

Pacing Measure: Training Frontier Agentic AI Foundation Models and at successive Generations of Capability

Given our purpose to forecast both National Power and Sovereignty in Agentic AI, being able to forecast a pacing measure is key.


TRAINING COMPUTE (FLOP) 2010-2025
n = 456 models
1E+26
1E+25
1E+24
1E+23
1E+22
1E+21
1E+20
1E+19
1E+18
1E+17
1E+16
1E+15
1E+14
1E+13
1E+12
1E+11
1E+10
70
60
50
40
30
20
10
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Data Source: Epoch AI
Notable AI Models Data Set
12 FEB 2026
Chart: Clancy, Timothy
COUNT of Models in Year
AVG Training compute (FLOP)
MAX Training compute (FLOP)
MIN Training compute (FLOP)


TRAINING TIME (Hours) 2010-2025
n= 222 models
10000
1000
100
10
1
45
40
35
30
25
20
15
10
5
0
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
Data Source: Epoch AI
Notable AI Models Data Set
12 FEB 2026
Chart: Clancy, Timothy
COUNT of Models in Year
AVG Training time (hours)
MAX Training time (hours)
MIN Training time (hours)



Calculating the Frontier Foundation Pacing Measure

Frontier Foundation Models at Each Generation:

Baseline for Pacing Measures Growth

We set the baseline in 2026 for our pacing measure using averages of Epoch AI's Notable AI Model dataset [3]. We want to understand how growth in training compute (FLOP), training dataset size (FLOP), and training time (hours) are changing as well as the number of models within each generation. Epoch AI's Frontier AI Model dataset tracks the top ten models by training compute at time of release. However, at best the sample size of each year's models is only ten and many fields of interest are marked with 'unspecified' or 'unreleased.' Epoch AI's Notable AI dataset meet a notability criteria and are the ones Epoch AI recommends for data analysis, and each year has hundreds of models so we use this dataset. This introduces a risk that our estimated parameters may be lower than actual values, but these can be adjusted upward easily as more data becomes available.

Table 4: Summary of Training Compute (FLOP) in Notable AI Models 2023-2025

Year	MIN	MEAN	MEDIAN	MAX	Models in Sample (n=)
2023	4.90E+16	2.07E+24	6.33E+22	5.00E+25	65
2024	5.85E+20	6.23E+24	3.03E+24	3.80E+25	48
2025	9.62E+21	4.04E+25	4.08E+24	5.00E+26	37

Table 5: Summary of Training Dataset Size (FLOP) in Notable AI Models 2023-2025

Year	MIN	MEAN	MEDIAN	MAX	Models in Sample (n=)
2023	6.90E+04	4.13E+12	7.68E+10	1.60E+14	63
2024	2.40E+09	9.53E+12	7.00E+12	1.14E+14	33
2025	1.44E+05	1.67E+13	1.48E+13	3.60E+13	29

Table 6: Summary of Training Time (hours) in Notable AI Models 2023-2025

Year	MIN	MEAN	MEDIAN	MAX	Models in Sample (n=)
2023	1	944	347	6,528	41
2024	14	897	612	2,500	12
2025	48	1,715	1,572	3,240	13

The findings of Table 4 - Table 6 show the wide range in values for any individual model, and how changes in technology, (or increasing secrecy and reluctance to publish) results in widely varying maximum and minimum values. However the Median values for training compute (FLOP), training dataset size (FLOP), and training time (hours) steadily increases. We therefore use the following median 2025 values as the baseline starting measure for our pacing growth.

Year	Training Compute (FLOP)	Training Dataset Size (FLOP)	Training Time (hours)
2025 Median Values	4.08E+24	1.48E+13	1,572

Pacing Measure Growth

Using our previous analysis of 2023-2025 on notable AI models and comparing median results then calculating growth factor

Table 7: Calculating Growth Factors from Median Results 2023-2025

Year	Median Training Compute (FLOP)	Year of Year Growth Factor	Median Data Set (FLOP)	Year of Year Growth Factor	Median Training Time (hours)	Year of Year Growth Factor
2023	6.33E+22		7.68E+10		347	
2024	3.03E+24	47.83	7.00E+12	91.15	612	1.76
2025	4.08E+24	1.35	1.48E+13	2.11	1,572	2.57

As Table 7 indicates the ushering in of LLM and their rapid evolution from 2023-2024 resulted in extremely high growth factors of x48 to x91 between Training Compute and Data Set Size. We therefore use the 2024-2025 growth factors in our comparison of year over year growth below.

Table 8: Year over Year Growth Factors Estimated

Source	Growth Factor of Training Compute (FLOP)	Growth Factor of Data Set Size (FLOP)	Growth Factor in Training Time (hours)
Notable AI 2023-2025 Analysis	x1.35/year	x2.11	x2.57
Hardware Acquisition Cost Analysis**[15, p. 5]**	x2.4-3.0 with a doubling time of 9 months		

From the same data set we also charted the values of Training Compute (FLOP), Dataset Size (FLOP), and Training Time (hours) 2010-2025. We arbitrarily selected 2010 as the starting point as the first year that had values in all three parameters. For each category we calculated max, min, average and count of models in the sample size, since not all records in the Notable AI Dataset always have the same data[3].

Frontier Foundation Models Needed to Innovate a New Generation of Capability

In our conceptual analogy, the evolution of one generation of military jet aircraft to the next is often best understood in retrospect or extensive analysis on well known capabilities. Given that we are just on the threshold of Agentic AI, it is difficult if not impossible to estimate what is required to ‘innovate’ a subsequent generation of Agentic AI capability. A good proxy might be the number of frontier foundation models, as these models often represent the cutting edge of capabilities. But there are few well researched estimates of how many frontier models must be innovated to achieve a next generation capability. The authors arbitrarily place this at five, and may revise in future editions as data.

Growth in Training Compute (FLOP)

Using hardware acquisition increases to approximate training growth requirements results in a range of 2.4-3.0x growth in training requirements since 2016 and a doubling time of 8-9 months[15, p. 5].

Data centers in the 1-5 GW range can support training runs from 1e28 to 3e29 FLOP while distributed data centers with aggregate 2 – 45 GW can support 2e28 to 2e30 FLOP, with a high end of 3e29 to 2e31 FLOP[16].

Technical Workforce

The technical workforce for Agentic AI are personnel who support both major operations: Training of models and Implementation through the serving of inference. The former workforce is much smaller than the latter. Designing and training Frontier models is a skillset possessed by so few that those highly skilled at it have been highly sought after gaining compensation packages of hundreds of millions of dollars to billions of dollars. But they are the tip of a very large pyramid of skilled workers necessary to serve the models, maintain the data centers, and integrate them into as an instrument of national power across industrial, business, and national security uses. Measuring this technical workforce is challenging in AI. Many programmers forgo completing formal education at the undergraduate, graduate, or PhD level instead opting to directly enter the workforce. Proxy measures are only roughly useful, such as the percentage of AI jobs among all new job openings, scholarly publications by country, or the attendance at AI conferences as reported in Figure 1, Figure 2, and Figure 3 respectively.

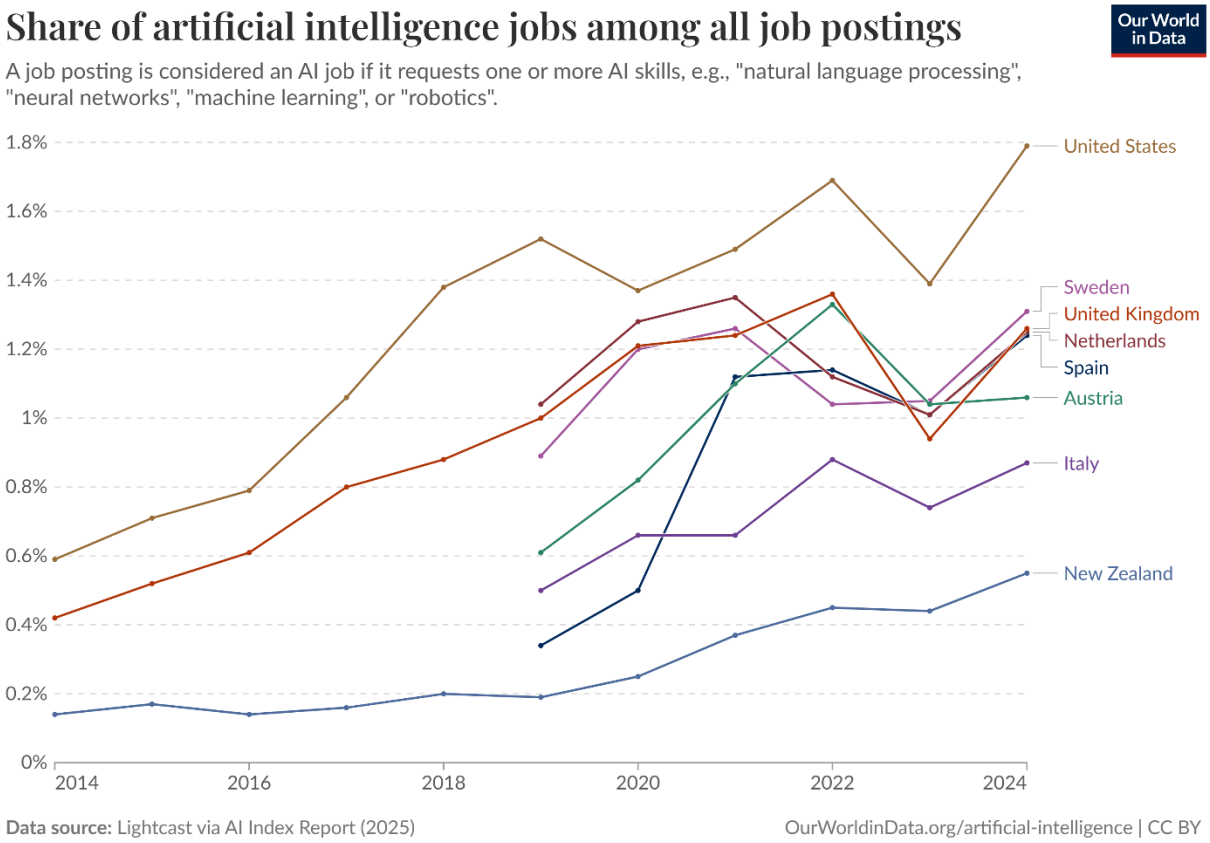


Figure 1: Percentage of AI Jobs among all Job Postings

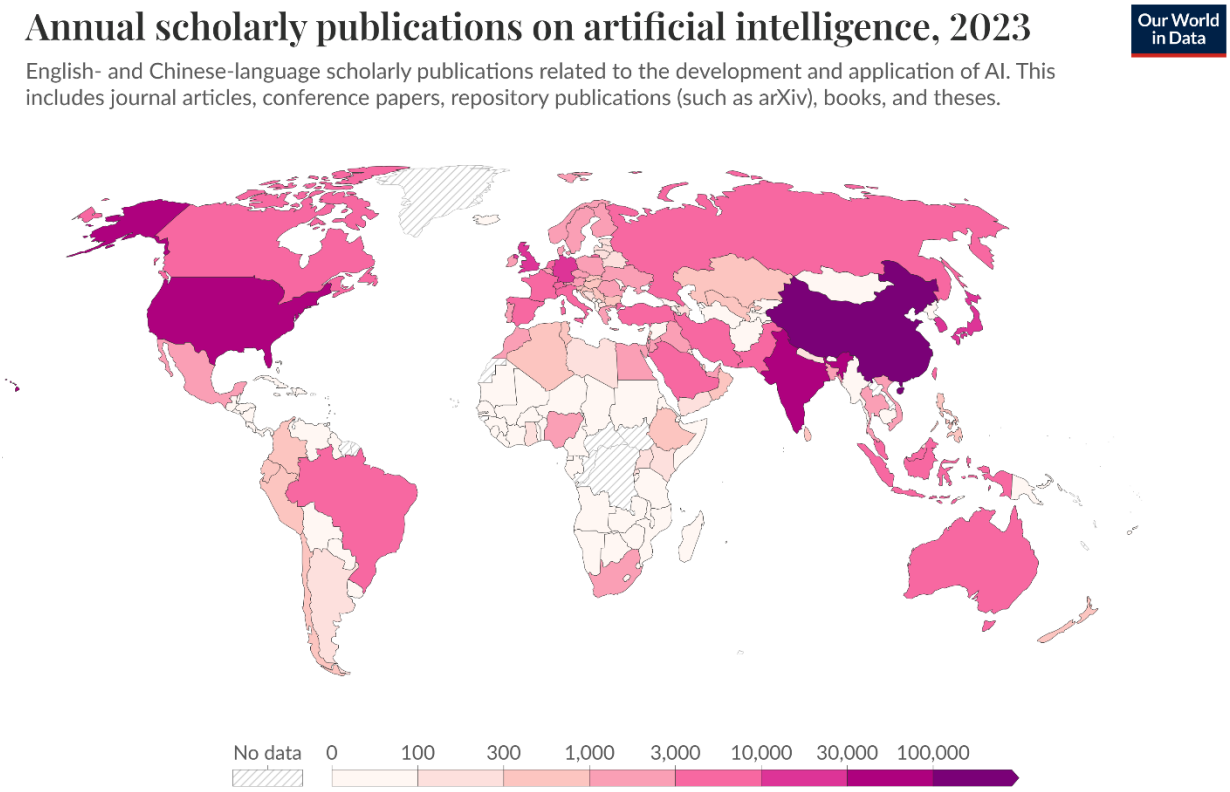


Figure 2: AI Scholarly Publications by Country for 2023

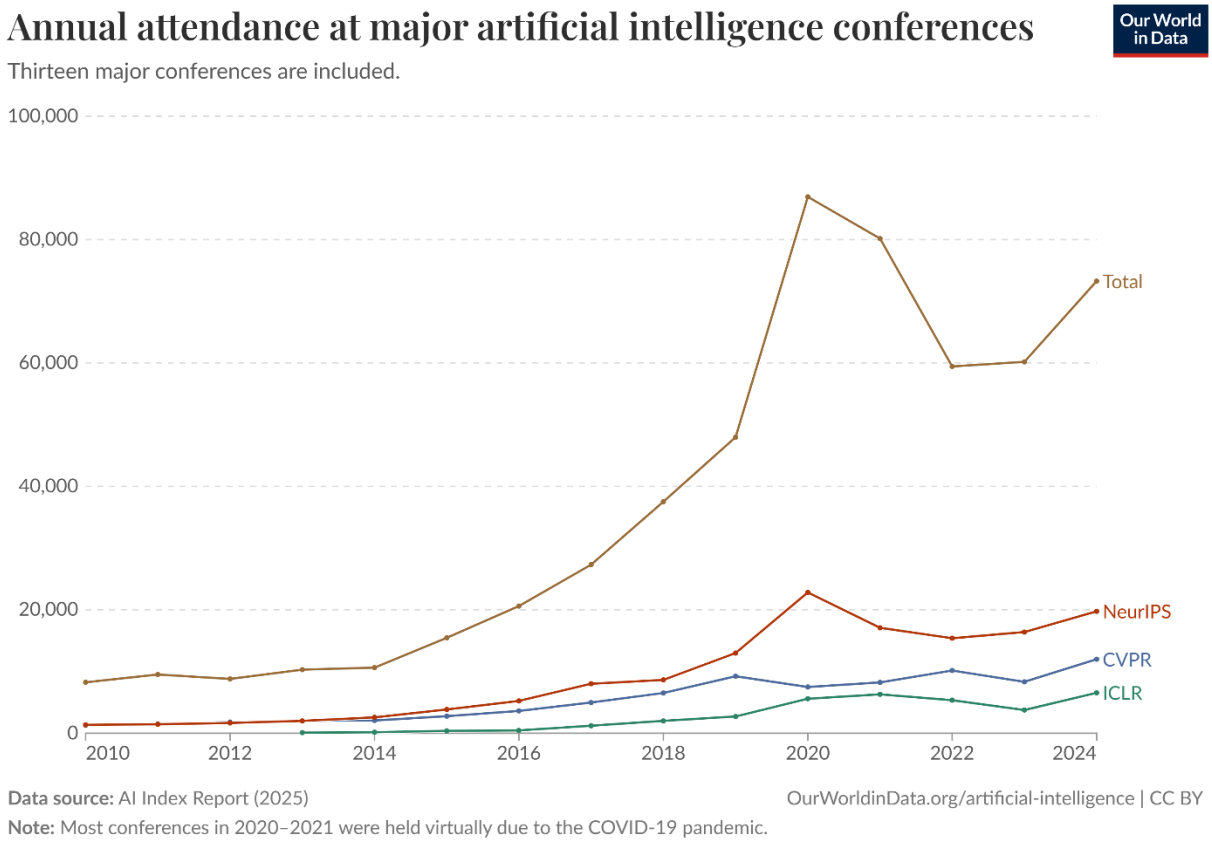


Figure 3: Annual attendance at major AI conferences

Growth in Dataset Training Size

Data sets at the abstraction of AI Sovereignty are measured in the total tokens available to train, converted into the FLOP of compute power to train. For example, Epoch AI estimates that total token availability for data training across text, multi-modal video, image, and audio will be 400 trillion to 20 quadrillion tokens, or 6e28 FLOP to 2e32 FLOP training[16].

SOURCES:

[1] G. Hamm, “2024 United States Data Center Energy Usage Report,” Lawrence Berkely National Laboratory, 2025. doi: 10.71468/P1WC7Q.
[2] “Frontier Data Centers,” Epoch AI. [Online]. Available: https://epoch.ai/data/data-centers
[3] D. Owen *et al.*, “Notable AI Models Dataset.” Feb. 12, 2026. Accessed: Feb. 14, 2026. [Online]. Available: https://epoch.ai/data/ai-models-documentation#downloads
[4] A. Laurent, “NVIDIA HGX Platform: Data Center Physical Requirements Guide,” Jan. 2026. [Online]. Available: https://intuitionlabs.ai/articles/nvidia-hgx-data-center-requirements

[5] S. Cruzes, “DATA CENTERS IN THE AGE OF AI: A TUTORIAL SURVEY ON INFRASTRUCTURE, SUSTAINABILITY, AND EMERGING CHALLENGES,” Oct. 27, 2025, *Preprints*. doi: 10.36227/techrxiv.176158592.23065552/v1.
[6] “Glossary of Data Center Terms.” [Online]. Available: https://www.pducables.com/resources/glossary-of-data-center-terms
[7] “NVIDIA DGX SuperPOD: Data Center Design Featuring NVIDIA DGX H100 Systems.” [Online]. Available: https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/index.html#dgx-superpod-data-center-design-featuring-dgx-h100
[8] “How many servers does a data center have?,” RackSolutions. [Online]. Available: https://www.racksolutions.com/news/blog/how-many-servers-does-a-data-center-have
[9] J. C. Hick and M. Izzo, “Futureproofing through 2035 for the AI and HPC Power Density Trend,” Los Alamos National Labratory, Los Alamos, Oct. 2025.
[10] A. Katal, S. Dahiya, and T. Choudhury, “Energy efficiency in cloud computing data centers: a survey on software technologies,” *Cluster Comput*, vol. 26, no. 3, pp. 1845–1875, Jun. 2023, doi: 10.1007/s10586-022-03713-0.
[11] D. Mytton, “Data centre water consumption,” *npj Clean Water*, vol. 4, no. 1, p. 11, Feb. 2021, doi: 10.1038/s41545-021-00101-w.
[12] E. Keski-Nisula, “Demand response potential of AI data center facilities — Physical flexibility of cooling, energy storage, and backup power in Finnish wholesale and reserve electricity markets,” Aalto University School of Electrical Engineering, 2025.
[13] A. Walters, A. Walker, and M. Kelley, “Rethinking AI Demand Part 1: AI Data Centers Are Experiencing a Surge of Training Demand - What Happens When the Surge is Over?” [Online]. Available: https://www.alvarezandmarsal.com/insights/rethinking-ai-demand-part-1-ai-data-centers-are-experiencing-surge-training-demand-what
[14] B. Srivathsan, M. Sorel, and P. Sachdeva, “AI power: Expanding data center capacity to meet growing demand,” McKinsey. [Online]. Available: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand#/
[15] B. Cottier, R. Rahman, L. Fattorini, N. Maslej, T. Besiroglu, and D. Owen, “The rising costs of training frontier AI models,” 2024, *arXiv*. doi: 10.48550/ARXIV.2405.21015.
[16] J. Sevilla *et al.*, “Can AI scaling continue through 2030?,” Epoch AI. [Online]. Available: https://epoch.ai/blog/can-ai-scaling-continue-through-2030