

Geographic Concentration of Machine Learning Models: How Compute Economics Re-Engineered the AI Map

Published: September 8, 2026 at 6:00 PM Author: Richard Djarbeng 15 min read

When you look at where notable artificial intelligence models come from, the global map looks empty.

In the [2024 Stanford AI Index Report](#), researchers tracked 149 notable machine learning systems built that year. The United States built 61 of them. China built 15. The European Union produced 21, the United Kingdom built 4, and Latin America, Africa, and Southeast Asia produced none.

Figure 2 shows the historical data behind this split.

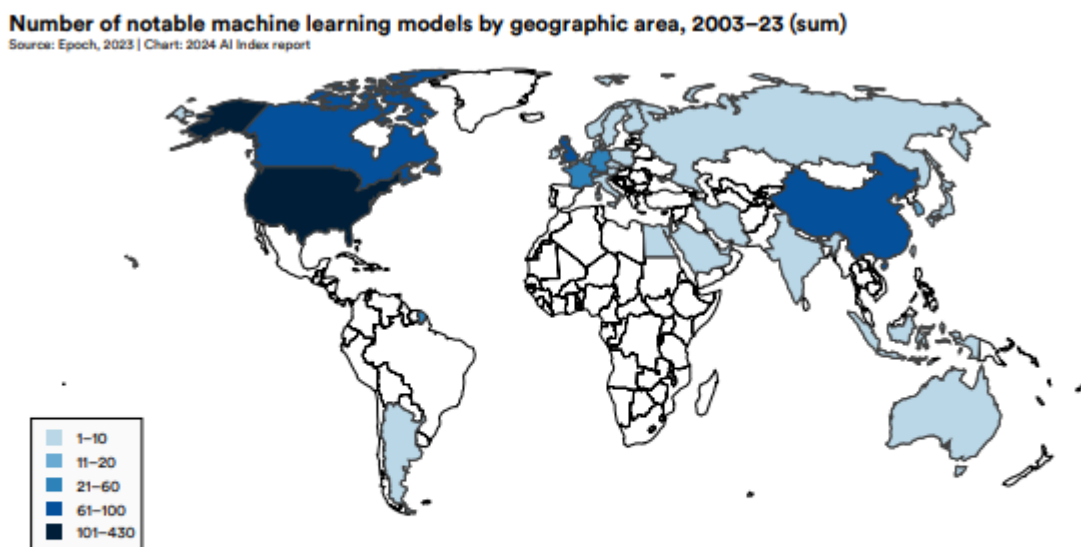


Figure 2: Number of notable machine learning models by geography (Maslej et al., 2024). The explosion in both necessary compute and associated cost has compounded concentration of breakthroughs in a few regions of the world.

Figure 2: Number of notable machine learning models by geographic area, 2003 to 2023. Data source: Epoch AI and the Stanford Institute for Human-Centered Artificial Intelligence 2024 AI Index Report.

The chart, compiled by [Epoch AI](#) and published in the Stanford report, shows that between 2003 and 2023 the United States produced around 430 notable models. China forms the second line with 85. The rest of the world sits near the bottom axis.

This is not a story about where talented computer scientists live. Skilled researchers work all over the world. The difference is that modern AI research requires electrical substations, specialized microchips, and hundreds of millions of dollars. The map in Figure 2 reflects who can pay for that hardware.

The disappearance of the academic map

The lines in Figure 2 used to look very different.

Before 2012, machine learning research happened mostly in public universities. Useful work came from wherever professors and graduate students tried new statistical ideas: early neural language models at [Université de Montréal](#), deep belief networks at the [University of Toronto](#), convolutional networks at [IDSIA in Switzerland](#), and learning theory at INRIA in France.

Most experiments ran on desktop computers with standard Intel processors or small departmental servers. A training run consumed under 10^{15} floating point operations (FLOPs), cost a few hundred dollars in power, and was funded by standard government grants. More than 80 percent of notable models built before 2012 came from university labs.

Pre-2012: University networks

[Montreal] <---> [Toronto] <---> [Lugano] <---> [Paris] <---> [Tokyo]

Hardware: Standard desktop computers, small server racks

Cost: Thousands of dollars in academic grants

Constraint: Algorithmic ideas

Post-2017: Industrial clusters

[Silicon Valley / Seattle] <=====> [Beijing / Shenzhen]

Hardware: 10,000 to 100,000 GPUs linked by fiber and liquid cooling

Cost: Tens to hundreds of millions of dollars

Constraint: Electrical substations, chip supplies, capital

University labs lost their lead when researchers showed that models got better as computers got bigger. According to [Epoch AI's institutional analysis](#), private company labs matched university output by 2014. By 2022, corporate labs produced 32 notable models

while independent university labs produced only 3. In 2023, industry produced 51 notable models, joint industry-university partnerships built 21, and universities working alone built 15.

A standard university research grant of \$500,000 cannot pay for a \$50 million training run. Researchers who wanted to work on the largest systems moved to the companies that owned the computers, and the numbers in Figure 2 shifted into the private sector.

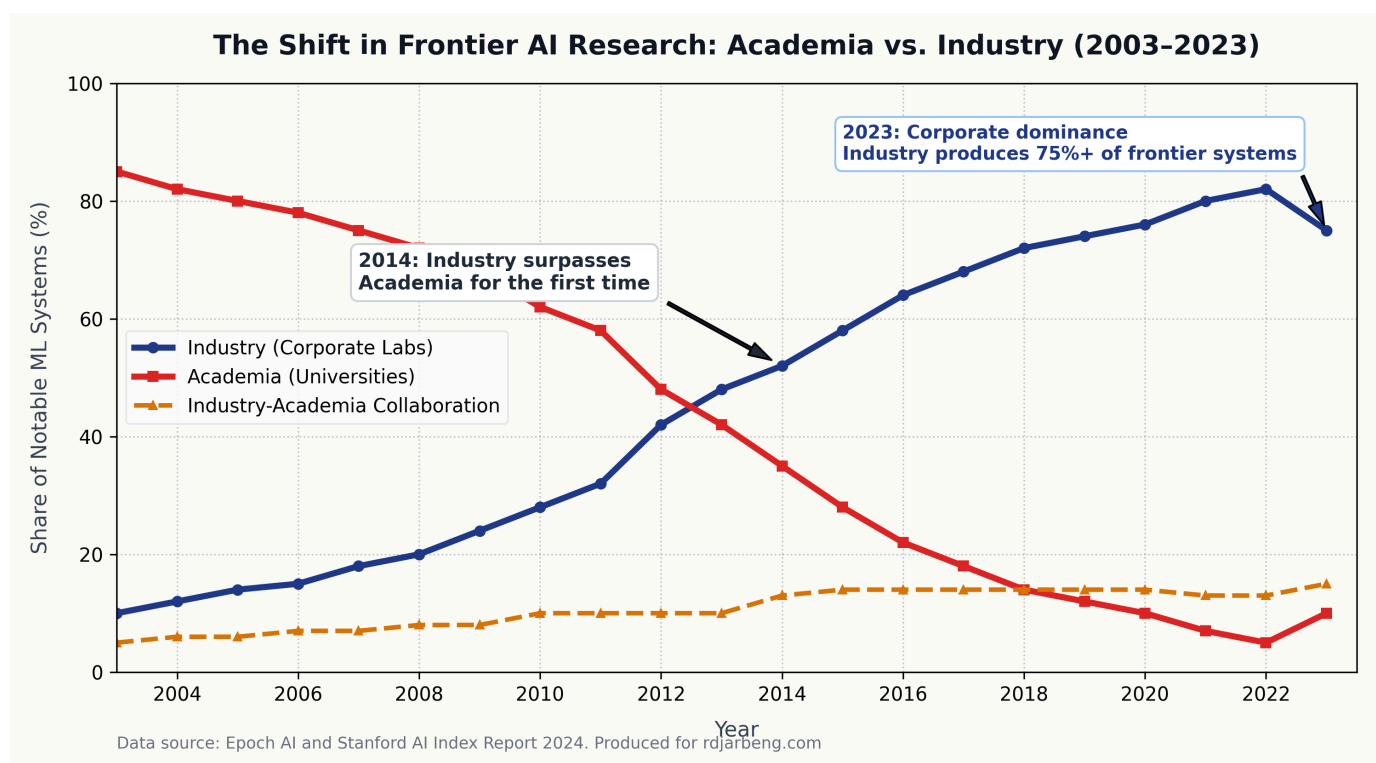


Figure 3: The shifting institutional balance in notable AI systems from 2003 to 2023. Industry output overtook universities in 2014, and by 2023 accounts for over 75 percent of notable models.

So what does this mean? The early period of machine learning was an exception. When research depended on algorithmic ideas, any university lab could lead the field. Once performance depended on hardware scale, control moved to corporate balance sheets.

The arithmetic behind the hardware cost

The shift from universities to corporate data centers comes down to an equation originally popularized in [Kaplan et al.'s scaling paper](#) and documented across [Epoch AI's compute accounting framework](#):

$$C \approx 6ND$$

$$C \approx 6ND$$

Here C is the total floating point operations needed to train a model, N is the parameter count, and D is the number of tokens (words or word fragments) in the training dataset.

The factor of 6 comes from the arithmetic of neural network training:

1. **Forward pass ($2N$ operations per token):** The model processes an input vector through its weights. Each weight requires one multiplication and one addition, so each parameter performs two operations per token.
2. **Backward pass ($4N$ operations per token):** The computer calculates two sets of numbers: the errors passed back through the layers ($2N$) and the adjustments made to each weight ($2N$). That makes the backward pass twice as large as the forward pass.

Combining the forward and backward passes gives $6N$ operations per token. For an entire dataset of D tokens, training costs $6ND$ operations.

Modern numbers make the scale clear. [Meta's Llama 3 405B](#) has 405 billion parameters ($N = 405 \times 10^9$) and trained on 15 trillion tokens ($D = 15 \times 10^{12}$):

$$C \approx 6 \times (405 \times 10^9) \times (15 \times 10^{12}) \approx 3.64 \times 10^{26} \text{ FLOPs}$$

$$C \approx 6 \times (405 \times 10^9) \times (15 \times 10^{12}) \approx 3.64 \times 10^{26} \text{ FLOPs}$$

That is 364 septillion arithmetic operations.

Training compute across model generations:

GPT-2 (2019):	1.5B params	x	40B tokens	-->	$\sim 3.6 \times 10^{20}$ FLOPs
GPT-3 (2020):	175B params	x	300B tokens	-->	$\sim 3.1 \times 10^{23}$ FLOPs
Llama 2 (2023):	70B params	x	2T tokens	-->	$\sim 8.4 \times 10^{24}$ FLOPs
Llama 3 (2024):	405B params	x	15T tokens	-->	$\sim 3.6 \times 10^{26}$ FLOPs

In 2022, DeepMind published the [Chinchilla scaling laws \(Hoffmann et al.\)](#). An earlier [2020 paper by Kaplan et al.](#) suggested that model size N should grow faster than data size D . Hoffmann showed that to get the most capability out of a given compute budget, parameters and tokens must grow together in equal proportion ($N \propto C^{0.5}$, $D \propto C^{0.5}$).

To finish 10^{26} operations in three months, thousands of specialized chips must run side by side without stopping.

So what does this mean? You cannot make up for missing hardware with clever code alone. Until someone finds a way to train models without this volume of arithmetic, frontier model creation remains tied to organizations that can run tens of thousands of processors at once.

Power, cooling, and facility costs

Running tens of thousands of processors in sync turns software work into utility engineering.

Frontier training relies on synchronous stochastic gradient descent, meaning thousands of chips exchange numbers every few milliseconds. If one network switch drops packets or a power line blinks, the whole cluster pauses.

The hardware bill for one 24,000-chip cluster looks like this:

Component	Quantity	Unit cost	Total cost
Nvidia H100/H200 GPUs	24,576 units	\$30,000 to \$40,000	\$737M to \$983M
High-Bandwidth Memory (HBM3e)	80 to 141 GB per GPU	\$15 to \$25 per GB	Included on board
Intra-node switches (NVLink 4.0)	900 GB/s per GPU	Custom silicon	~\$50M
Network fabric (InfiniBand/Spectrum-X)	400G/800G switches, optics	\$1,500 to \$2,500 per port	\$80M to \$150M
Electrical gear and liquid cooling	Chillers, pumps, cold plates	Utility grade	\$100M to \$200M
Complete facility investment	Single campus	Turnkey build	\$1.0B to \$1.5B

These facilities run into electrical and cooling limits:

- 10,000 GPUs draw about 12 megawatts of chip power, or roughly 15 to 18 megawatts for the whole building.
- 50,000 GPUs require 75 to 90 megawatts.
- A planned 100,000-chip facility approaches 150 to 180 megawatts, roughly the power used by a city of 150,000 homes.

Air cooling no longer works on racks drawing 40 to 140 kilowatts. Operators pump liquid coolant through copper plates directly against the silicon dies.

Connecting that much power takes time. According to the [Lawrence Berkeley National Laboratory](#), transmission regions like PJM in the eastern United States face queues of four to seven years to connect new high-voltage substations. In Europe, municipal authorities in

Dublin, Amsterdam, and Frankfurt have [curbed new data center power connections](#) to protect local electrical grids.

To secure power, American cloud companies have made deals directly with power plants:

- Microsoft signed a 20-year power agreement with Constellation Energy to [reopen the Unit 1 reactor at Three Mile Island](#) in Pennsylvania.
- Amazon Web Services purchased the [Cumulus data center campus](#), located beside the Susquehanna nuclear station in Pennsylvania.
- Google signed an agreement with [Kairos Power](#) to purchase electricity from seven small modular nuclear reactors by 2035.

According to [financial reporting by the Wall Street Journal](#), Microsoft, Alphabet, Amazon, and Meta each spend \$30 billion to \$50 billion annually on capital expenses. Few companies or national governments can match that spending.

So what does this mean? The bottleneck in AI is now physical infrastructure. When training requires city-scale electrical capacity and billion-dollar budgets, the geography of AI follows the balance sheets of a few large companies.

Export controls and sovereign computing

Because most frontier capacity sits in the United States and China, computing power has become part of trade and foreign policy.

The [Bureau of Industry and Security \(BIS\)](#), at the U.S. Department of Commerce used export rules to restrict chip sales:

1. **October 2022 rules:** The BIS published [interim final export rules](#) restricting exports of high-end accelerators (Nvidia A100 and H100) to China based on interconnect speeds and throughput, and prohibited foundries using American tools from making advanced chips designed by Chinese firms.
2. **October 2023 updates:** After Nvidia released modified chips for the Chinese market (the A800 and H800), the BIS issued [updated controls](#) with performance density limits, restricting those models as well.

China funded domestic manufacturing in response. Semiconductor Manufacturing International Corporation (SMIC) manufactured the Huawei Ascend 910B processor using deep ultraviolet lithography. According to [analysis by the Center for Strategic and International Studies \(CSIS\)](#) and hardware teardowns by [TechInsights](#), the chip reaches

roughly 70 to 80 percent of an Nvidia A100's performance in 16-bit math, but SMIC reports lower wafer yields (around 30 to 50 percent compared to over 90 percent at TSMC), which raises the cost per working chip. Chinese cloud operators also have to write custom software to replace Nvidia's CUDA platform and build larger networks to offset chip communication speeds.

Other governments realized that depending entirely on American cloud services meant losing control over their own data and technological capability. Several started national programs:

- **United Kingdom:** Put £300 million into the AI Research Resource to build [Isambard-AI at the University of Bristol](#), using 5,448 Nvidia GH200 processors, alongside the Dawn supercomputer at Cambridge.
- **France:** Supported Paris-based [Mistral AI](#) to build independent models, alongside the speech research lab [Kyutai](#).
- **United Arab Emirates:** Funded the Falcon model series through the [Technology Innovation Institute in Abu Dhabi](#).
- **Japan:** [Subsidized Sakura Internet and SoftBank](#) through the Ministry of Economy, Trade, and Industry to build domestic GPU clusters for Japanese language models.
- **European Union:** Set up the [EuroHPC AI Factories](#) program to add GPU partitions to public supercomputers like LUMI in Finland, Leonardo in Italy, and MareNostrum 5 in Spain.

So what does this mean? A national cluster of 5,000 GPUs cannot match a 100,000-GPU private data center. These government investments do not beat commercial frontier models; their goal is to keep domestic research and public services from relying completely on foreign providers.

Algorithmic alternatives to brute force

Researchers without massive budgets have focused on making architectures more efficient:

1. **Sparse Mixture of Experts (MoE):** In a dense model, every parameter runs for every token. Mixture of Experts models route each token to a small subset of specialized sub-networks. Mistral's [Mixtral 8x7B \(Jiang et al.\)](#) has 46.7 billion total parameters, but activates only 12.9 billion per token, reducing the compute needed during inference.

2. **Memory optimizations (FlashAttention):** Graphics processors are often held up by the time it takes to move data between off-chip memory and on-chip cache. Tri Dao's [FlashAttention](#) calculates attention directly in on-chip SRAM, cutting memory read/write cycles and speeding up attention layers by two to four times.
3. **Lower precision formats:** Using 8-bit (FP8) and 4-bit (FP4) formats instead of 16-bit floats reduces memory requirements and increases throughput on modern tensor cores. Methods like [AWQ \(Lin et al.\)](#) and [bitsandbytes](#) let researchers run models on smaller hardware.
4. **Open-weights releases:** When Meta released weights for [Llama 3](#), and teams like Mistral, Qwen, and DeepSeek published theirs, they separated model training from model use. A developer in Nairobi, Warsaw, or Tokyo does not need \$100 million to train a base model. They can download existing weights and fine-tune them on a single machine using parameter-efficient methods like [LoRA \(Hu et al.\)](#).

Some have suggested training frontier models by connecting consumer PCs across the internet, similar to [Folding@home](#). Network speeds prevent this for initial pre-training:

Metric	Datacenter fabric (NVLink / InfiniBand)	Home broadband internet	Difference
Latency	Under 2 microseconds	20 to 80 milliseconds	10,000x to 40,000x slower
Bandwidth	400 Gbps to 900 GB/s	20 to 100 Mbps upload	4,000x to 36,000x narrower
Packet loss	Zero-loss fabric	0.5% to 3% loss	Causes stalls in training
Synchronization	Synchronous SGD	Asynchronous protocols	Slower convergence

Pre-training requires nodes to synchronize gradients after every step. In a data center, that happens in microseconds. Over residential internet, waiting for thousands of PCs across the world to sync would leave processors idle most of the time.

So what does this mean? Open weights and efficient architectures let people fine-tune and run models anywhere. But training new base models from scratch still requires large, tightly connected physical facilities.

The current map

The geographic split in Figure 2 is the result of capital costs, power supplies, and network physics. As long as frontier models require septillions of operations (\$\$C \approx 6ND\$\$), and as long as training requires microsecond-level connections between chips, frontier

development stays where money, silicon, and electricity meet.

Today the field has settled into three groups:

1. **The American cloud companies:** Microsoft, Alphabet, Amazon, and Meta, backed by large capital budgets, private energy contracts, and clusters of 50,000 to 100,000 GPUs.
2. **The Chinese state and industry ecosystem:** Using domestic chip manufacturing and custom network layouts to build frontier systems under trade restrictions.
3. **The regional and open-source perimeter:** Governments funding domestic clusters to keep strategic independence, and developers using open weights to build software without paying for pre-training.

Until someone develops an architecture that learns without massive matrix arithmetic, the distribution in Figure 2 will persist. The map of artificial intelligence is drawn by electrical grids, chip manufacturing, and corporate capital.

Related Concepts & Posts

The following cards provide further definitions of concepts discussed above, and possible links to related posts.

Training Compute FLOPs

Hardware Scale

Exponential Growth

Cumulative floating point operations expended during model training, growing from 10^{18} FLOPs in 2012 to over 10^{26} FLOPs in modern frontier systems.

[Explore Dataset](#)

Chinchilla Optimal Scaling

Parametric Math

$C \approx 6ND$

Hoffmann et al. formulation demonstrating that compute-optimal training requires scaling model parameters and training dataset tokens in equal proportion.

[Read Scaling Paper](#)

Sovereign Compute Infrastructure

Geopolitics

Capital Concentration

National and regional investments in high-bandwidth GPU clusters to counter bilateral dominance and secure autonomous foundational AI capability.

[View 2024 AI Index](#)