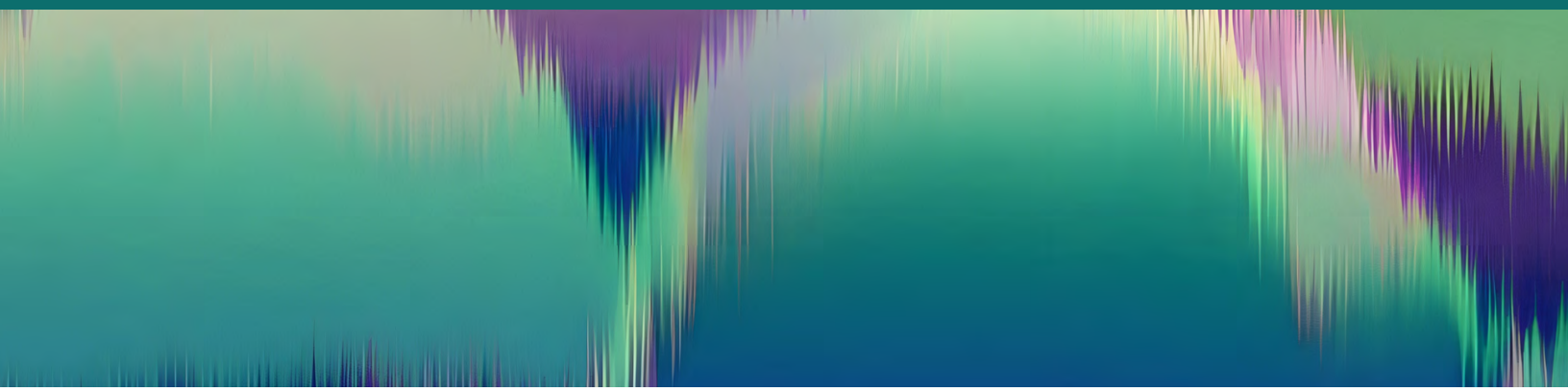
1

Research and Development

Overview

The resources powering AI development continued to grow in 2025, but fewer notable models were released than the year before, and the systems at the frontier are increasingly concentrated among a small number of organizations. Industry now accounts for over 90% of notable AI models, and the most capable systems are also the least transparent, with training code, dataset sizes, and parameter counts increasingly withheld. The computing power behind these models has grown roughly 3.3 times per year since 2022, yet almost all of it flows through a single chip foundry in Taiwan, making the global hardware supply chain fragile. Open-source development and AI publications continued to grow, and the research landscape is becoming more geographically distributed. China now leads in publication volume, citation share, and patent grants, while smaller countries like Switzerland and Singapore lead in AI researchers per capita. Yet some dimensions of the field have not changed at all. Gender gaps in AI talent remain deeply entrenched, with no meaningful progress in any country since 2010. This chapter covers the research and development pipeline, from the landscape of AI models through the compute, data centers, energy, and open-source software that support them, to the broader research ecosystem of publications, patents, and talent.

Contents



Chapter Highlights

1. **Industry produced over 90% of notable AI models in 2025, but the most capable models are now the least transparent.** Training code, parameter counts, dataset sizes, and training duration are no longer disclosed for several of the most resource-intensive systems, including those from OpenAI, Anthropic, and Google.

2. **China leads in research, while the U.S. leads in notable model development.** China leads in publication volume, citations, and patent grants, while the U.S. retains higher-impact patents and produced 50 notable models in 2025 to China's 30. South Korea leads in AI patents per capita, and China's share of the top 100 most-cited AI papers grew from 33 in 2021 to 41 in 2024.

3. **Reported parameters held in the trillions as disclosure dropped.** Parameter counts have stayed near 1 trillion for three years, though reporting from frontier labs has stopped. Training compute, which can be estimated independently, has continued to rise.

4. **Synthetic data is still not replacing real data in pre-training, but data quality and post-training techniques are showing promise.** OLMo 3.1 Think 32B, with nearly 90 times fewer parameters than Grok 4, achieves comparable results on several benchmarks through pruning, deduplication, and curation alone.

5. **Global AI compute capacity grew 3.3x per year since 2022, reaching 17.1 million H100-equivalents.** Nvidia accounts for over 60% of total compute, with Google and Amazon supplying much of the remainder and Huawei holding a small but growing share. The buildout is being driven by hyperscaler data center expansion and sustained demand for frontier model training and inference.

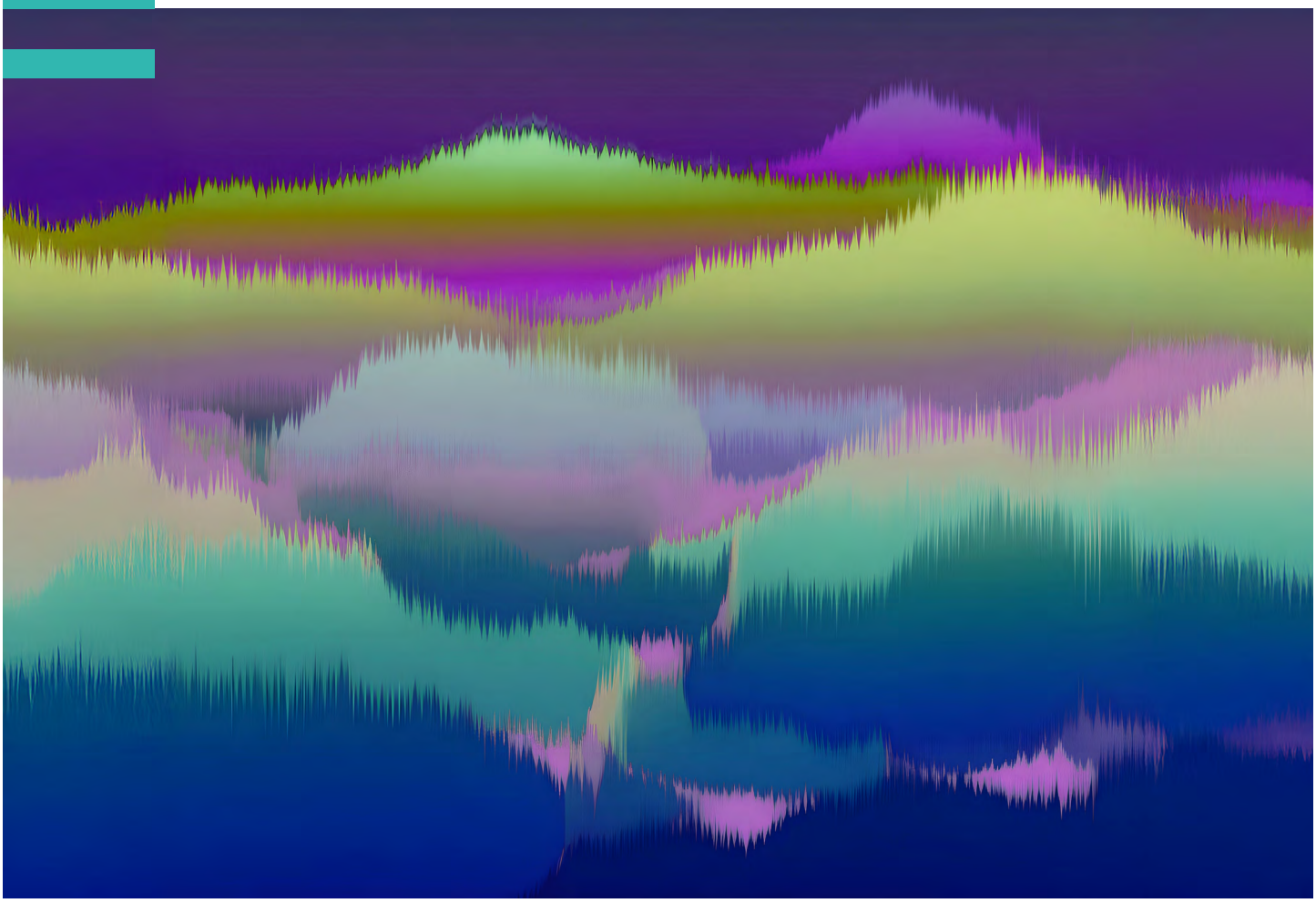
6. **The United States leads in AI data centers, and one Taiwanese foundry fabricates the majority of chips inside them.** The United States hosts 5,427 data centers, more than ten times any other country, consuming more energy than any other region. A single company, TSMC, fabricates almost every leading AI chip and makes the global AI hardware supply chain dependent on one foundry in Taiwan, though a TSMC-U.S. expansion began to operate in 2025.

7. **AI's environmental footprint increases across power, water, and emissions.** In 2025, Grok 4's estimated training emissions reached 72,816 tons of CO_2 equivalent. AI data center power capacity rose to 29.6 GW, comparable to New York state at peak demand, and annual GPT-4o inference water use alone may exceed the drinking water needs of 12 million people.

8. **Open-source AI development continues to scale, with 5.6 million projects on GitHub and Hugging Face uploads tripling since 2023.** U.S.-based projects still attract the most engagement, with 30 million cumulative GitHub stars across projects that have crossed the 10-star threshold.

9 **The number of AI researchers and developers moving to the United States has dropped 89% since 2017.** The decline is accelerating, down 80% in the last year alone. The U.S. is still home to more AI talent than any other country, but it is attracting new talent at the lowest rate in over a decade.

10 **The AI talent map is shifting, but gender gaps remain deeply entrenched.** Switzerland and Singapore lead the world in AI researchers and developers per capita and some countries show relatively higher female representation, including Saudi Arabia (32.3%), Canada (29.6%), and Australia (30.1%), though no country approaches gender parity.



1.1 Notable AI Models

This section starts with the models themselves. Using Epoch AI's curated dataset of notable models, this section examines where frontier AI models are coming from, how they are deployed, and what it takes to build them. Epoch AI designates models as noteworthy based on criteria such as state-of-the-art advancements, historical significance, or high citation rates. This is a manual curation, so the dataset is not a census of all AI models or a full map of all model development activity.[1] Trends should be read as patterns within the domain. The sections that follow track the infrastructure and inputs behind these systems, including compute, data centers, energy costs, and open-source software, before looking at the broader research ecosystem through publications, patents, and talent.

This chapter focuses on the research and development pipeline and its inputs. The next chapter, Technical Performance, reviews model capabilities and benchmark performance in detail.

By National Affiliation[2]

Notable model production remains concentrated within a small number of countries (Figures 1.1.1–1.1.3). Historically, the United States has produced the largest in total output numbers, followed by China. This pattern continued in 2025 as the United States led with the release of 50 notable AI models, China with 30, and South Korea with 5. The number of new model releases declined year over year across all major geographic areas.

Number of notable AI models by select geographic areas, 2025

Source: Epoch AI, 2026 | Chart: 2026 AI Index report

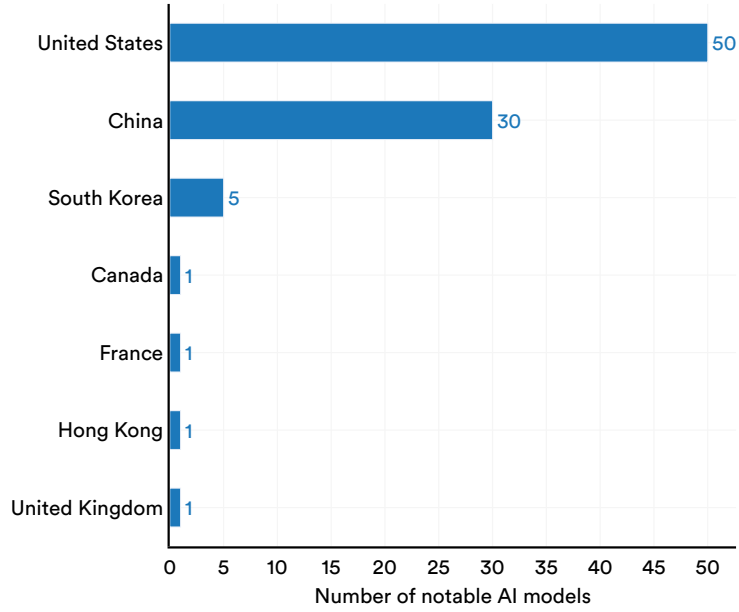


Figure 1.1.1[3]

1 New and historic models are continually added to the Epoch AI database, so the total year-by-year counts of models included in this year's AI Index might not exactly match those published in last year's report. The data is based on a snapshot taken on February 12, 2026.

2 A machine learning model is associated with a specific country if at least one author of the paper introducing it is affiliated with an institution based in that country. In cases where a model's authors come from several countries, double-counting can occur.

3 This chart highlights model releases from a select group of geographic areas. More comprehensive data on model releases by country will be available in the upcoming AI Index Global Vibrancy Tool.

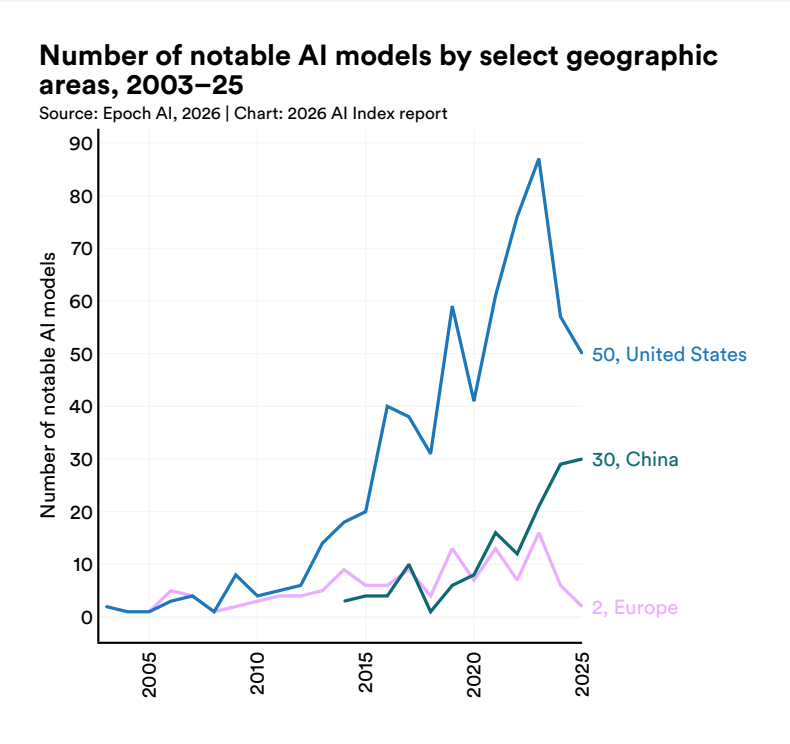


Figure 1.1.2

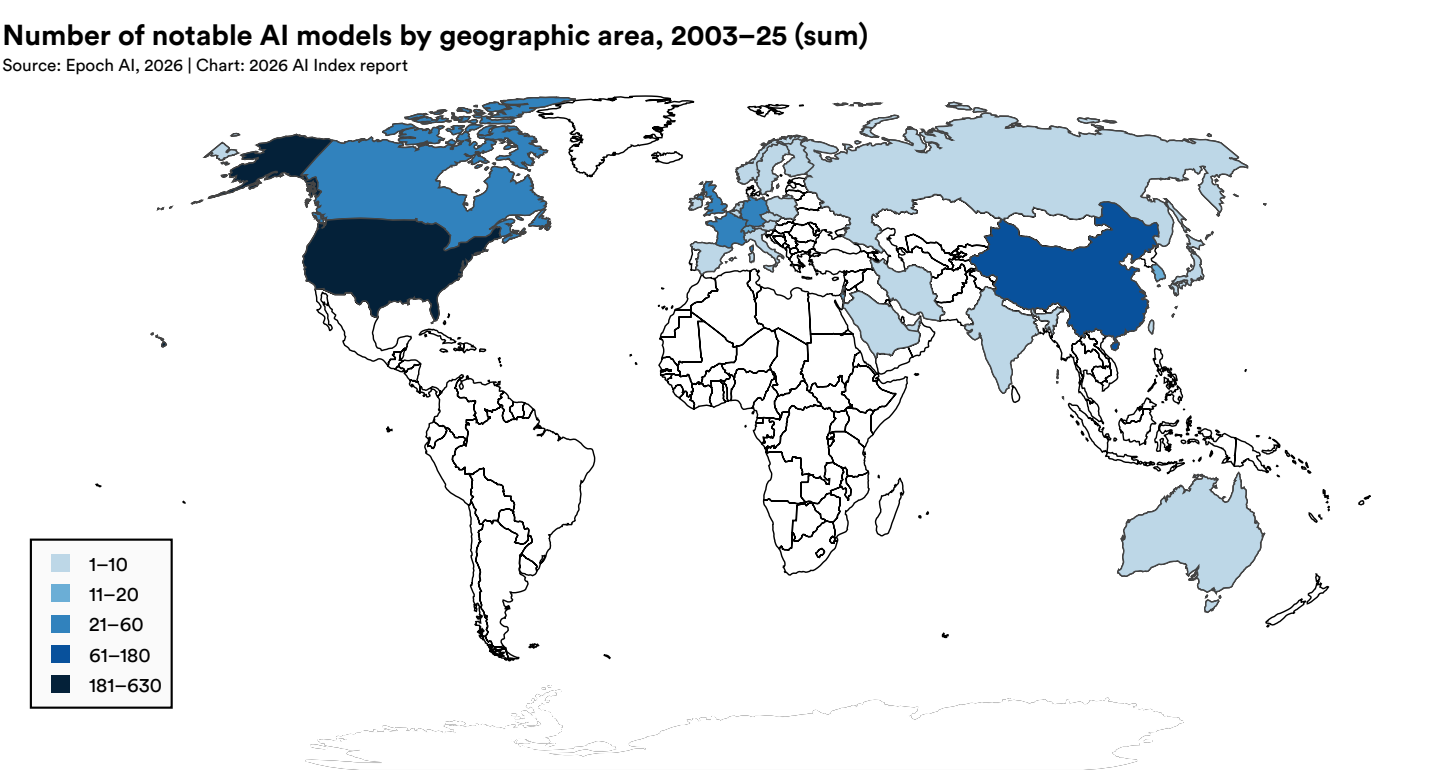


Figure 1.1.3

By Sector and Organization

The development of notable AI models continues to be predominantly concentrated in industry (Figures 1.1.4 and 1.1.5). Over the past decade, the share produced by industry has grown steadily and now represents the largest share by a wide margin (91.6%). In 2025, Epoch AI identified one notable AI model originating from academia, compared to 87 from industry.

Within industry, a small set of organizations account for a large share of releases (Figures 1.1.6 and 1.1.7). In 2025, the top contributors were OpenAI (19), Google (12), and Alibaba (11). Since 2014, Google has produced the largest number of notable models, followed by Meta and OpenAI. Within academia, Tsinghua University (26), Stanford University (26), and Carnegie Mellon University (25) have been the most prolific over the past decade.

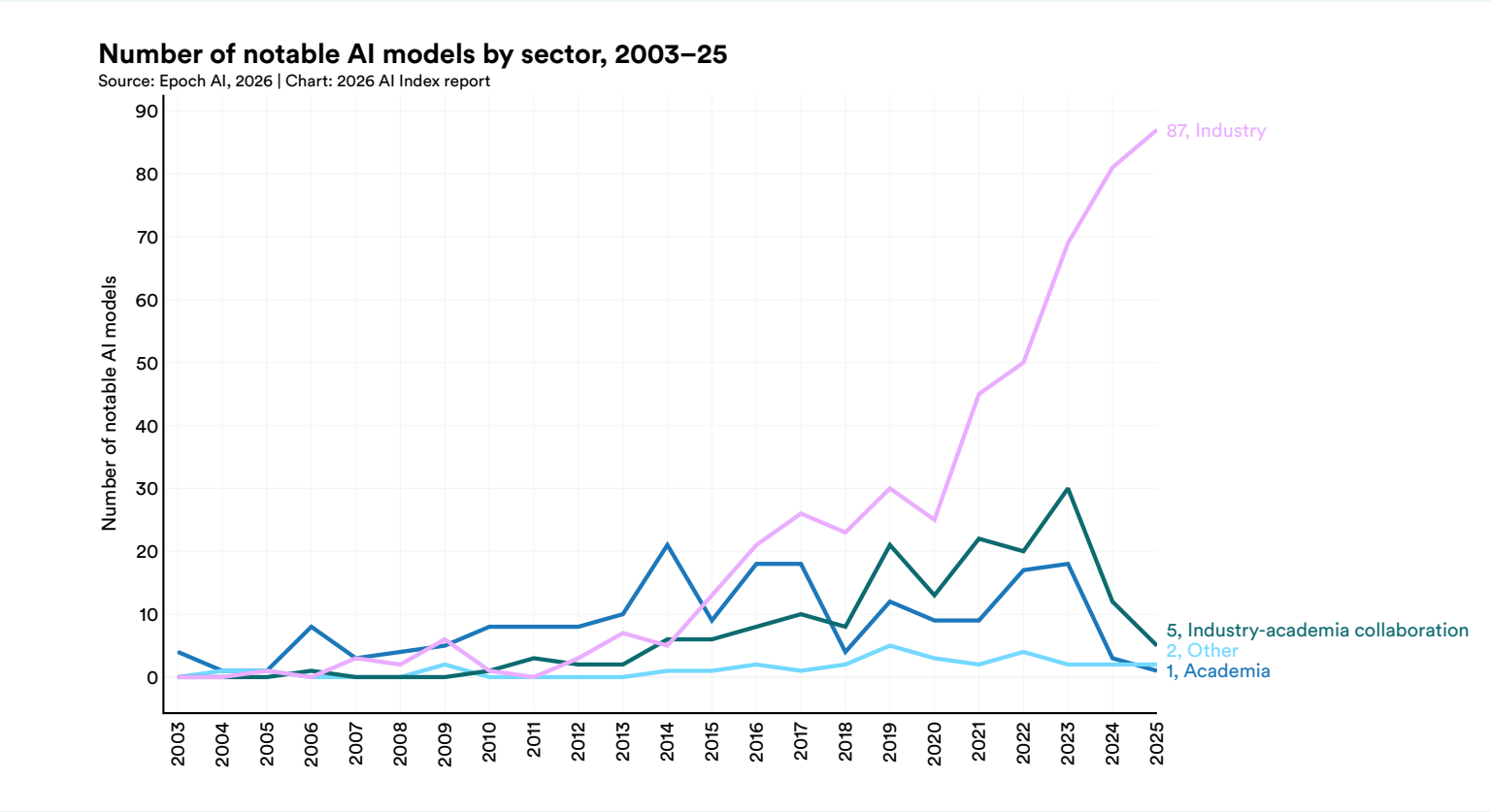


Figure 1.1.4

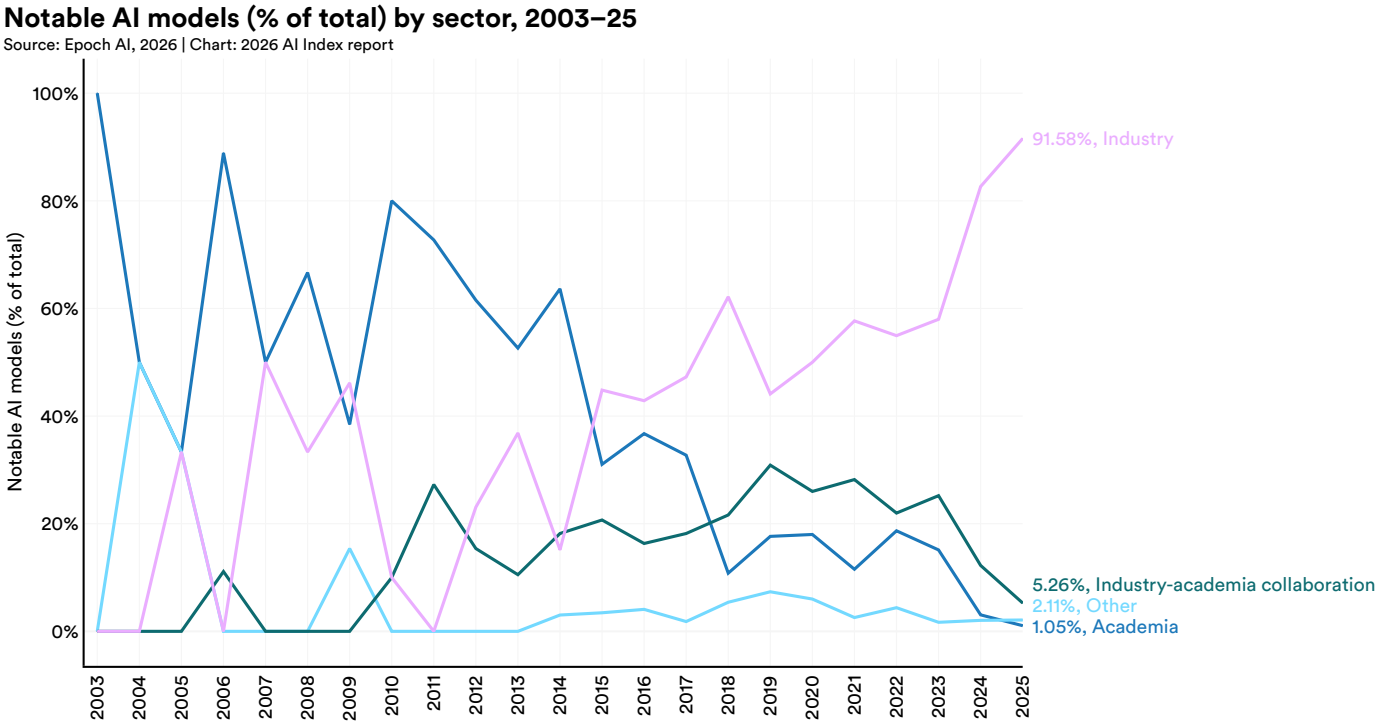


Figure 1.1.5

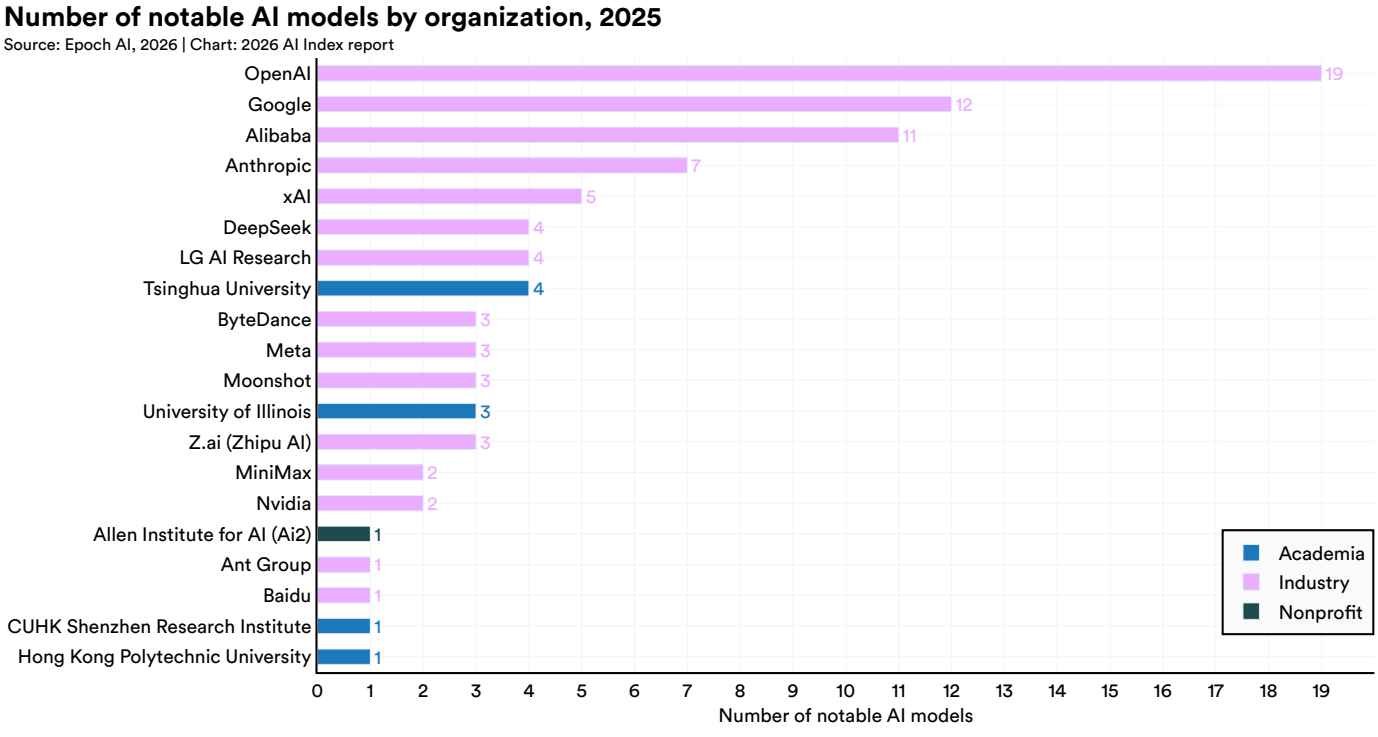


Figure 1.1.6[4]

4 In the organizational tally figures, research published by DeepMind is classified under Google.

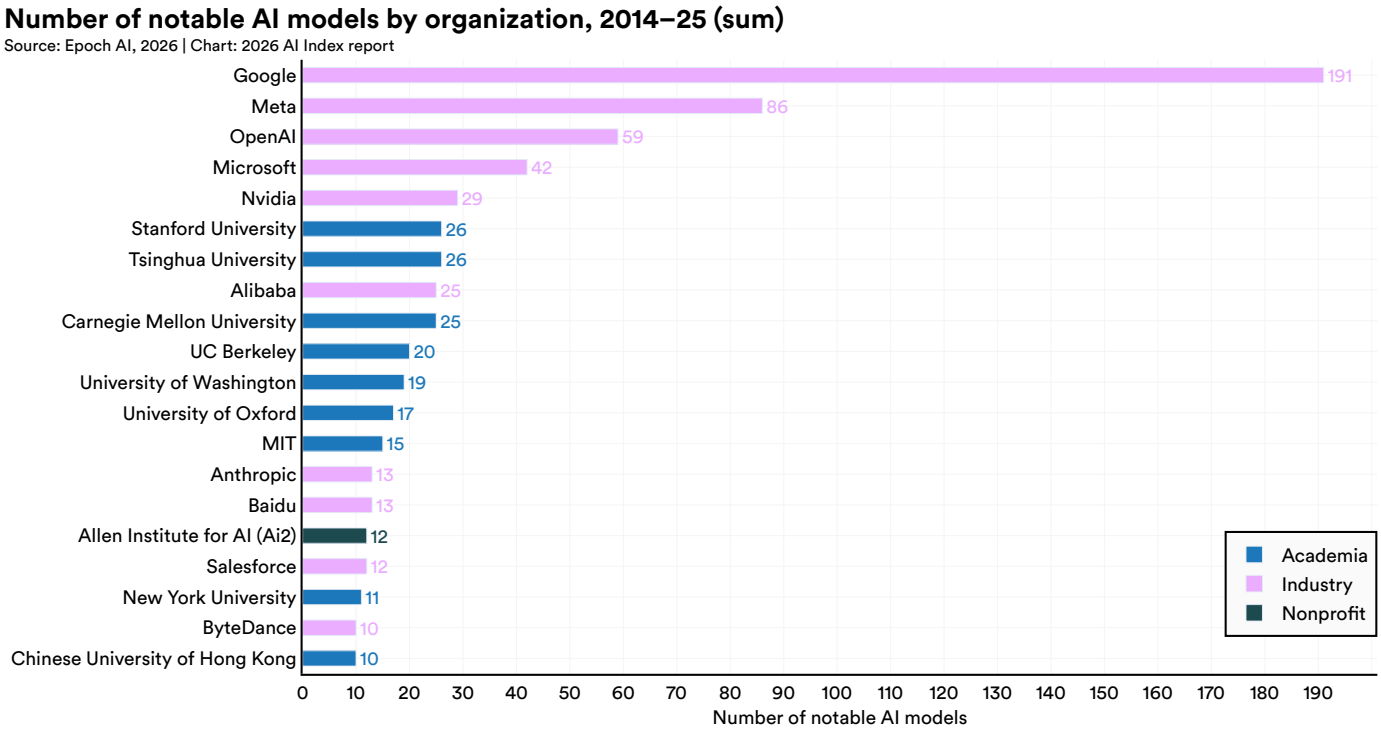


Figure 1.1.7

Model Release

Release patterns for notable AI models have continued to shift toward controlled access (Figure 1.1.8). In 2025, API access was the most common release type, with 45 of 95 models made available this way. and API-only releases have steadily increased since 2020. The second most common release type was "open weights (unrestricted)," meaning the models are fully available for use, modification, and redistribution. The remaining models were released in a mix of access types, including "hosted access (no API),"[5] "open weights (restricted use),"[6] and "open weights (noncommercial)." The "unknown" designation refers to models that have unclear or undisclosed access types, and "unreleased" models remain proprietary, accessible only to their developers or select partners.

Training code is becoming even less accessible than model code overall (Figure 1.1.9). In 2025, 80 of 95 notable models were released without their corresponding training code, compared to 4 that made their code "open source." In 2020, models with open source and unreleased training code were about the same in number, but by 2023, the majority were unreleased and the gap has continued to widen. This growing opacity limits the ability of external researchers to reproduce results, audit development, and validate safety claims. These challenges are central to the responsible AI and governance discussions in Chapter 3 and Chapter 8.

5 Hosted access refers to using computing resources or services (such as software, hardware, or storage) provided remotely by a third party, rather than personally owning or managing them. Instead of running software or infrastructure locally, hosted access involves accessing these resources via the cloud or another remote service, typically over the internet. For example, using GPUs through platforms like AWS, Google Cloud, or Microsoft Azure—rather than running them on one's own hardware—is considered hosted access.

6 Open weights models share their architecture at varying levels of restriction, "noncommercial" limits use to research purposes, "restricted use" permits broader use with some conditions, and "unrestricted" places no limitations on use, modification, or redistribution.

Number of notable AI models by access type, 2014–25

Source: Epoch AI, 2026 | Chart: 2026 AI Index report

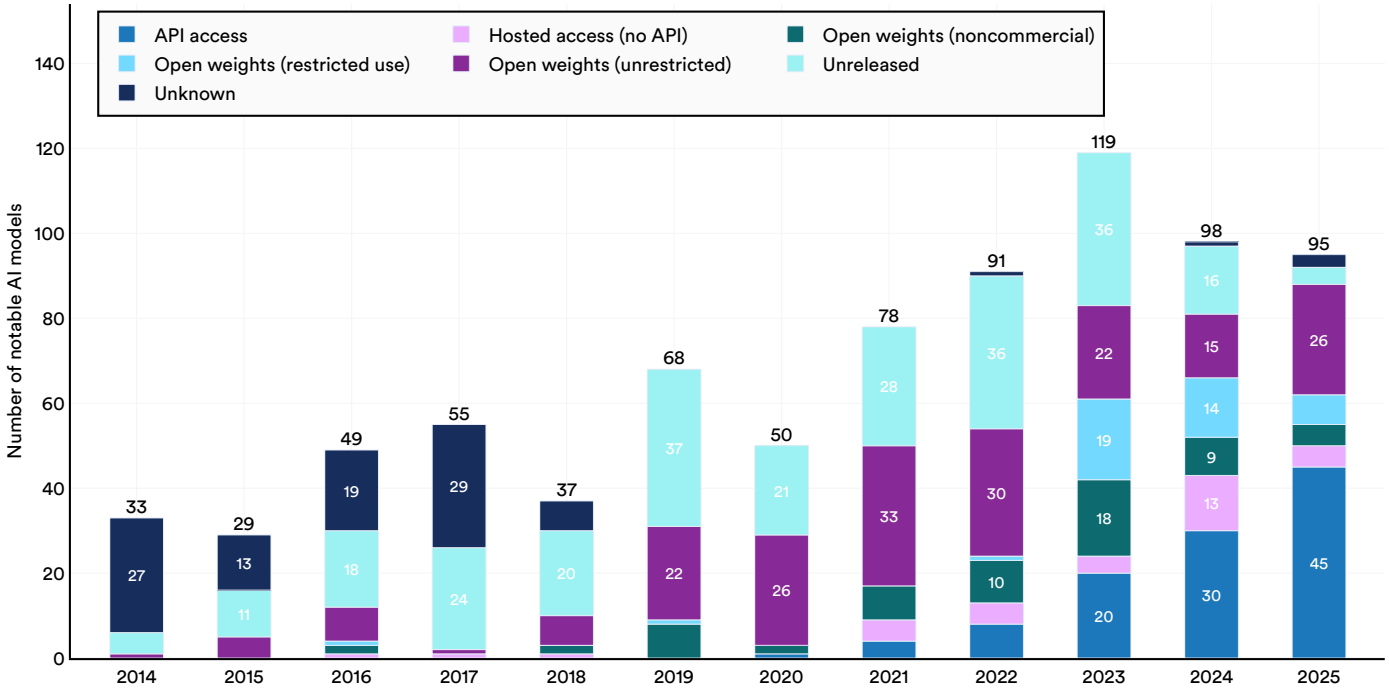


Figure 1.1.8[7]

Number of notable AI models by training code access type, 2014–25

Source: Epoch AI, 2026 | Chart: 2026 AI Index report

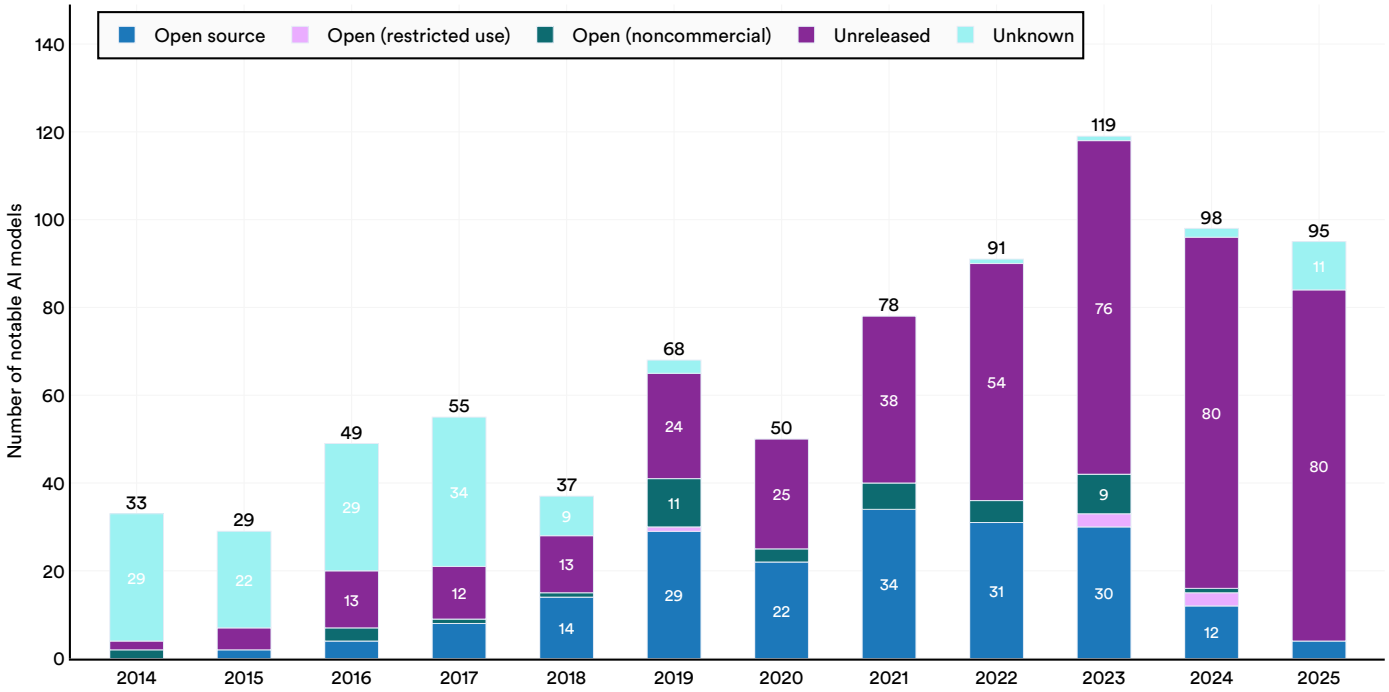


Figure 1.1.9

7 Not all models in the Epoch database are categorized by access type, so the totals in Figures 1.1.8 and 1.1.9 may not fully align with those reported elsewhere in the chapter.

Parameter and Compute Trends

Parameter counts for notable AI models have increased significantly from the early 2010s through 2022, driven by the growing complexity of model architecture, greater data availability, improvements in hardware, and proven efficacy of larger models (Figures 1.1.10–1.1.12[8]). Since then, growth in reported parameter counts has flattened, but this is likely understating actual growth due to the absence of certain data points. Several of the most resource-intensive models released in recent years, including those from OpenAI, Anthropic, and Google, have not publicly disclosed parameter counts, training dataset sizes, or training duration.

Similarly, training dataset sizes and training duration increased through the early 2020s, with leading models training on tens of trillions of tokens over periods exceeding 100 days. Again, due to limited disclosure from major frontier labs, the more recent data is incomplete.

Number of parameters of notable AI models by sector, 2003–25

Source: Epoch AI, 2026 | Chart: 2026 AI Index report

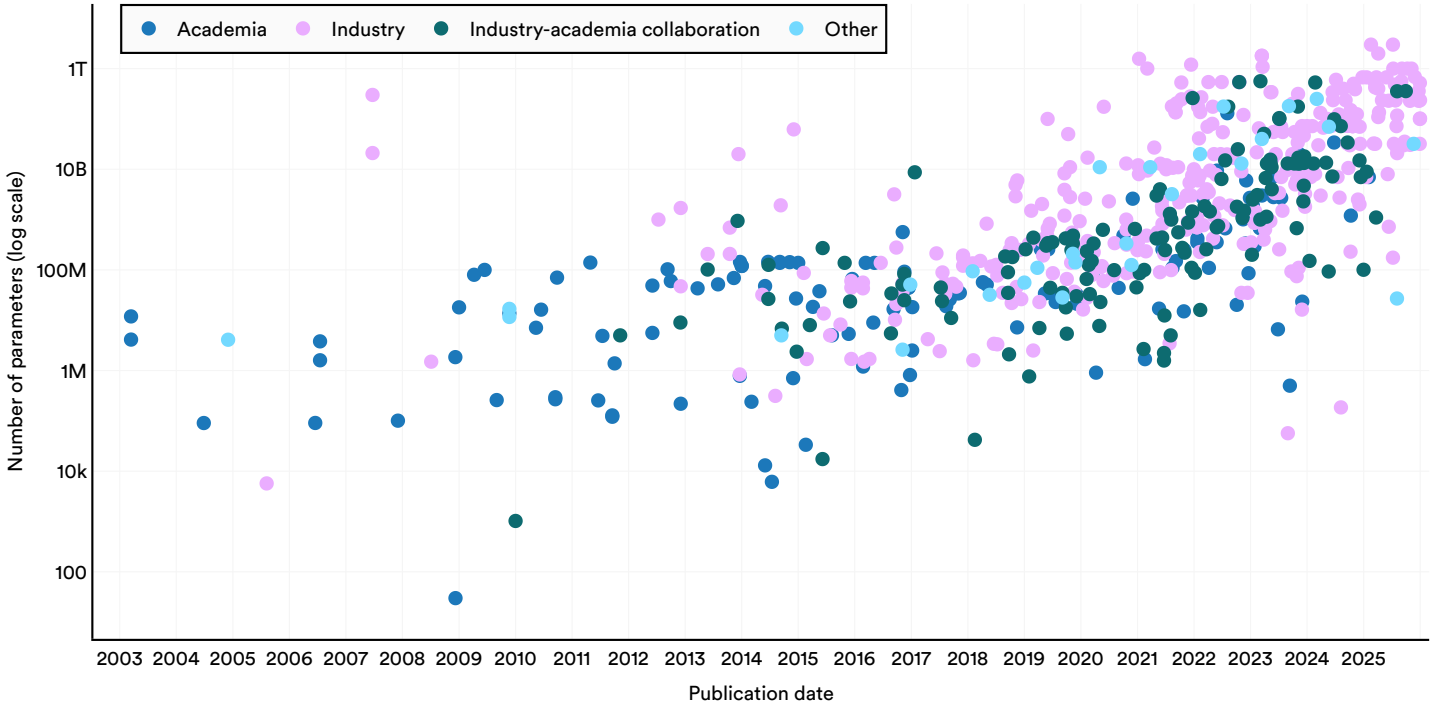


Figure 1.1.10

8 Several of the figures in this section use a log scale to reflect the exponential growth in AI model parameters and compute in recent years.

Training dataset size of notable AI models, 2010–25

Source: Epoch AI, 2026 | Chart: 2026 AI Index report

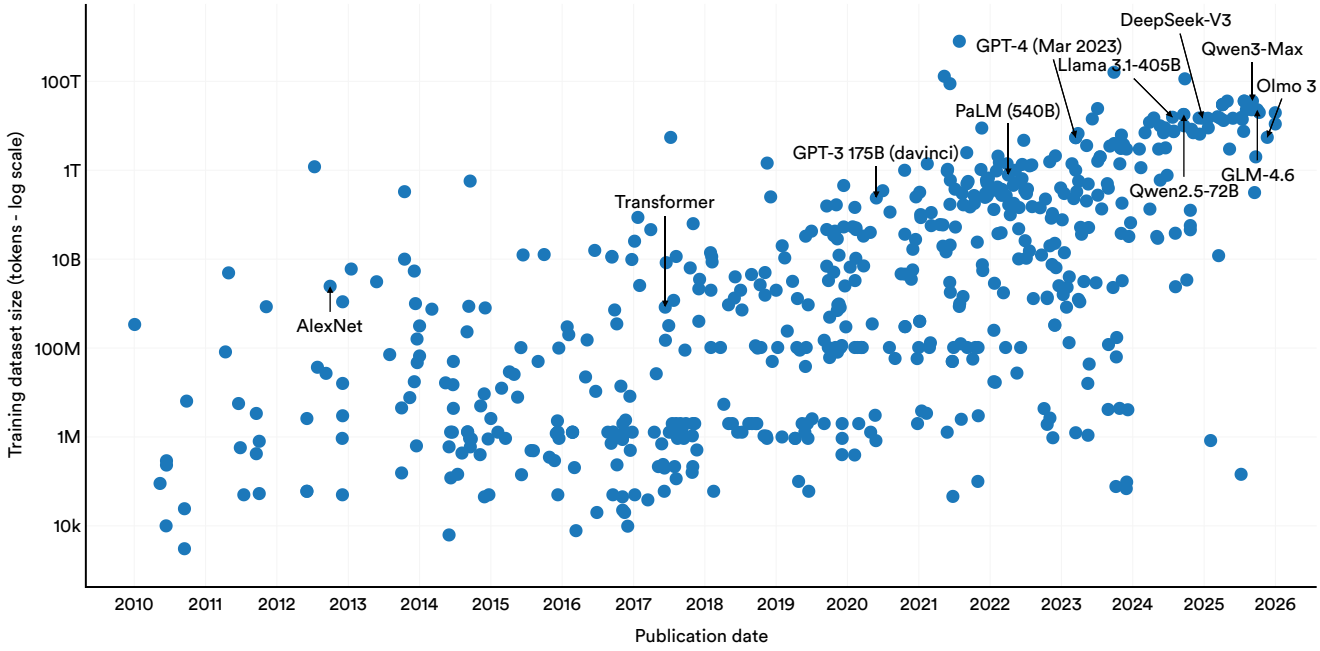


Figure 1.1.11

Training time of notable AI models, 2010–25

Source: Epoch AI, 2026 | Chart: 2026 AI Index report

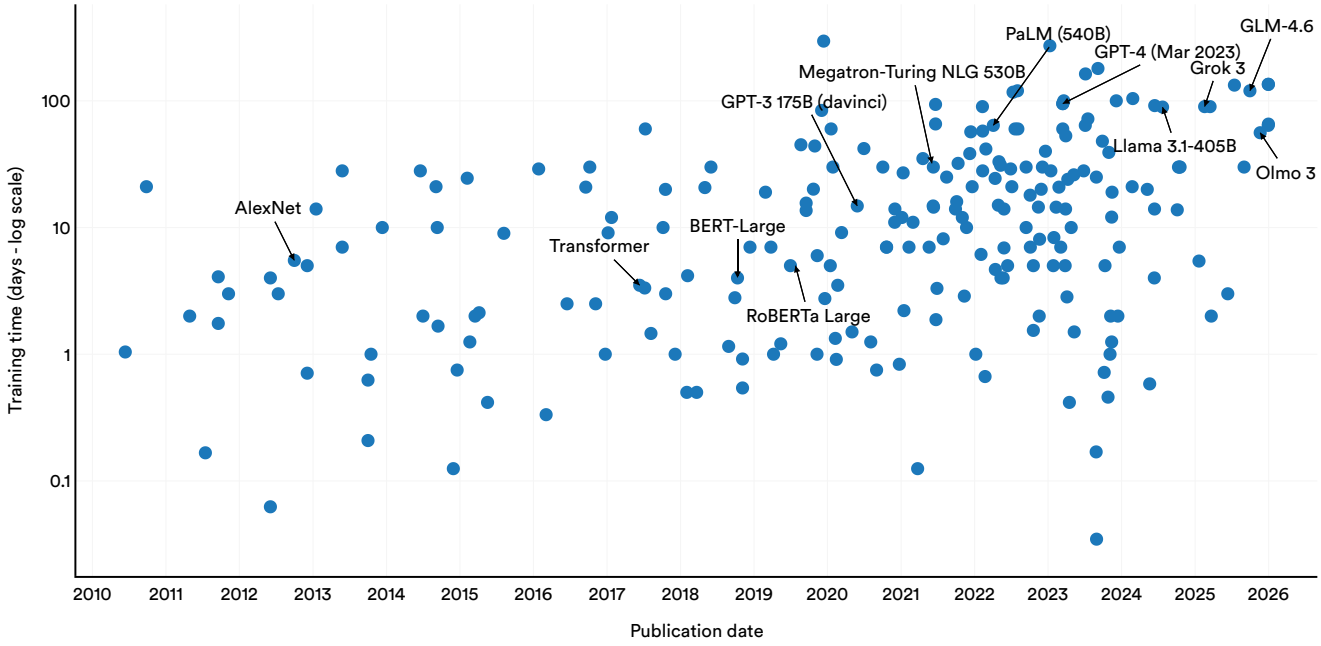


Figure 1.1.12

Since compute can be estimated even when not directly reported, training compute trends for notable models show clear growth over the same period (Figures 1.1.13 and 1.1.14). Compute requirements for notable models have risen by several orders of magnitude, with industry accounting for the highest values. When comparing the two countries with highest model output, U.S. models continue to be the most computationally intensive compared to Chinese models. However, the comparison in recent years cannot be fully substantiated because U.S. models have not directly reported their training compute.

Training compute of notable AI models by sector, 2003–25

Source: Epoch AI, 2026 | Chart: 2026 AI Index report

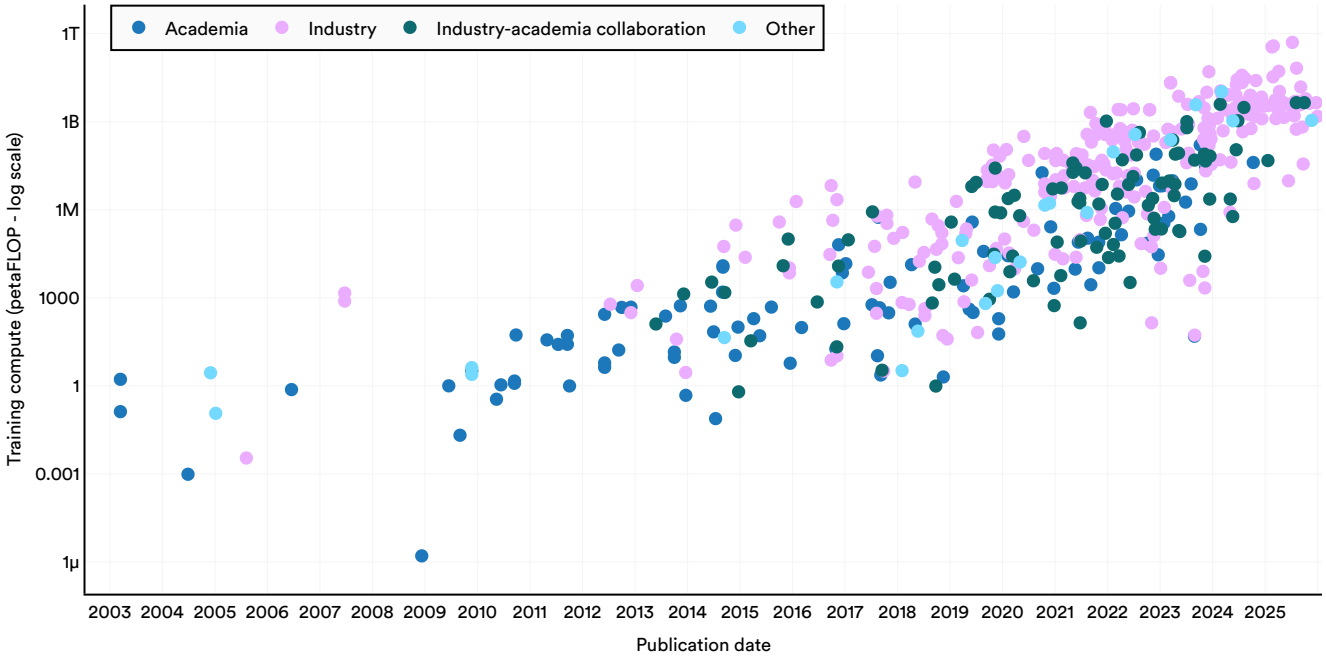


Figure 1.1.13[9]

Training compute of select notable AI models in the United States and China, 2018–25

Source: Epoch AI, 2026 | Chart: 2026 AI Index report

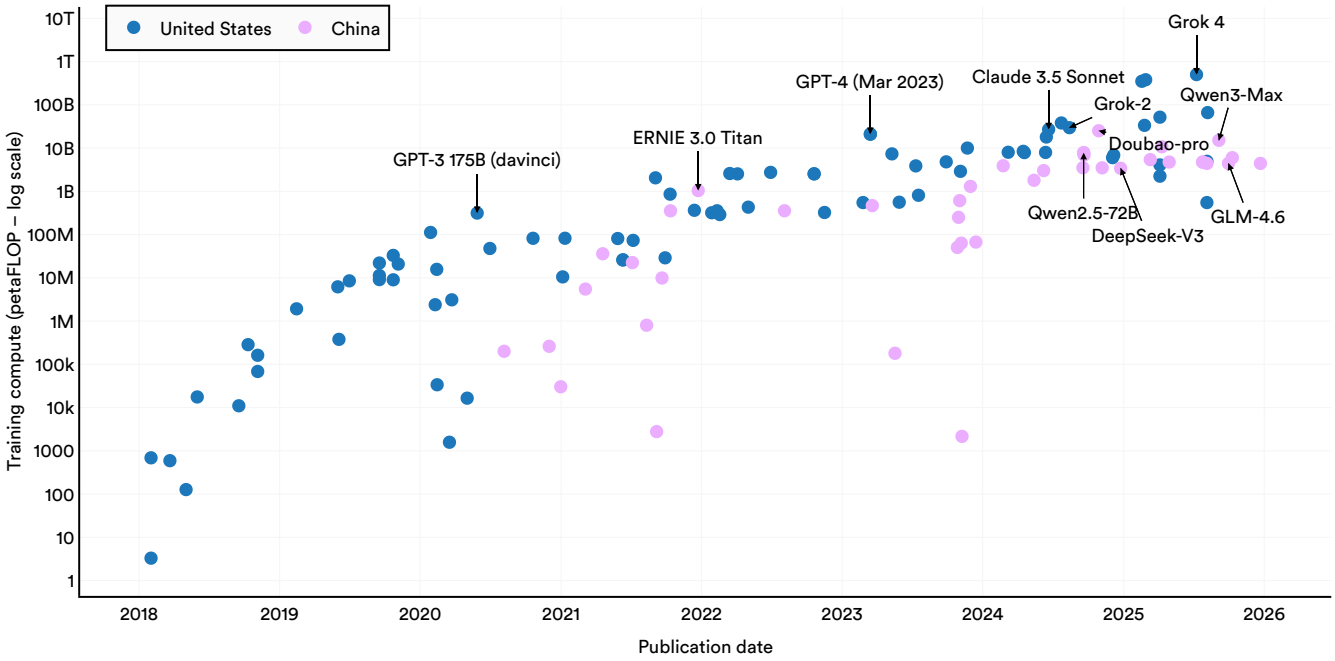


Figure 1.1.14

9 Estimating training compute is an important aspect of AI model analysis, yet it often requires indirect measurement. When direct reporting is unavailable, Epoch estimates compute by using hardware specifications and usage patterns or by counting arithmetic operations based on model architecture and training data. In cases where neither approach is feasible, benchmark performance can serve as a proxy to infer training compute by comparing models with known compute values. Full details of Epoch's methodology can be found in the documentation section of their website.

HIGHLIGHT:

Will Models Run Out of Data?

Last year, the AI Index highlighted concerns around data bottlenecks and the sustainability of the scaling approach as it relates to training data. Leading AI researchers have publicly claimed that the available pool of high-quality human text and web data for training large models has been exhausted, a state often referred to as "peak data." This has continued to raise industry-wide concerns about the sustainability of scaling laws, which have historically depended on ever-larger datasets. One set of projections from Epoch AI suggests that, under certain assumptions, the estimated depletion date could fall between 2026 and 2032.

Synthetic Data in Pre-training

Limits on the availability of real-world data may be less consequential if synthetic data (data generated by AI systems) can be used to improve the performance of subsequent models. Previous editions of the AI Index found no definitive evidence that synthetic data improves model performance during the pre-training phase.[10] The 2024 report referenced research suggesting that model performance can collapse when real training data is replaced with synthetic data. The 2025 report noted more recent findings that such collapse can be avoided if real data remains part of the training set, but that simply adding more data does not necessarily lead to performance gains.

The consensus remains largely unchanged. There is still no definitive evidence that synthetic data can fully offset real-data depletion in pre-training contexts. However, recent research suggests that synthetic data may offer value in more limited settings. Hybrid training approaches, which combine real and synthetic data, can significantly accelerate training, sometimes by a factor of five to 10 at scale, without surpassing real data in final model performance. Training on purely synthetic data has shown promise for smaller models or narrowly defined tasks, such as classification, code generation, or work in low-resource languages, but these gains have not generalized to large, general-purpose language models. Where synthetic-only training has achieved performance comparable to real data, it has typically involved substantially smaller models that are not directly comparable to current state-of-the-art systems. For example, the SYNTHLLM family of models, trained entirely on synthetic data, achieves strong results yet still lags behind leading models on major benchmarks (Figure 1.1.15).

Model	GSM8K	MATH	Minerva	Olympiad	College	Gaokao	Average
Llama-3.2-1B-Instruct	47.2	28.0	5.9	5.5	18.8	25.2	21.8
SynthLLM-1B (3.2M)	45.7	32.9	6.6	7.6	27.3	28.8	24.8
SynthLLM-1B (7.4M)	**50.4**	**37.4**	**8.1**	**8.3**	**31.7**	**30.6**	**27.4**
Llama-3.2-3B-Instruct	78.0	47.5	17.3	14.8	31.8	37.7	37.9
SynthLLM-3B (3.2M)	78.1	54.3	16.9	17.5	38.3	46.5	41.9
SynthLLM-3B (7.4M)	**80.7**	**60.0**	**18.8**	**21.9**	**42.3**	**50.9**	**45.8**
Llama-3.1-8B-Instruct	84.2	48.9	25.7	13.2	32.1	43.4	41.2
Llama-3.1-70B-Instruct	**94.5**	66.1	**34.2**	29.6	41.4	56.6	54.2
NuminaMath-CoT-7B	75.4	55.2	19.1	19.9	36.9	47.5	42.3
NuminaMath-CoT-72B	90.8	66.7	25.0	32.6	39.7	54.0	51.5
MAmmoTH2-Plus-8B (10M)	78.4	41.9	10.7	11.3	16.1	31.9	31.7
JiuZhang3.0-8B (6M)	88.7	51.2	21.7	18.8	37.5	43.4	43.6
NaturalReasoning-8B (2.8M)*	-	55.6	-	-	-	-	-
OpenMathInstruct-2-8B (14M)	91.1	67.5	22.5	27.7	39.2	53.5	50.3
SynthLLM-8B (3.2M)	88.4	66.1	25.4	30.2	44.3	56.9	51.9
SynthLLM-8B (7.4M)	92.1	**71.3**	26.5	**33.0**	**45.3**	**61.0**	**54.9**

Source: Qin et al., 2025

Figure 1.1.15

10 Pre-training refers to the initial phase of model development in which a model is trained (typically via self-supervised learning) on large, general-purpose datasets to acquire broad linguistic or multimodal representations. Post-training refers to subsequent refinement of the base model, through techniques such as supervised fine-tuning or reinforcement learning, to specialize behavior, improve alignment, or optimize performance on particular tasks.

HIGHLIGHT:

Data-centric Methods

Discussions on data availability often overlook an important shift in recent AI research. Performance gains are increasingly driven by improving the quality of existing datasets, not by acquiring more. Rather than scaling data indiscriminately, researchers are spending more effort in pruning, curating, and refining training inputs. Data-centric methods emphasize performance improvements through practices such as cleaning labels, deduplicating samples, and constructing higher-quality datasets. A growing body of research shows that training models on low-quality or polluted data can significantly degrade performance. Likewise, recent evidence illustrates that data pruning, selecting the most informative training inputs, often outperforms approaches that train on all available data indiscriminately.

Recent large-scale model development illustrates this paradigm in practice. Olmo 3 researchers prioritized large-scale deduplication, quality-aware data selection, and stage-specific training curricula rather than indiscriminate data scaling. These interventions, combined with iterative feedback loops to evaluate and refine candidate data mixes, allowed their models to achieve competitive performance despite training on substantially fewer tokens than other leading state-of-the-art models (Figure 1.1.16). Olmo 3.1's Think 32B model, for example, contains roughly 32 billion parameters, nearly 90 times fewer than Grok 4's 3 trillion, yet it achieves comparable performance on several benchmarks, including American Invitational Mathematics Examination (AIME)[11] 2025.

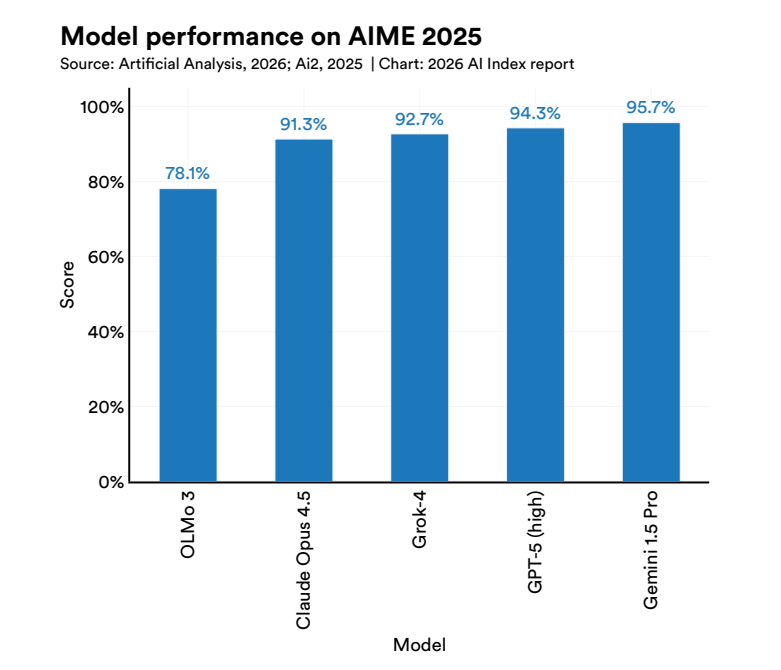


Figure 1.1.16

Synthetic Data in Post-training

Recent research shows that synthetically generated data can be effective for improving model performance in post-training settings, including fine-tuning, alignment, instruction tuning, and reinforcement learning. A growing body of research released in 2025 supports this finding. Evidence suggests that synthetic post-training data is effective in few-shot generation settings, for improving long-context capabilities, for optimizing reinforcement learning workflows, and for strengthening reasoning more broadly.

Prevalence of Synthetic Content

Since the launch of ChatGPT in November 2022, there have been predictions that the internet would soon become overrun by AI-generated content. Recent research from Graphite suggests that beginning in January 2025, over 50% of newly published online content was generated by AI (Figure 1.1.17). Others have projected that the share in 2026 could be even higher.

11 The American Invitational Mathematics Examination (AIME) is an annual high school math competition widely used as a benchmark for AI mathematical reasoning, with each year's exam providing a fresh test set.

HIGHLIGHT:

AI-generated content vs. human content
Source: Graphite.io, 2025 | Chart: 2026 AI Index report

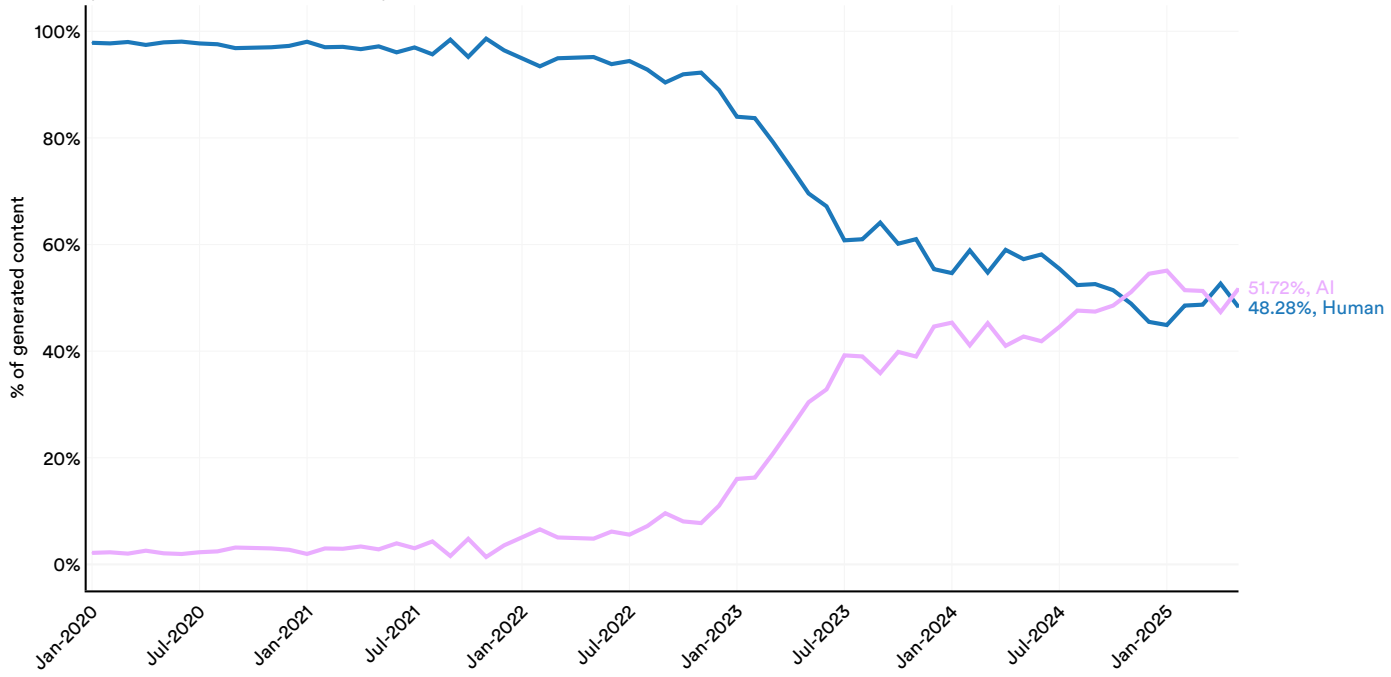


Figure 1.1.17

Given growing concerns about the suitability of synthetic data for training AI systems, this trend raises questions about the long-term reliability of current scaling trajectories. In response, many firms that depend on high-quality training data have increasingly turned to proprietary sources. In May 2025, the New York Times entered into a licensing agreement with Amazon to allow its content to be used for training purposes. By mid-2025, Meta was reportedly engaged in similar discussions with news organizations, while health and life sciences companies such as Bristol Myers Squibb have pursued comparable strategies. These developments suggest that firms training frontier AI systems are adjusting their data acquisition strategies as the volume of openly available training data continues to decline.

1.2 Compute and Infrastructure

The development of AI models requires significant infrastructure investment. As training processes have expanded in scale and complexity, the underlying hardware has also improved in both speed and efficiency. In turn, these gains shape what kinds of models researchers and labs can realistically build. The growth in training compute discussed in the previous section would not have been possible without corresponding improvements in hardware capabilities. This section leverages data from Epoch AI to track hardware performance, adoption, and aggregate computing capacity over time.

Performance and Efficiency

Peak computational performance of machine learning hardware has increased exponentially across releases between 2008 and 2025 (Figure 1.2.1). The gains are especially visible at lower precision types, where precision refers to the number of bits used to represent numerical values. Lower precision formats such as FP16 and Tensor-FP16/BF16 now show the highest performance levels and have become standard in many training and inference settings.

Peak computational performance of ML hardware for different precisions, 2008–25

Source: Epoch AI, 2026 | Chart: 2026 AI Index report

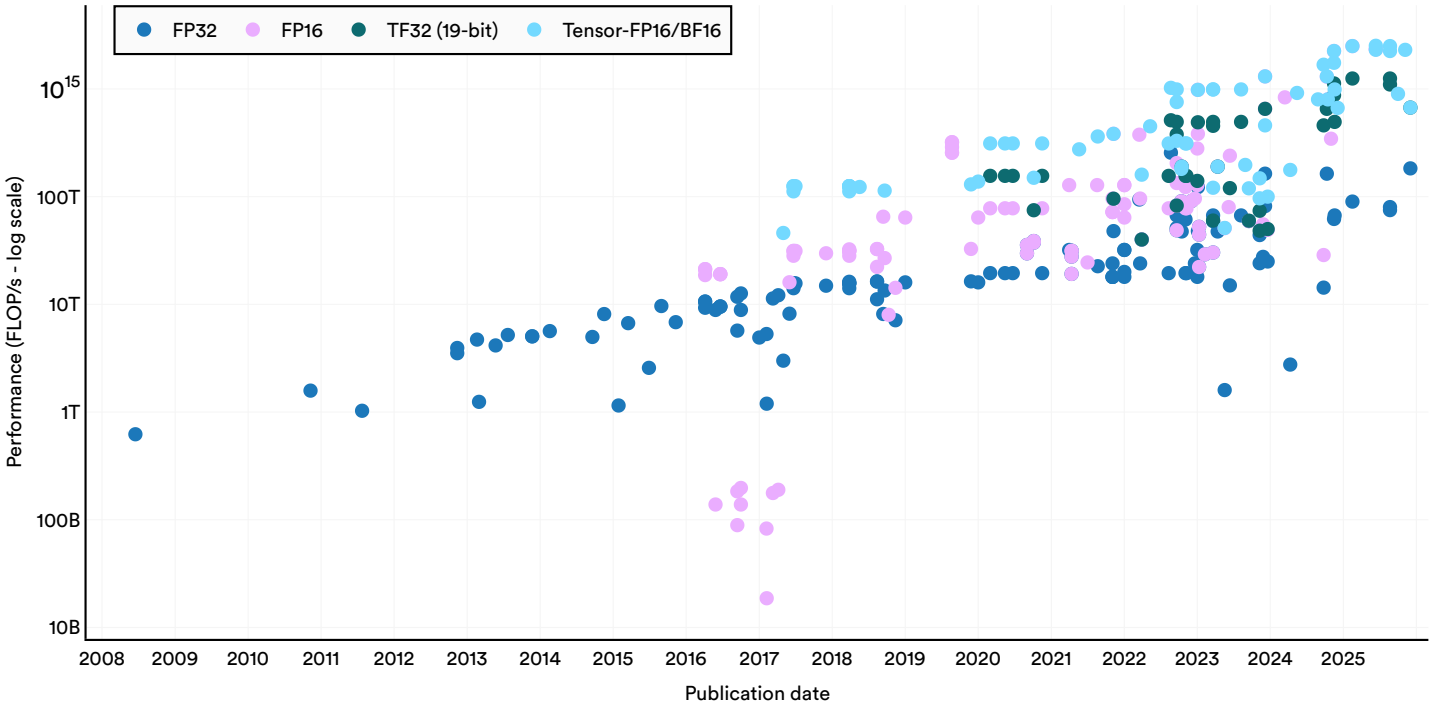


Figure 1.2.1

Hardware for Notable Models

Hardware adoption patterns among notable AI models reflect the gains in performance and efficiency (Figure 1.2.2). Since 2017, the cumulative number of notable models trained on A100-class hardware has increased, with 84 models trained in 2025. The previous generation, V100, continues to power a sizable share (69 models). Newer hardware, such as the H100, has seen early rapid adoption (28), while other categories, such as TPU v3 and TPU v4, show stable curves.

Cumulative number of notable AI models trained by accelerator, 2017–25

Source: Epoch AI, 2026 | Chart: 2026 AI Index report

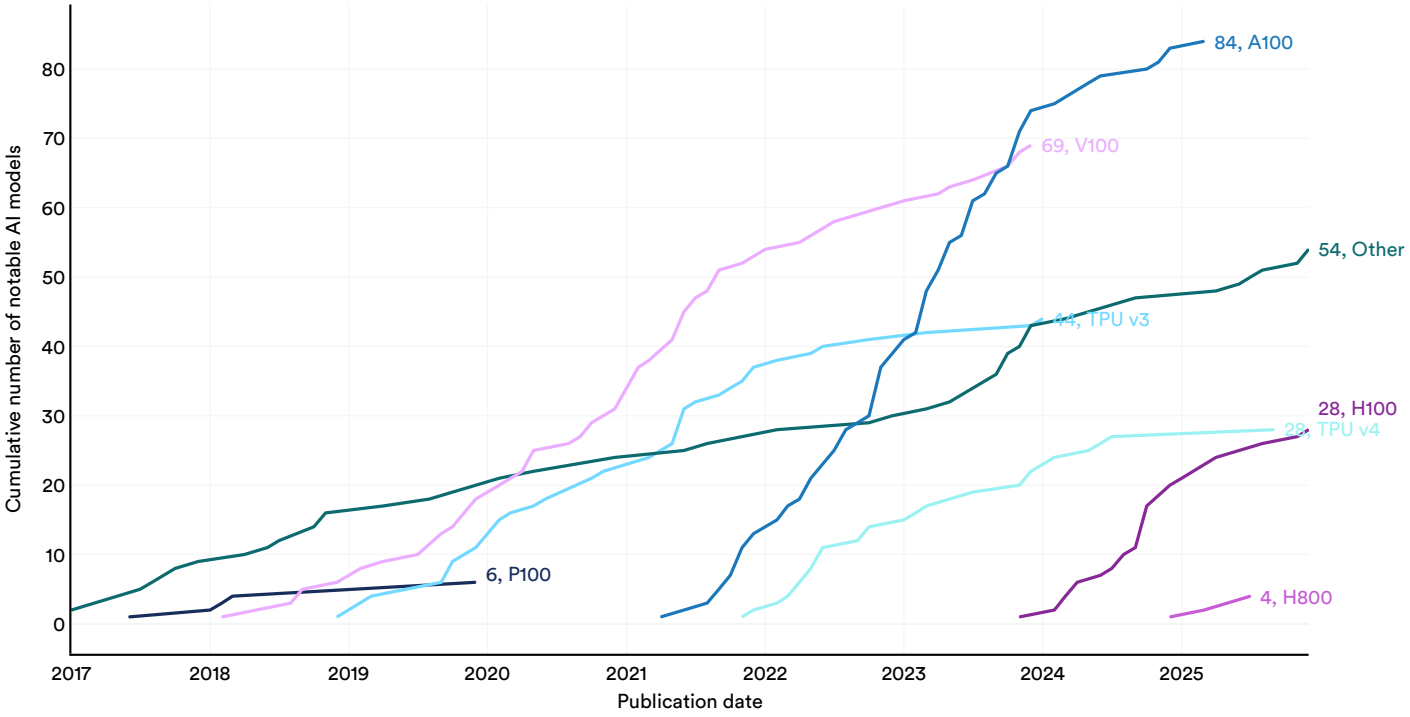


Figure 1.2.2

Global Computing Capacity

The supply of AI computing capacity from major chip designers has continued to increase (Figure 1.2.3). Total capacity has increased by an estimated 3.3x per year since 2022, reaching approximately 17.1 million H100-equivalents.[12] Nvidia AI chips currently account for over 60% of total compute, with Google and Amazon supplying much of the remainder and Huawei holding a small but growing share. The growth in compute capacity tracks closely with investment patterns described in Chapter 4, where leading AI companies have increased their capital expenditure and infrastructure has become the fastest growing focus area of private AI funding.

12 Since these estimates are inferred from revenue data, financial disclosures and analyst reports, they reflect broader trends rather than exact counts. Data coverage also varies by manufacturer; Nvidia and Google data starts in 2022, while others start in 2024.

Global computing capacity from AI chips across major designers, 2022–25
Source: Epoch AI, 2026 | Chart: 2026 AI Index report

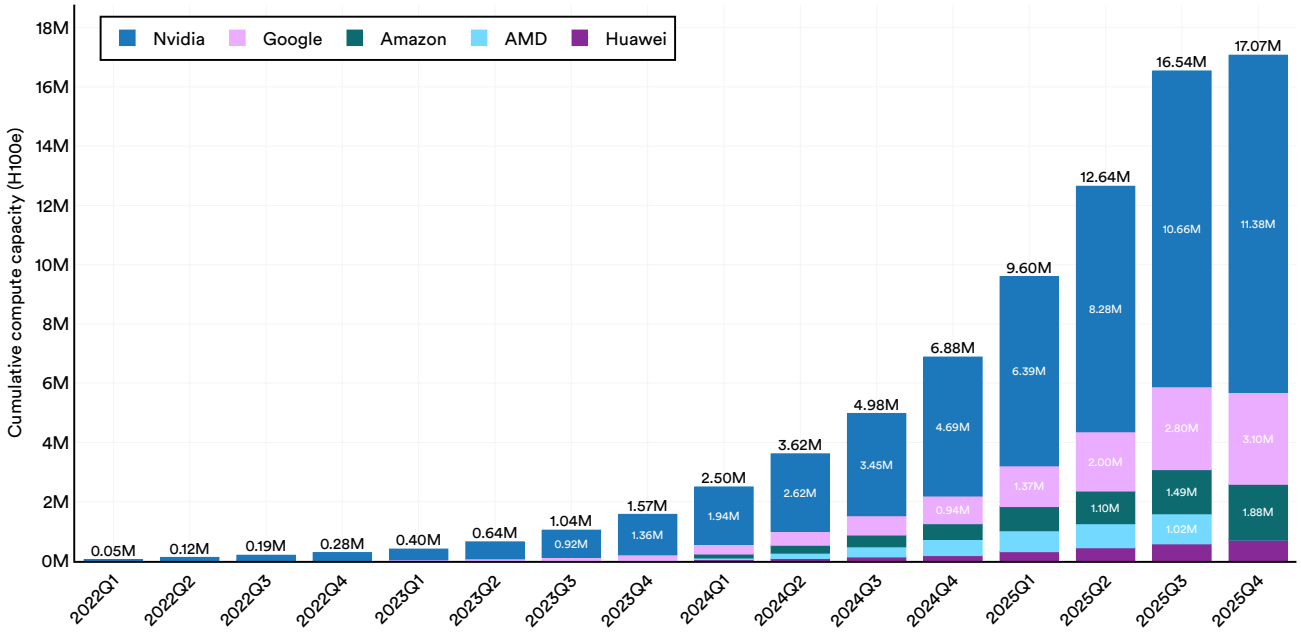


Figure 1.2.3

Data Center Power Capacity

The expansion of computing capacity carries a direct energy cost. Total AI data center power capacity reached approximately 29.6 GW by Q4 2025, enough to power all of New York state at peak demand (Figure 1.2.4). AI chip power, measured by thermal design power, accounted for roughly 11.8 GW of the total, with the remainder attributed to cooling, networking, and other data center infrastructure. This estimate is based on the rated power capacity of leading AI chips sold over time, with a multiplier of approximately 2.5 applied to account for the additional requirements of powering infrastructure.

Global AI data center power capacity, 2022–25
Source: Epoch AI, 2026 | Chart: 2026 AI Index report

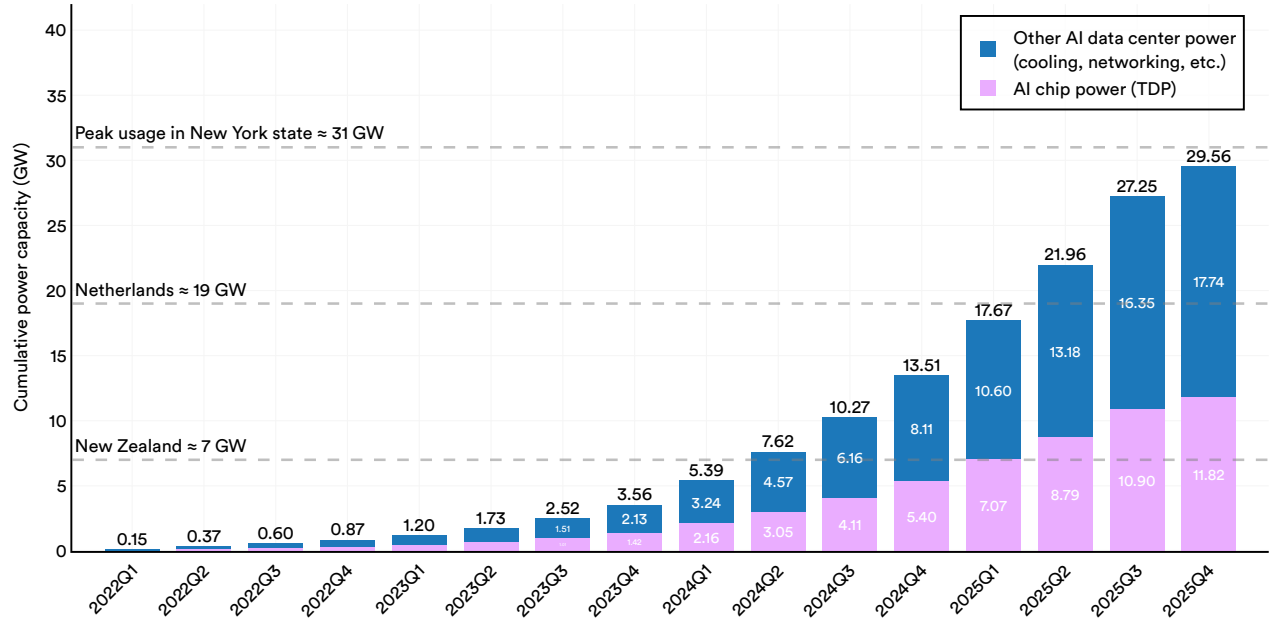


Figure 1.2.4

1.3 Data Centers

The physical infrastructure underlying AI development extends beyond models and compute described in the previous section. Data centers are where compute is housed, and their capacity, geographic distribution, and underlying supply chains shape what AI systems can be built and where. This section draws on data from Cloudscene to track the global distribution of data centers and introduces an overview of the broader AI infrastructure ecosystem to provide context for the geographic and supply chain dynamics.

AI Infrastructure: Beyond GPUs

Modern AI data centers depend on a combination of compute, storage, communications, and specialized hardware that enables AI systems to run at large scale. GPUs and custom accelerators such as Tensor Processing Units (TPUs) are the most widely discussed, but they are only one layer of a broader infrastructure stack. All data processed by these chips is held in high-bandwidth memory (HBM), which supports moving large volumes of data in and out efficiently. The leading manufacturers of HBM are SK Hynix (South Korea), Samsung (South Korea), and Micron (USA). During training, GPUs must continuously share data with one another, which requires fast, high throughput network connectivity achieved with fiber-optic cables running high-bandwidth networking architectures such as InfiniBand.

The supply chain behind this hardware adds another dimension. Companies like Nvidia and SK Hynix design but do not manufacture chips. Instead, they provide designs to specialized semiconductor foundries, primarily the Taiwan Semiconductor Manufacturing Company (TSMC) and Samsung Foundry, which fabricate the chips at the nanometer scales modern AI hardware requires. The fabricated chips are then packaged and tested by assembly companies such as ASE Group (Taiwan) and Amkor Technology (United States). TSMC is a single point of dependency in the global AI supply chain, as it fabricates virtually every leading AI chip, including Nvidia's Blackwell GPUs and AMD's MI300X. There are high barriers to entry at every layer—requiring decades of accumulated expertise, specialized equipment, and significant capital investment to overcome.

The infrastructure ecosystem is relevant beyond AI capabilities, as it shapes education priorities and workforce development. Chapter 7 (Education) distinguishes between AI software-related and AI hardware-related degrees. That distinction is also relevant here, where different countries play different roles across the hardware supply chain.

Geographic Distribution

Most of the world's data center infrastructure is located in a small number of countries (Figures 1.3.1 and 1.3.2). In 2025, the United States led by a wide margin, with 5,427 data centers, more than 10 times the count of any other country. Germany (529), the United Kingdom (523), and China (449) followed, while the majority of the remaining countries each had fewer than 300 facilities. The U.S. may show a clear lead, but the other country rankings should be assessed with the understanding that data center counts do not capture differences in facility size, computing capacity, or utilization.

Global distribution of data centers, 2025

Source: Cloudscene, 2025 | Chart: 2026 AI Index report

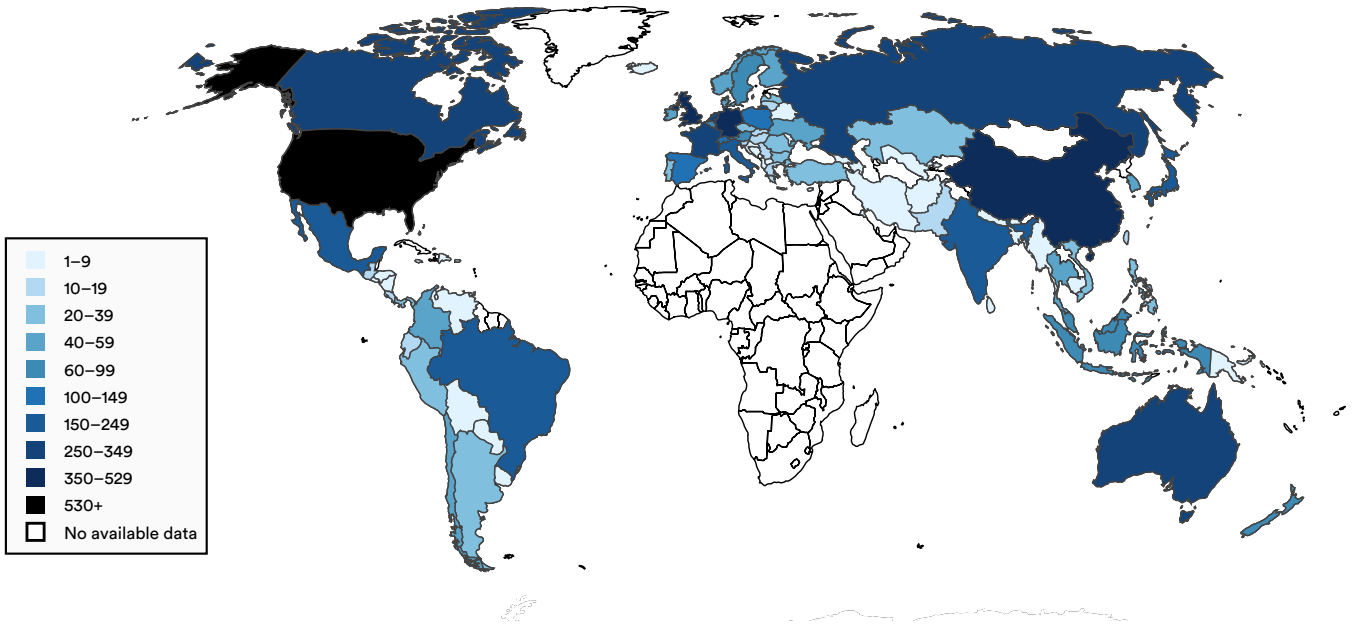


Figure 1.3.1

Number of data centers by geographic area, 2025

Source: Cloudscene, 2025 | Chart: 2026 AI Index report

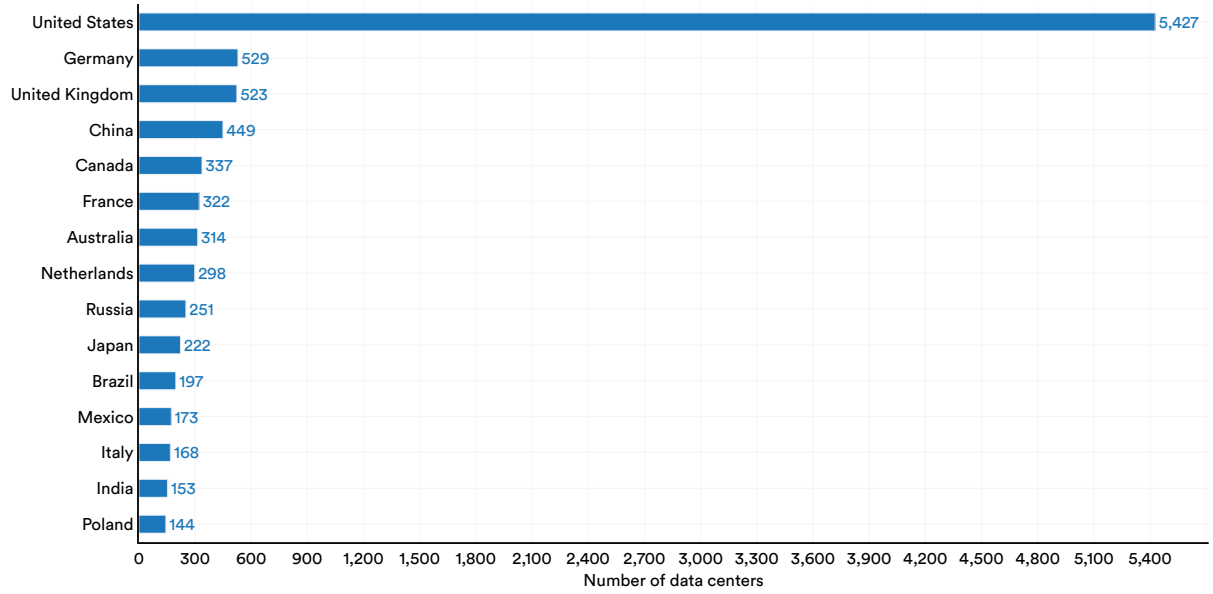


Figure 1.3.2

1.4 Energy and Environmental Impact

As AI systems have scaled and become more widely deployed, their energy consumption and environmental footprint have become very visible. The compute and infrastructure trends described in the preceding sections translate into heavy demands on energy, water, and carbon emissions. This section examines those costs across three areas of AI development: training, inference, and data center energy usage. The analysis draws from Epoch AI's model-level data, recent academic benchmarking research (Jegham et al., 2025), the International Energy Agency's reporting on data centers (IEA, 2025), and de Vries and Gao (2025).

Training

Leading machine learning hardware has grown more efficient since 2016, as measured in FLOP/s per watt (Figure 1.4.1). Leading chips deliver about 10 times more computation per watt than those available a decade ago, with Nvidia B200 and Google TPU v5e among the most efficient. However, models have scaled faster than efficiency has improved, so total power required to train frontier systems has continued to increase. Total power draw for training models has grown by several orders of magnitude since the early 2010s (Figure 1.4.2). The most compute-intensive models in the data set, such as Grok 3 and Llama 4 Behemoth, required upward of 100 million watts during training. Due to limited disclosure by their developers, power draw information is not available for many of the newest models that have been released.

Carbon emissions from training have increased even more sharply (Figure 1.4.3). Training AlexNet in 2012 produced an estimated 0.01 tons of CO_2 equivalent, while training Grok 4 in 2025 produced about 72,816 tons. To put this into context, that is more than the lifetime carbon emissions of an average car (63 tons). Larger models generally produce more emissions but not always, as it can also depend on hardware efficiency, training duration, and the carbon intensity of the energy sources used. DeepSeek v3, for example, produced approximately 597 tons, which is much less than models of comparable size (Figure 1.4.4).

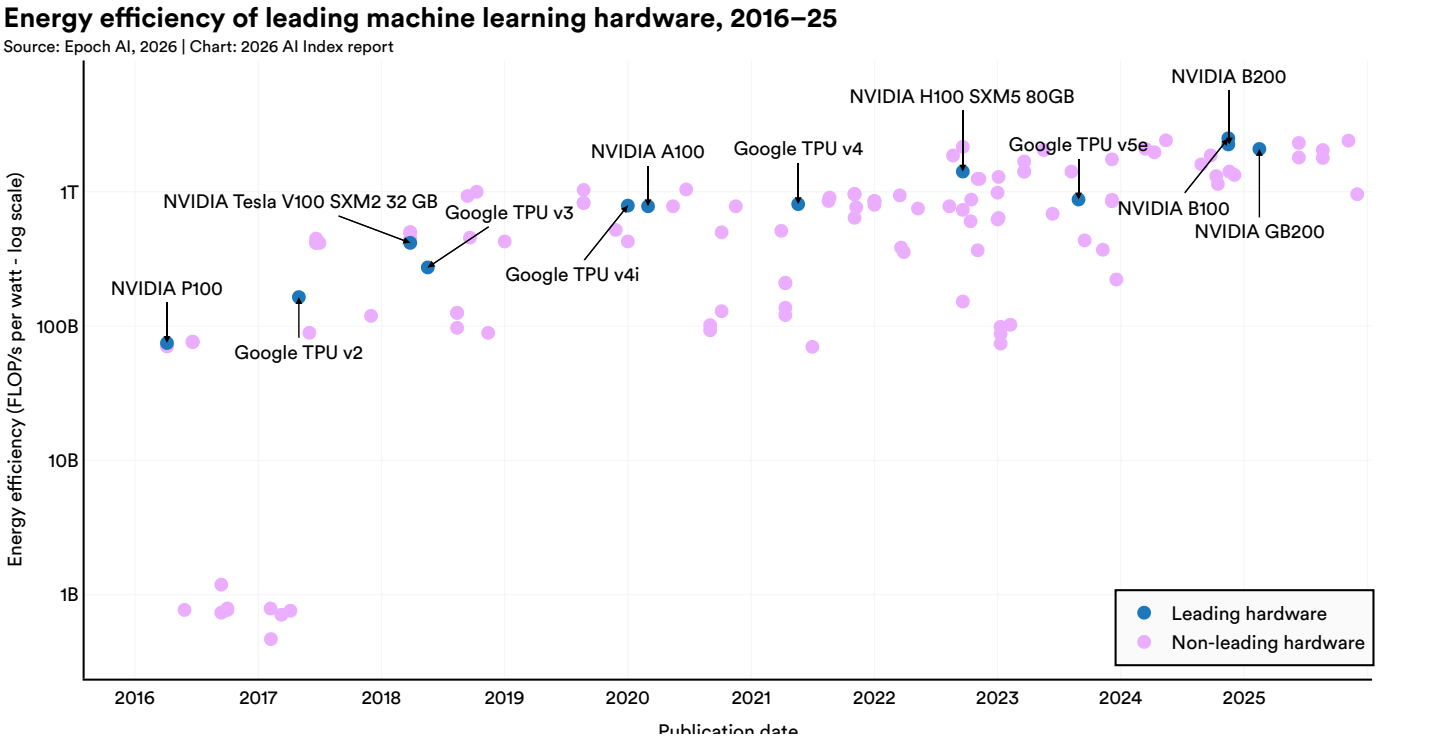


Figure 1.4.1

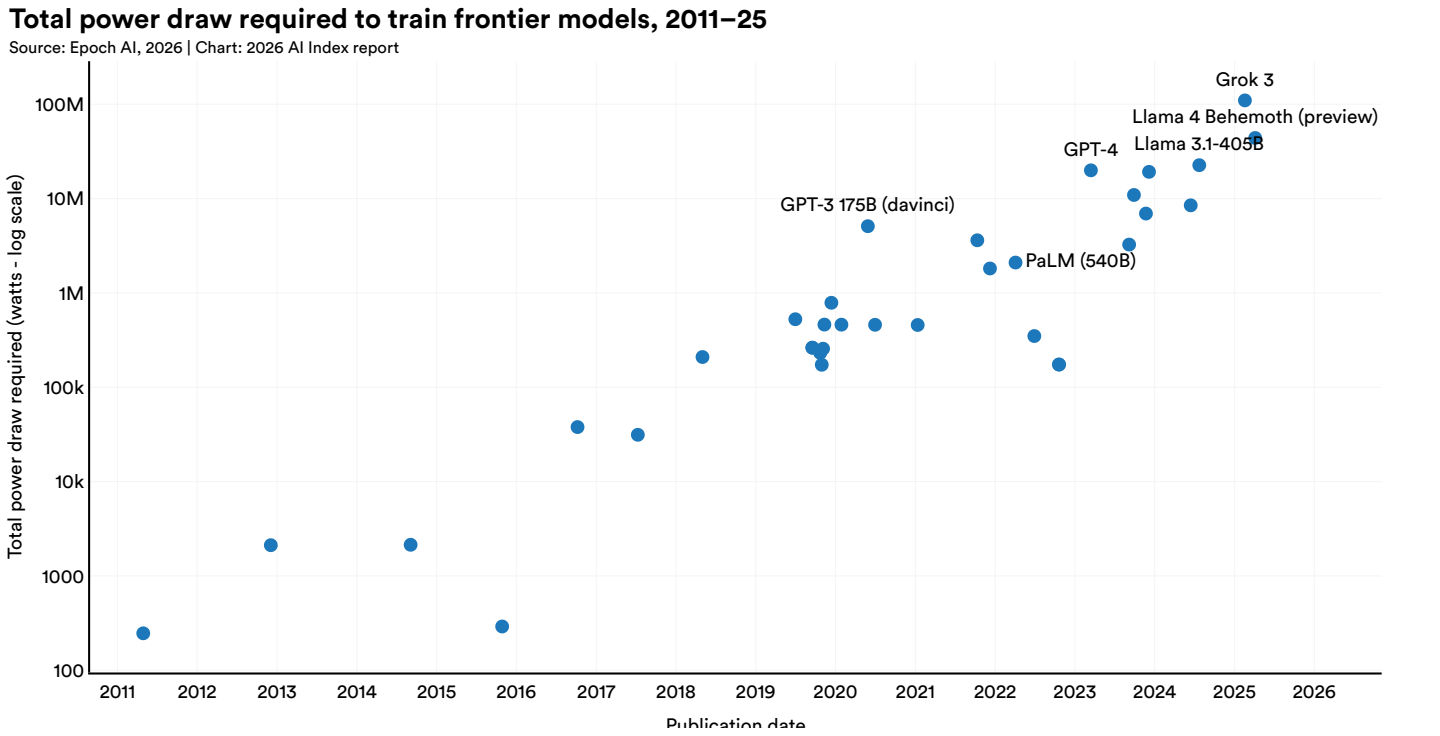


Figure 1.4.2

Estimated carbon emissions from training select AI models and real-life activities, 2012–25

Source: AI Index, 2026; Strubell et al., 2019 | Chart: 2026 AI Index report

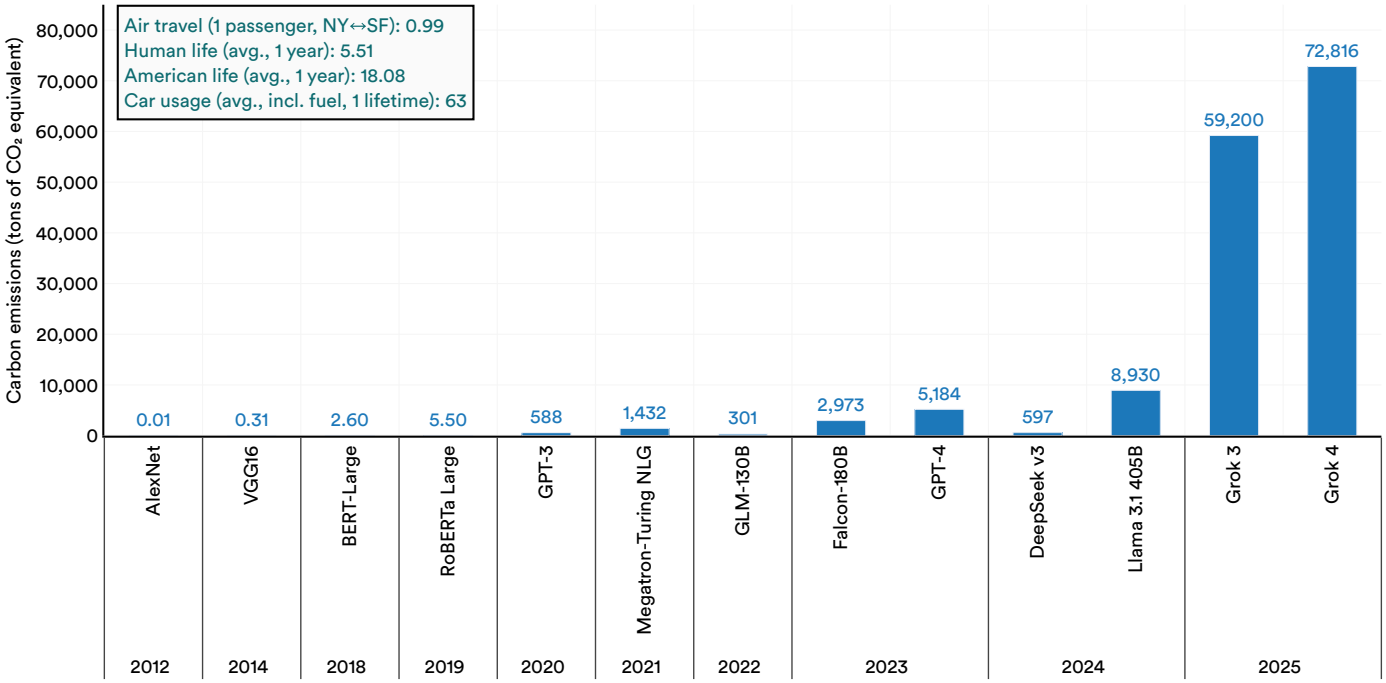


Figure 1.4.3

Estimated carbon emissions and number of parameters by select AI models

Source: AI Index, 2026 | Chart: 2026 AI Index report

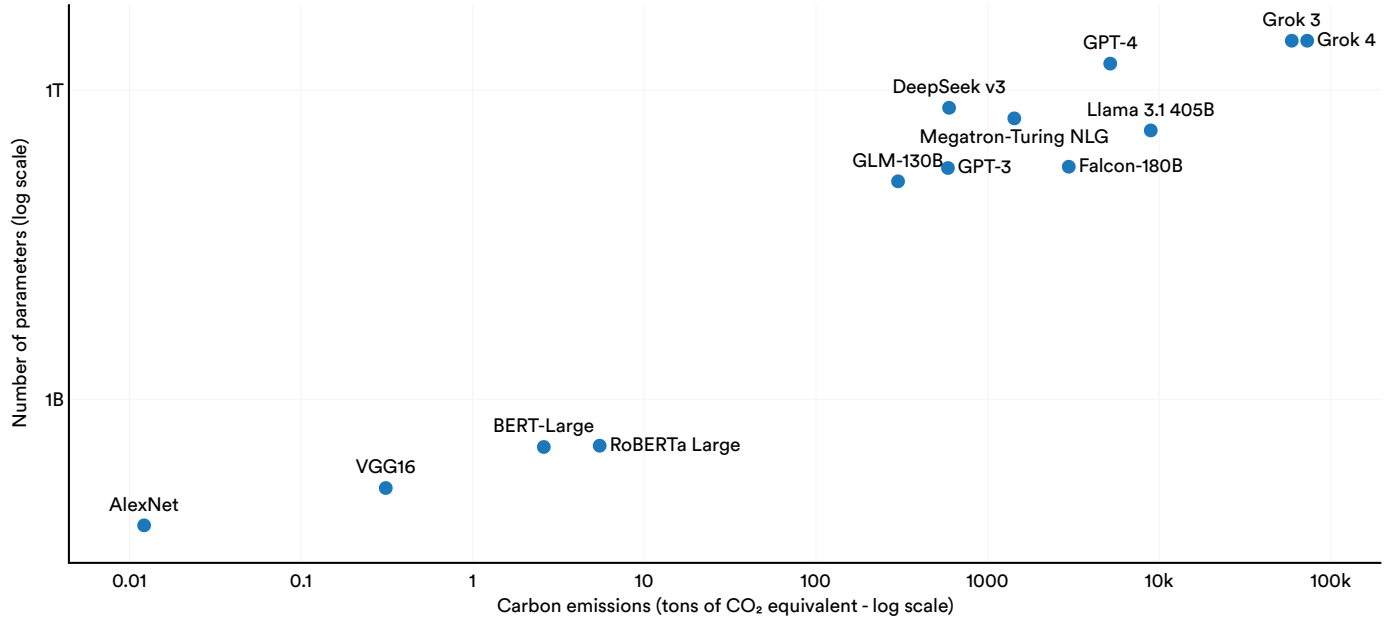


Figure 1.4.4

Inference

Training costs have typically received the most attention, but inference represents a growing share of AI's total energy footprint. Once a model is deployed at scale, the cumulative energy required to serve queries can exceed the one-time cost of training within months.

Recent benchmarking by Jegham et al. (2025) provides per model estimates of inference energy consumption and carbon emissions for medium-length prompts (defined as approximately 1,000 input tokens and 1,000 output tokens). Among the top 15 models by energy consumption in 2025, DeepSeek V3.2 Exp and DeepSeek V3.2 consumed the most per query (23 Wh), followed by GPT-5 (high) at 21.9 Wh (Figure 1.4.5). Models such as Claude 4 Opus and GPT-5 min (medium) sit at the lower end, consuming between 5 and 6 Wh. When ranked by carbon emissions, the models also follow a similar pattern (Figure 1.4.6). DeepSeek V3.2 Exp and DeepSeek V3.2 produced the highest per medium-length prompt, approximately 14 grams of CO2 equivalent each. For comparison, Claude 4 Opus and Mistral Medium 3 were the lowest at 1.6 and 1.5 grams, respectively. There is a wide spread even among models released in the same year, showing not only that inference efficiency varies but that higher capability is not necessarily proportional to the environmental cost.

Model energy consumption for medium-length prompts

Source: Jegham et al., 2025 | Chart: 2026 AI Index report

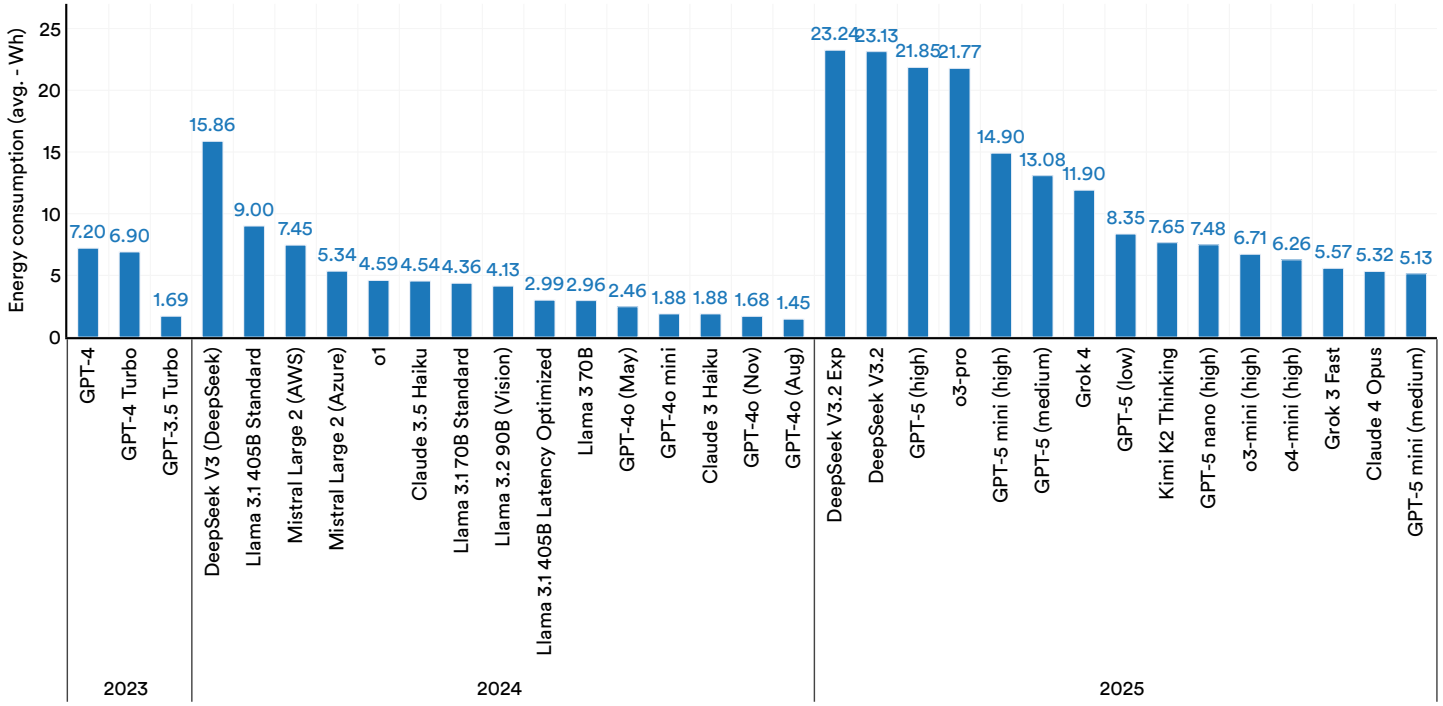


Figure 1.4.5[13]

13 This figure shows the top 15 models by energy consumption for 2024 and 2025. The full set of models is available through the source dashboard.

Model carbon emissions for medium-length prompts
Source: Jegham et al., 2025 | Chart: 2026 AI Index report

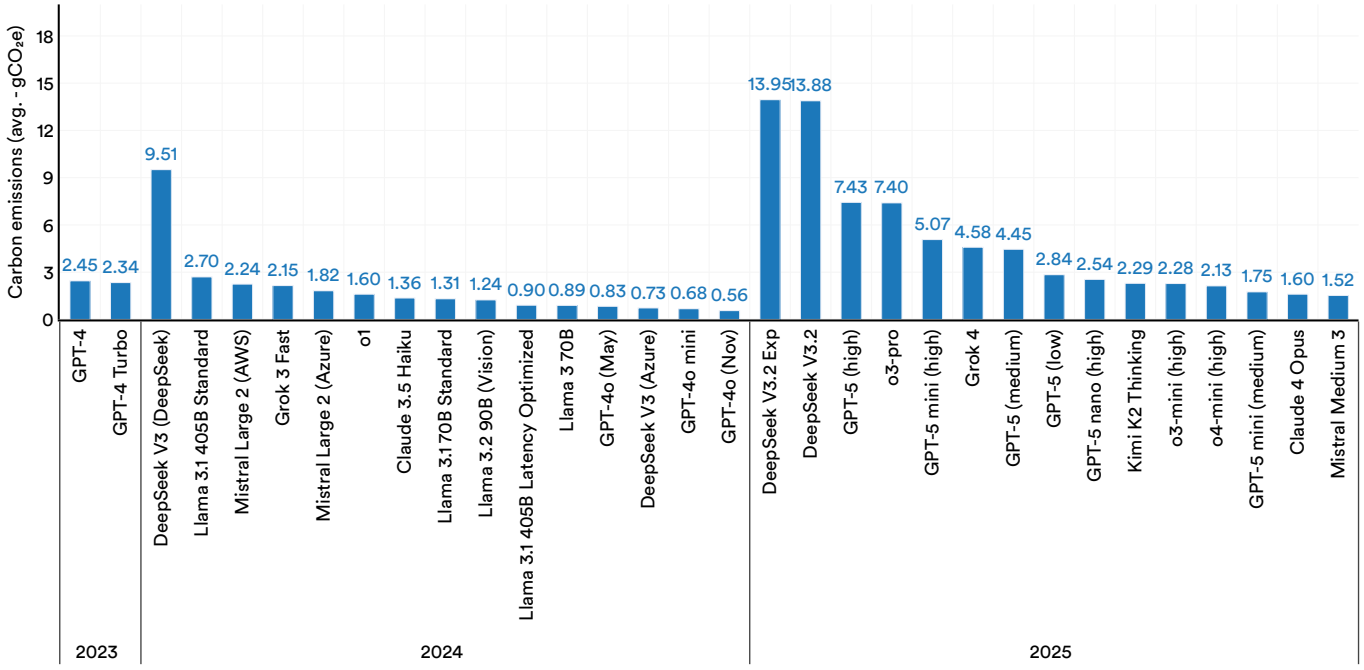


Figure 1.4.6[14]

At the level of a single query, the numbers seem more modest. A short GPT-4o query consumes approximately 0.42 Wh, which is 40% more than a Google search at 0.3 Wh (Figure 1.4.7). A daily session of eight medium-length queries uses the energy comparable to charging two smartphones (9.7 Wh). But across hundreds of millions of daily queries, the consumption scales into something much larger.

The same scaling dynamic is true for water consumption (Figure 1.4.8). Annual estimates for GPT-4o inference range from about 1.3 to 1.6 kiloliters, which, at the high end, exceeds the annual drinking water needs of 12 million people.

14 This figure shows the top 15 models by energy consumption for 2024 and 2025. The full set of models is available through the source dashboard.

Per-query and daily energy consumption: GPT-4o vs. common activities

Source: AI Index, 2025 | Chart: 2026 AI Index report

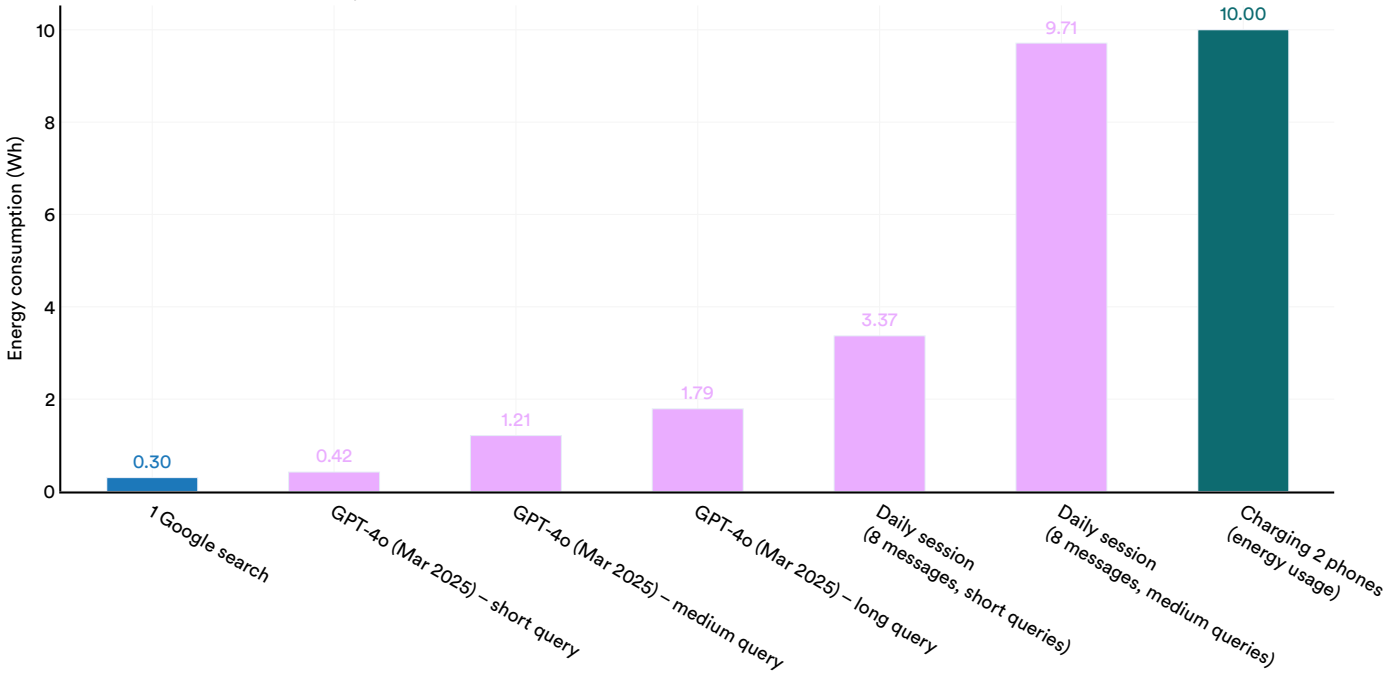


Figure 1.4.7

Annual water consumption: GPT-4o vs. real-world baselines

Source: AI Index, 2025 | Chart: 2026 AI Index report

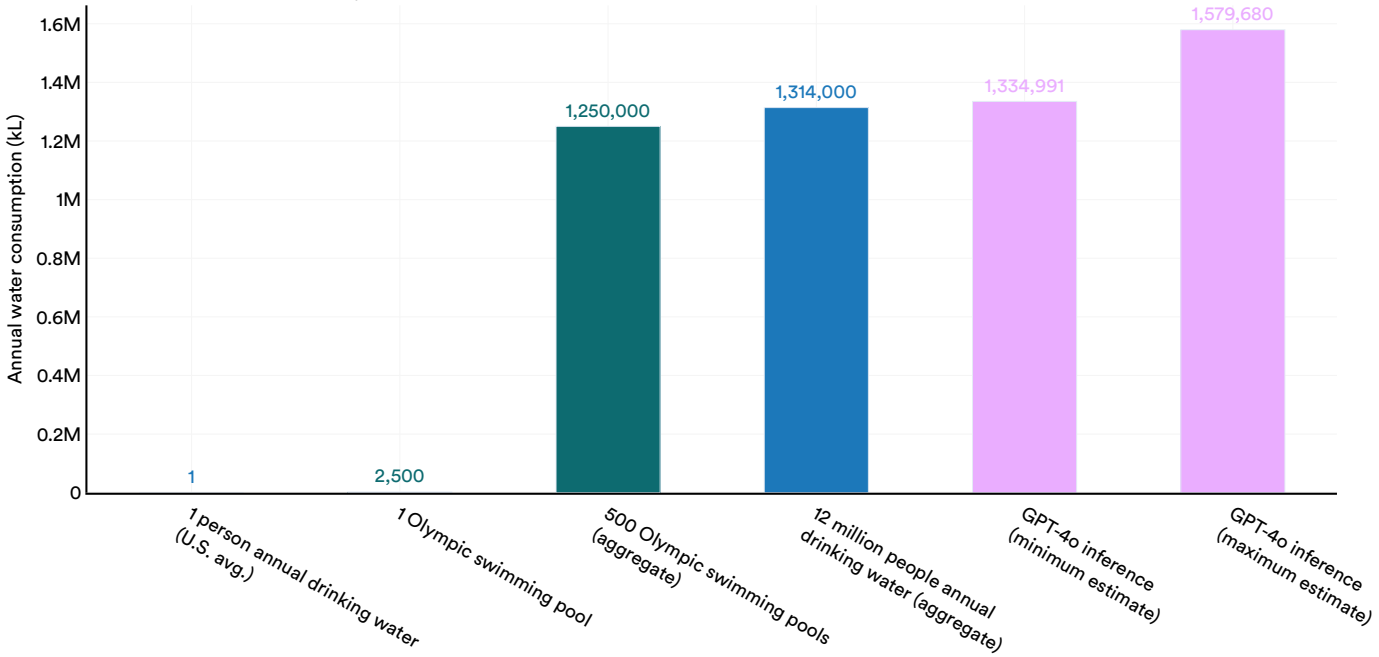


Figure 1.4.8

Data Center Usage

The power demands of models and queries add up to a much larger infrastructure footprint. The estimated power demand from AI accelerator modules reached approximately 5,200 MW cumulatively through 2024 (Figure 1.4.9). Nvidia accounted for the largest share, which is consistent with the company's leading position in global AI chip capacity (as discussed in Section 1.2). When including the full systems supporting those accelerators (servers, cooling, networking), estimated demand reached approximately 9,400 MW (Figure 1.4.10). However, these figures from de Vries and Gao (2025) carry uncertainty from variation in utilization rates and facility-level efficiency, as reflected in the error bars on the charts.

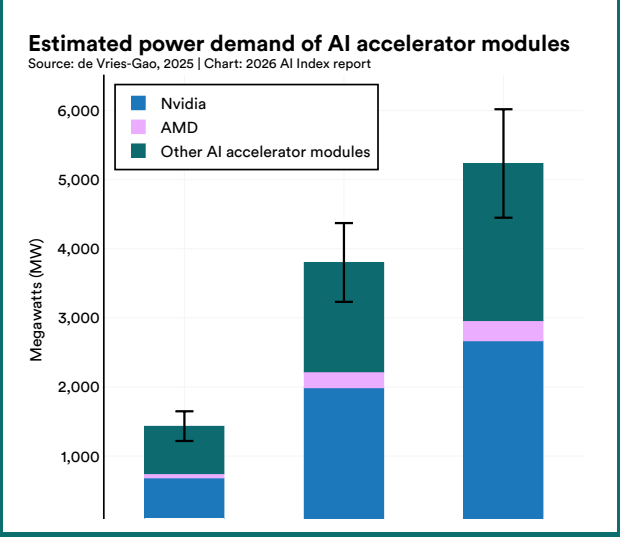


Figure 1.4.9

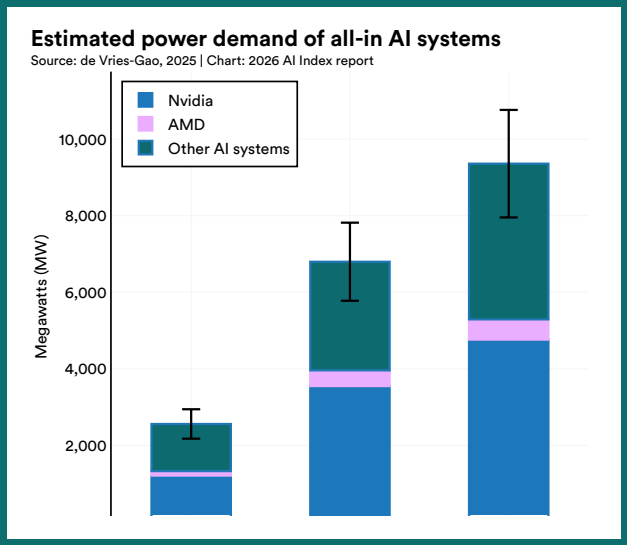


Figure 1.4.10

To put that scale in perspective, the cumulative power demand of all-in AI systems is comparable to the national electricity consumption of Switzerland or Austria, and roughly half that of Bitcoin mining (Figure 1.4.11). Excluding crypto, global data centers accounted for the highest estimated power demand at around 47,000 MW, with AI hardware making up a growing share of that total.

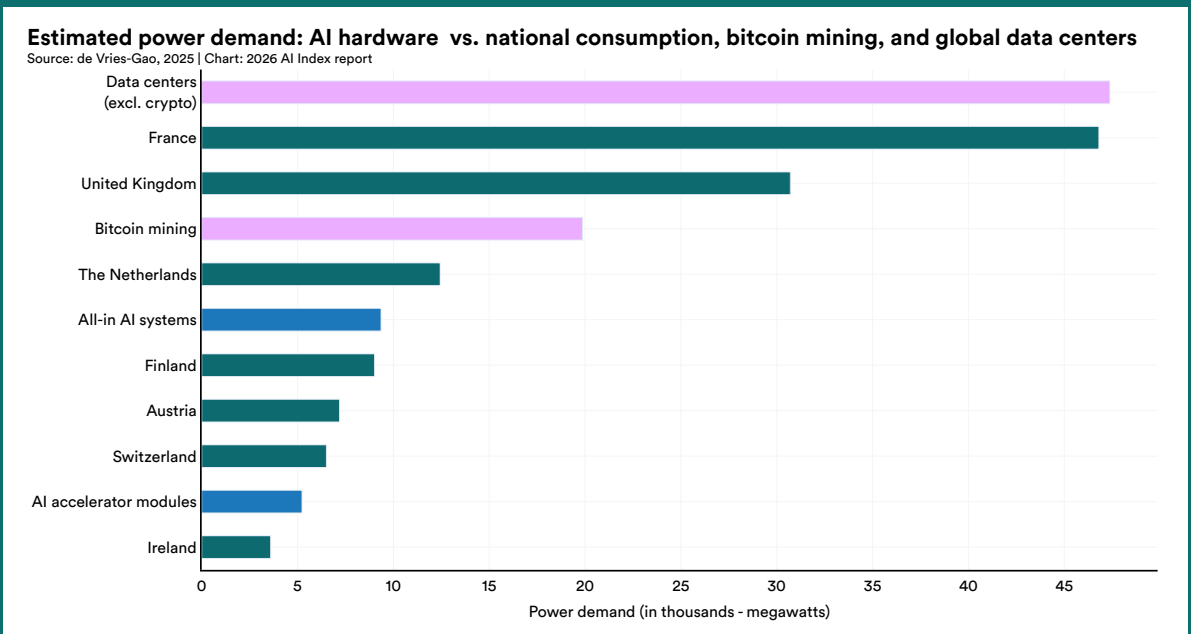


Figure 1.4.11

Cost, however, has been moving in the opposite direction. Since 2006, the cost of GPU computation has fallen by more than 99% (Figure 1.4.12). This decline has been key to enabling the scaling trends described throughout this chapter, making it economically feasible to train and deploy models at levels that would have been cost prohibitive even a decade ago. At the regional level, data center electricity consumption has increased across all major regions, and it is projected to continue to rise through 2030 (Figure 1.4.13). The United States accounts for the largest share, followed by China, Europe, and the rest of Asia.

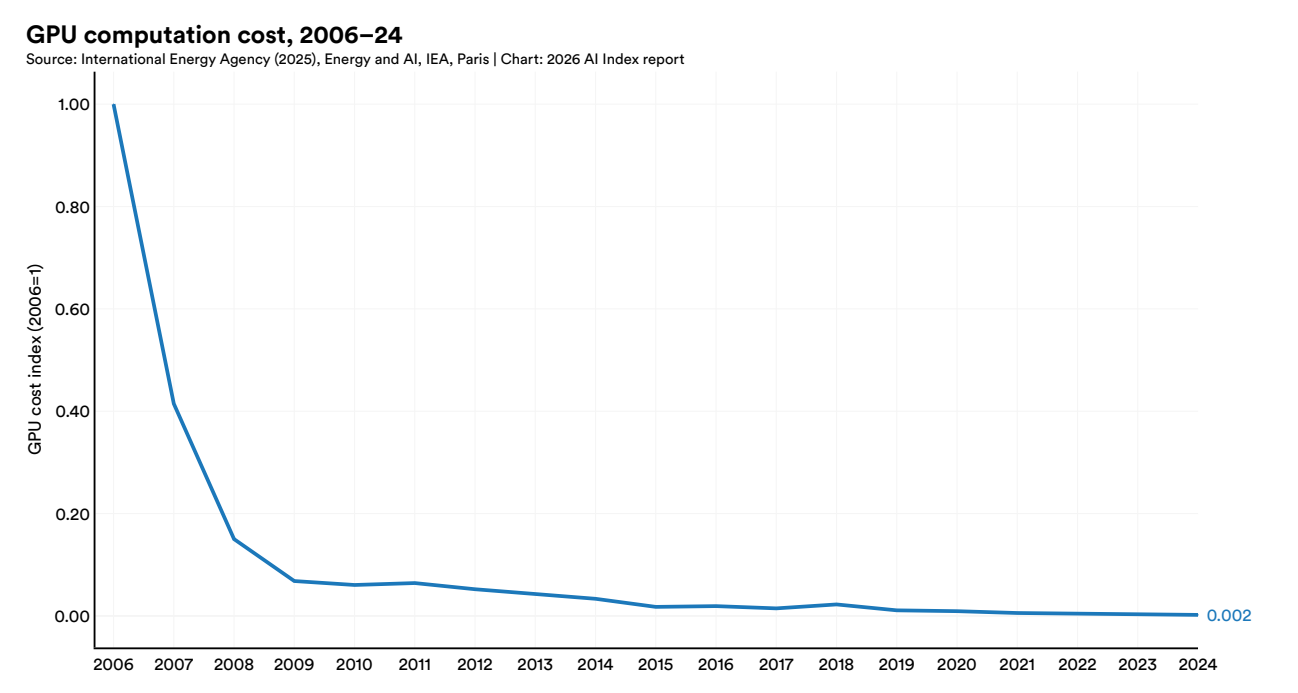


Figure 1.4.12

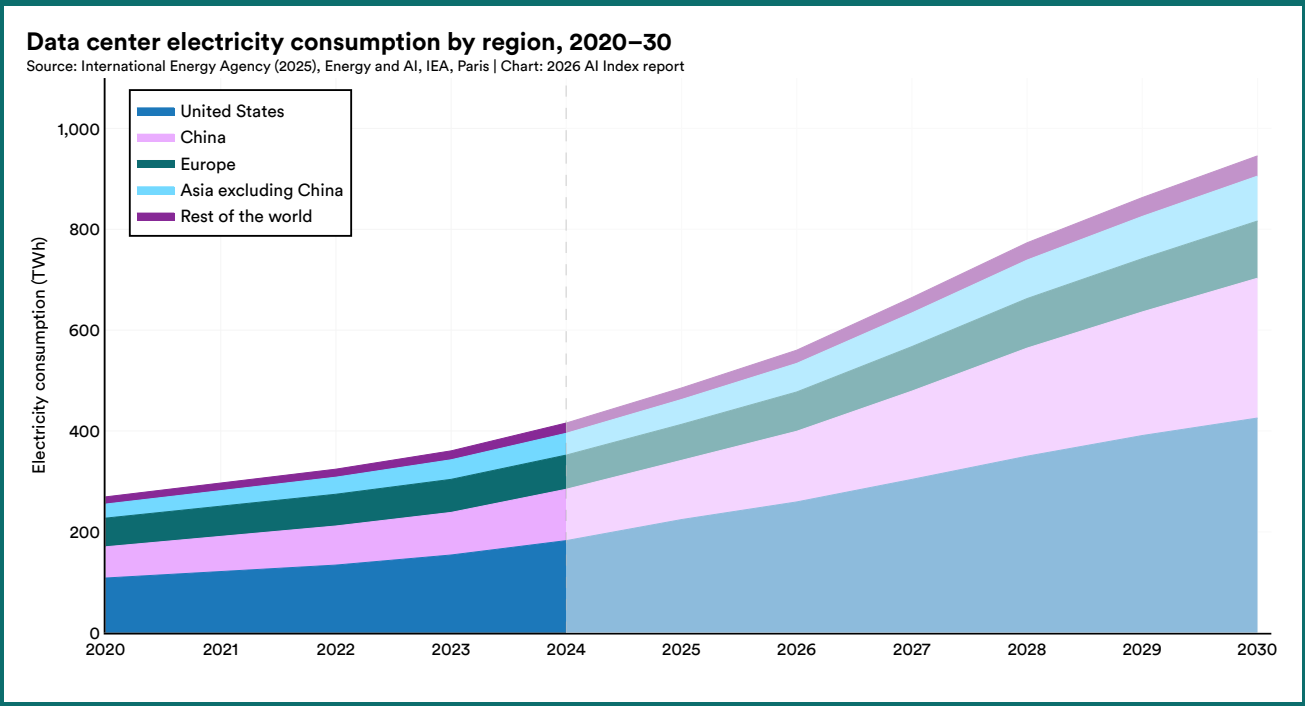


Figure 1.4.13[15]

15 Data in this chart reflects IEA projections rather than observed consumption.

1.5 Open-Source AI Software

AI Development Activity Overview

The preceding sections have focused on notable frontier models and the infrastructure required to build and maintain them. Open-source platforms like GitHub and Hugging Face offer a different view that captures the developer ecosystem experimenting with and building on AI models. Much of this activity is not reflected in academic publications or frontier model releases. The AI Index analyzes data from both platforms[16] to better understand how open-source AI development is evolving over time.

Projects

The scale of open-source development has grown steadily. The number of AI-related GitHub projects increased from 1,549 in 2011 to approximately 5.6 million in 2025, with year-over-year growth accelerating 23.7% from 2024 (Figure 1.5.1). However, most repositories often consist of personal or experimental work and receive minimal attention. When filtering for projects with at least 10 stars, a rough proxy for community engagement, the count drops to 206,880 in 2025 (Figure 1.5.2). The growth trajectory is similar for both measures.

Number of GitHub AI projects, 2011–25

Source: GitHub, 2025 | Chart: 2026 AI Index report

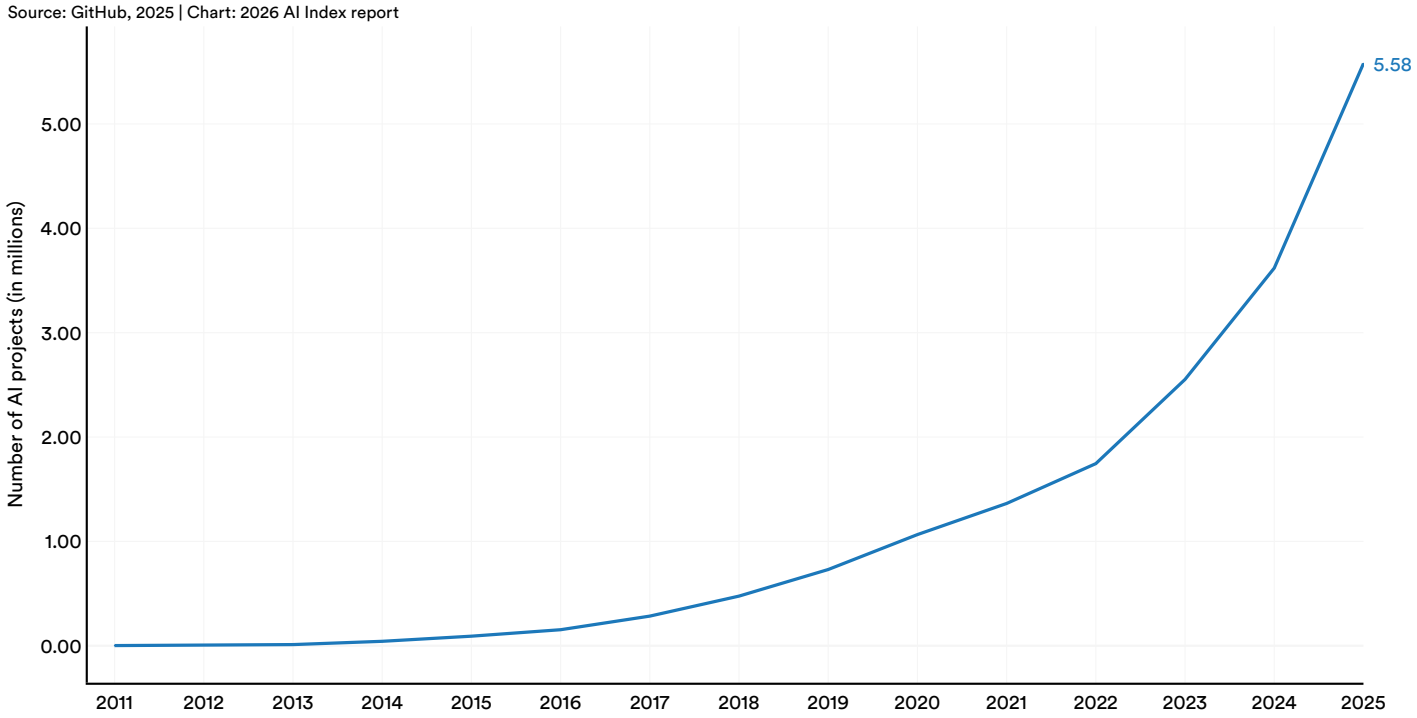


Figure 1.5.1

16 Chinese researchers often use alternatives to GitHub, such as Gitee and GitCode, for code sharing, but the data from those sites is not included in this report. A full methodological description is available in the Appendix.

Number of GitHub AI projects with at least 10 stars, 2011–25

Source: GitHub, 2025 | Chart: 2026 AI Index report

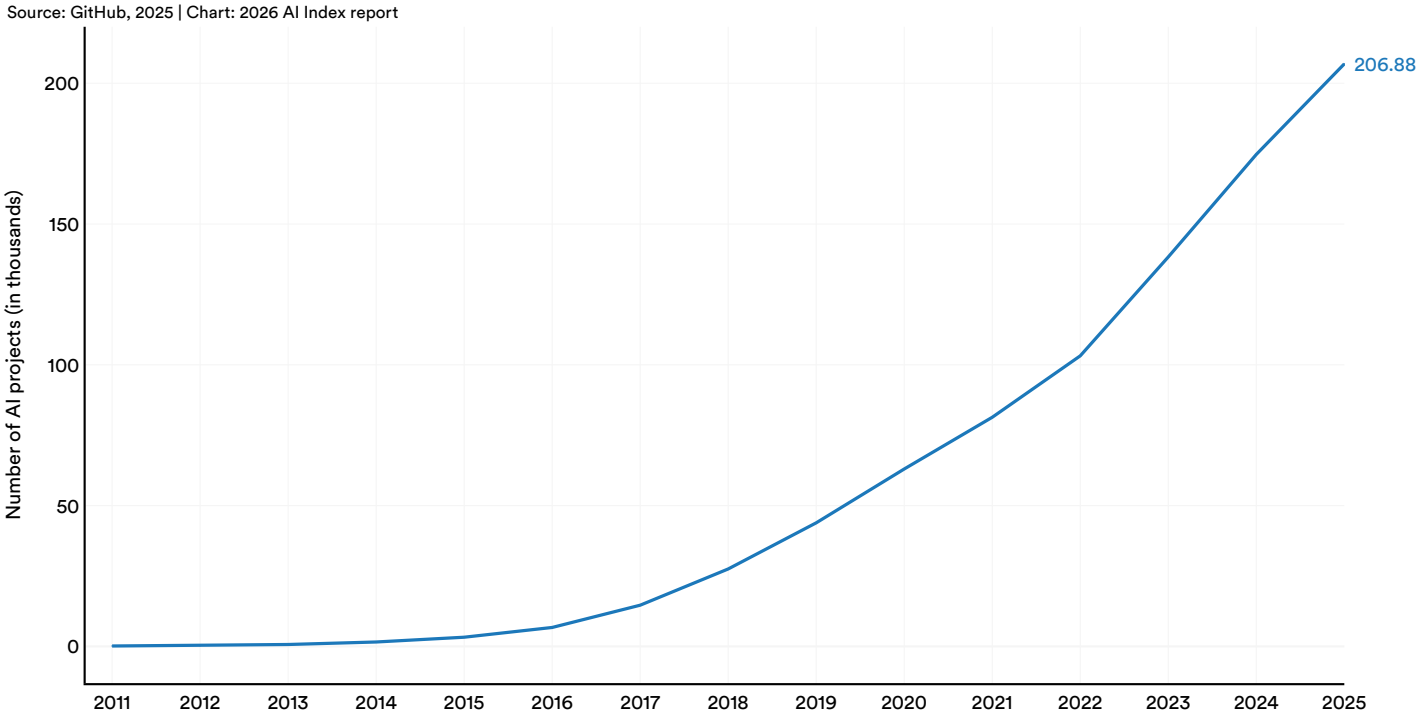


Figure 1.5.2

The geographic distribution of more visible open-source AI projects has shifted over time (Figure 1.5.3). Among projects with at least 10 stars, the United States accounted for the largest share in 2025 (31.7%), though that has declined steadily from nearly 80% in 2011 as developers in other regions have increased their presence on the platform. Europe and the rest of the world have grown in number of projects, while China's share has leveled off since 2019. India remains a growing contributor, representing 5.2% of projects with at least 10 stars. Because GitHub data does not capture Chinese developers who use domestic platforms such as Gitee or GitCode, China's share of global open-source AI activity is likely understated. The existing geographic attribution for China uses self-reported location rather than IP-based geolocation.

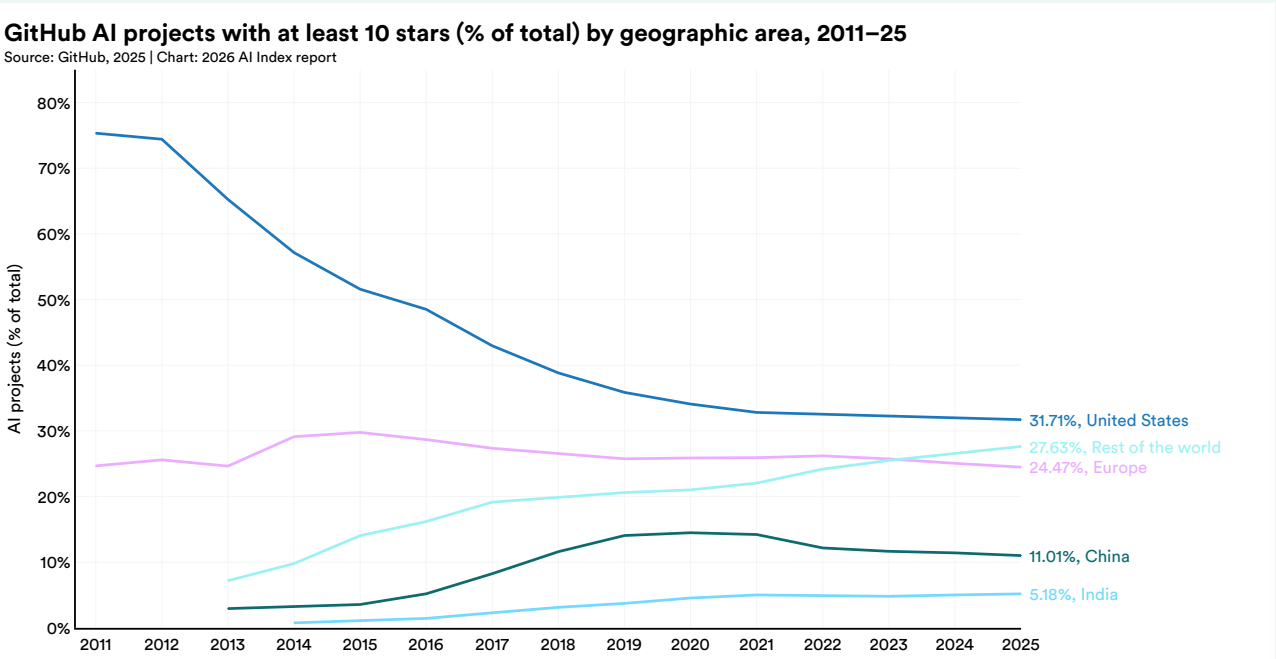


Figure 1.5.3[17]

Stars

Beyond project counts, GitHub stars provide another signal of developer interest and engagement in open-source communities (Figure 1.5.4). The total number of stars for AI projects increased from 14 million in 2023 to 18.2 million in 2025.[18] All major geographic regions saw year-over-year increases. However, the geographic pattern for stars differs from the project share data above. Despite its declining share of projects, the United States accumulated the highest number of stars at 30 million cumulatively (Figure 1.5.5). So while open-source activity becomes more geographically distributed, the projects with the most engagement remain disproportionately U.S.-based.

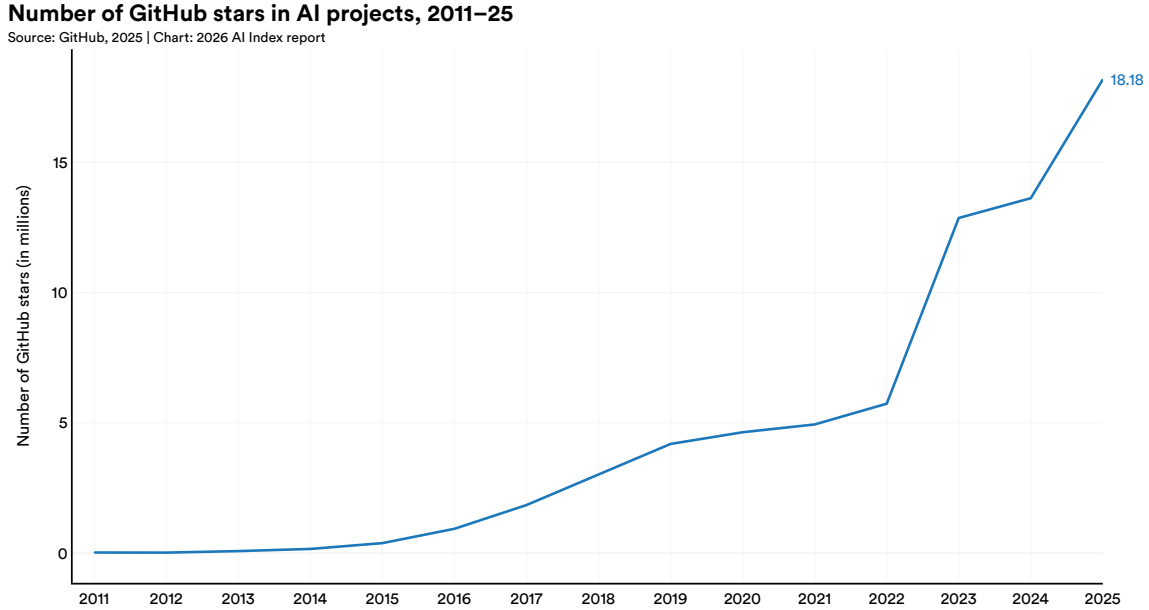


Figure 1.5.4

17 In previous AI Index reports, project locations for China and Hong Kong were determined using IP-based geolocation. Due to frequent VPN usage in these regions that resulted in systematic misclassification, self-reported profile locations are now used for China and Hong Kong, while IP-based geolocation continues to be applied for all other countries.

18 Figure 1.5.4 shows new stars given to GitHub projects within a year, not the total accumulated over time.

Number of GitHub stars by geographic area, 2011–25
Source: GitHub, 2025 | Chart: 2026 AI Index report

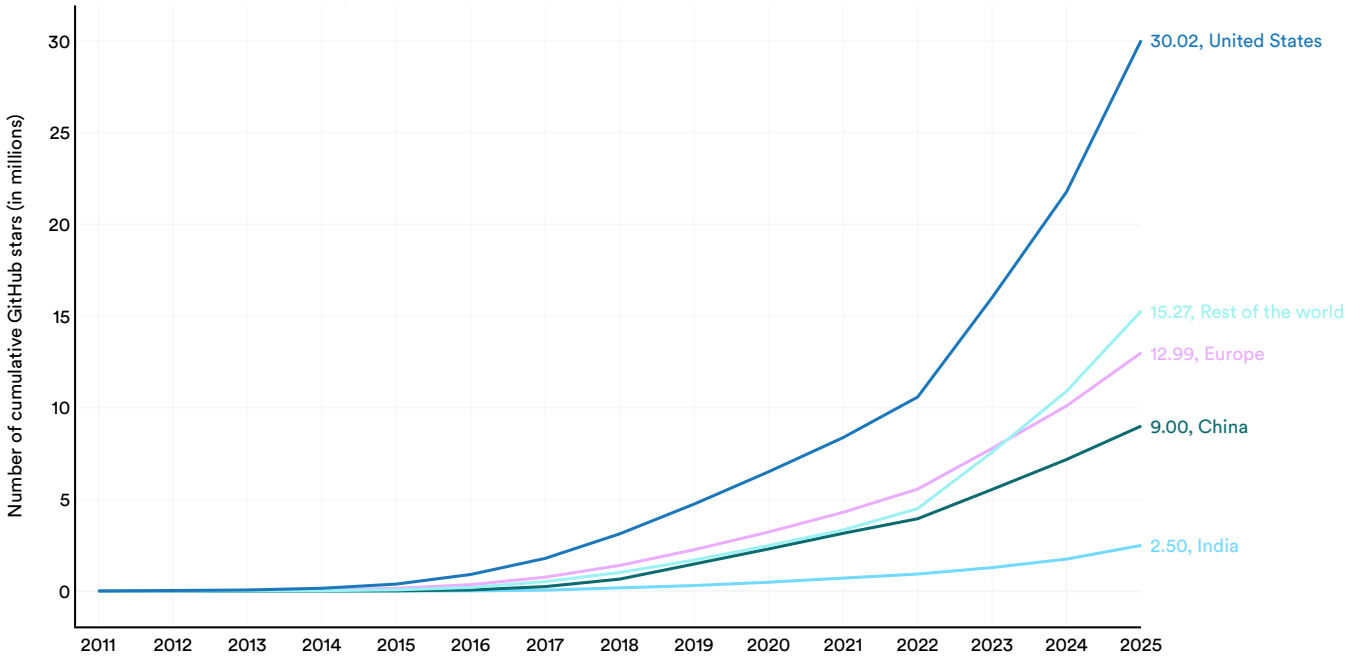


Figure 1.5.5

Model and Dataset Ecosystem

To complement the GitHub view, this section uses metadata from Hugging Face, a widely used community platform and open repository for AI models and datasets. The analysis focuses on assets created or uploaded between 2022 and 2025 to understand recent activity and adoption trends (Figures 1.5.6 and 1.5.7). Upload activity has continued to rise over the last few years, with a marked increase after the second quarter of 2024. From 2023 to 2025, model uploads more than tripled, while dataset uploads grew fourfold. Download distribution also shifted after 2023. Geographically[19] U.S.-developed models lost share to unaffiliated users. On the developer side, major private actors such as Google and Meta have shifted from being the principal authors to accounting for a relatively small share of downloads, while communities such as Sentence Transformers and the BERT community have grown (Figure 1.5.8). A large share of total model downloads fell into an "Others" category, reflecting the wider distribution of development activity even as the most downloaded models were tied to a small number of sources.

19 Data was obtained in collaboration with researchers from Longpre et al. (2025). Their dataset provides Hugging Face model download data that the authors describe as consistent and relatively complete. It was validated with the Hugging Face team, is reported to be less noisy than raw counts, and includes cleaned and imputed missing metadata. It is released as a weekly panel rather than an all-time-downloads cross-section. Coverage spans March 2020 to August 2025 and includes the top 200 most-downloaded Hugging Face models per week. These models account for 49.6% of total normalized, filtered downloads. This restriction focuses the analysis on models with higher observed download volume, reduces long-tail variation, and may support more stable estimates.

Number of models and datasets on Hugging Face, 2022–25

Source: Hugging Face, 2025 | Chart: 2026 AI Index report

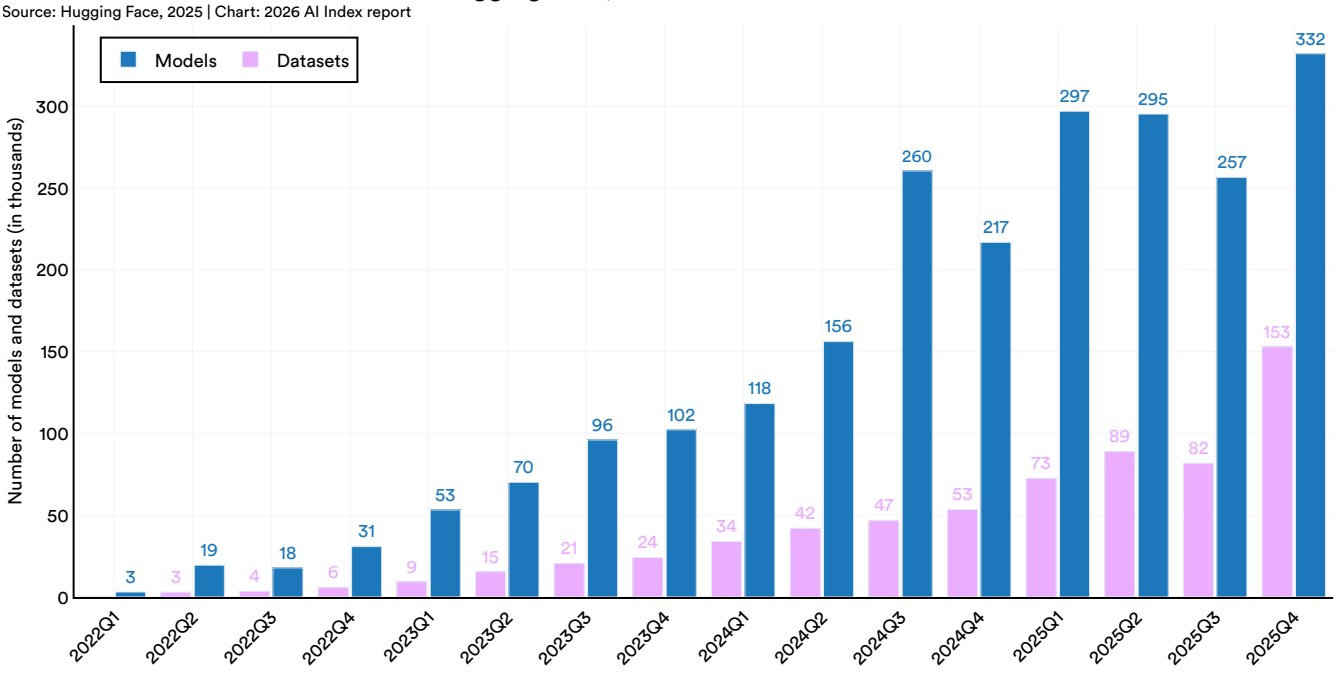


Figure 1.5.6[20]

Global distribution of downloads among top Hugging Face models, Q2 2020–Q3 2025

Source: Longpre et al., 2025; Hugging Face, 2025 | Chart: 2026 AI Index report

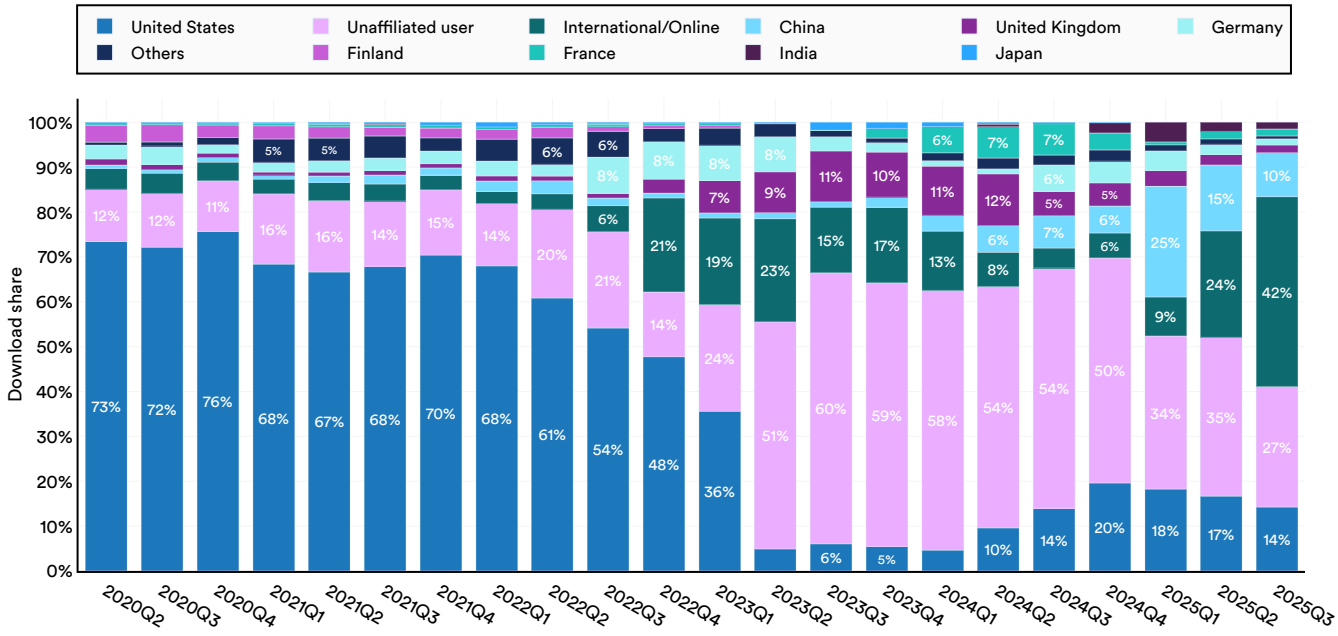


Figure 1.5.7[21]

20 The data shown in this chart comes from the publicly accessible Hugging Face repository. For more details, refer to the Appendix.

21 Data source: Longpre et al. (2025). For more details, refer to the Appendix.

Download share by developer among top Hugging Face models, Q2 2020–Q3 2025
Source: Longpre et al., 2025; Hugging Face, 2025 | Chart: 2026 AI Index report

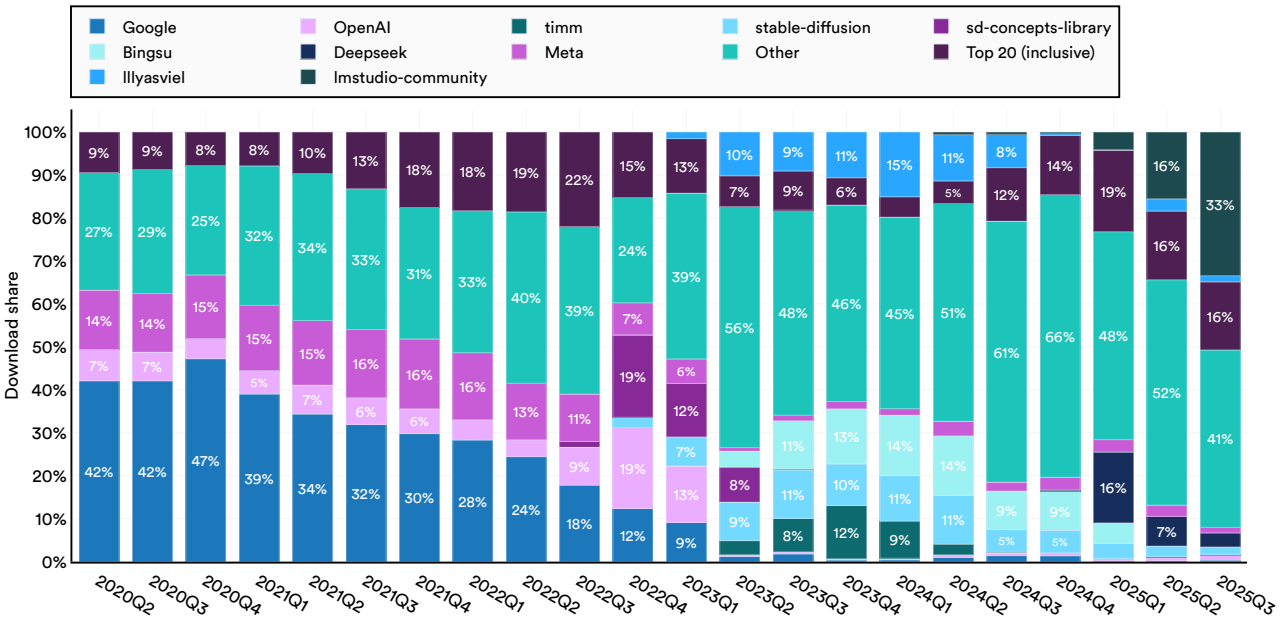


Figure 1.5.8[22]

The most popular model types have shifted over the last three years. Text embedders, classifiers, and audio models, which together accounted for nearly 70% of downloads in 2022, fell to less than 6% in 2025 (Figure 1.5.9). Text generation, multimodal, and video generation models have grown in their place. Text generation led in 2025, accounting for more than 42% of total downloads. Image generation models also increased steadily, remaining the second most downloaded category. Despite these shifts, downloads remain highly concentrated, with nearly 80% associated with the top three categories.

Download share by modality among top Hugging Face models, Q3 2022–Q3 2025
Source: Longpre et al., 2025; Hugging Face, 2025 | Chart: 2026 AI Index report

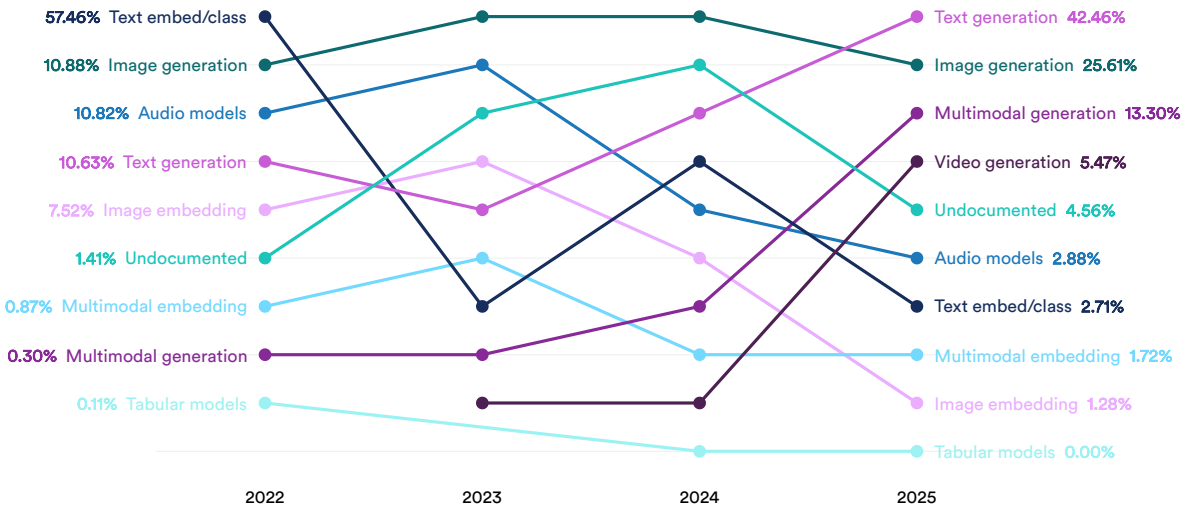


Figure 1.5.9[23]

22 Data source: Longpre et al. (2025). For more details, refer to the Appendix.

23 Data source: Longpre et al. (2025). For more details, refer to the Appendix.

1.6 Publications

The first half of this chapter tracked the models, infrastructure, and energy behind AI development. This section shifts to research output, specifically English-language AI publications and citations. Publications offer a longitudinal signal of AI research activity at scale, and the AI Index has tracked them consistently over time. While publication volume is not a measure of research quality, and not all research appears in indexed databases, this approach offers a consistent method for tracking the research frontier year over year. The analysis draws from OpenAlex, a bibliographic database[24] the AI Index has used since 2025, and considers both publication volume and downstream influence through citation patterns.

Total Number of AI Publications

Total AI publication output continues to rise. AI publications more than doubled between 2013 and 2024, increasing from roughly 102,000 to about 258,000 (Figure 1.6.1). Growth continued in 2024, though at a slower rate, with publications increasing 6.3% from 2023. AI research now makes up a substantial portion of the broader computer science ecosystem, accounting for 40.9% of all computer science publications in OpenAlex.

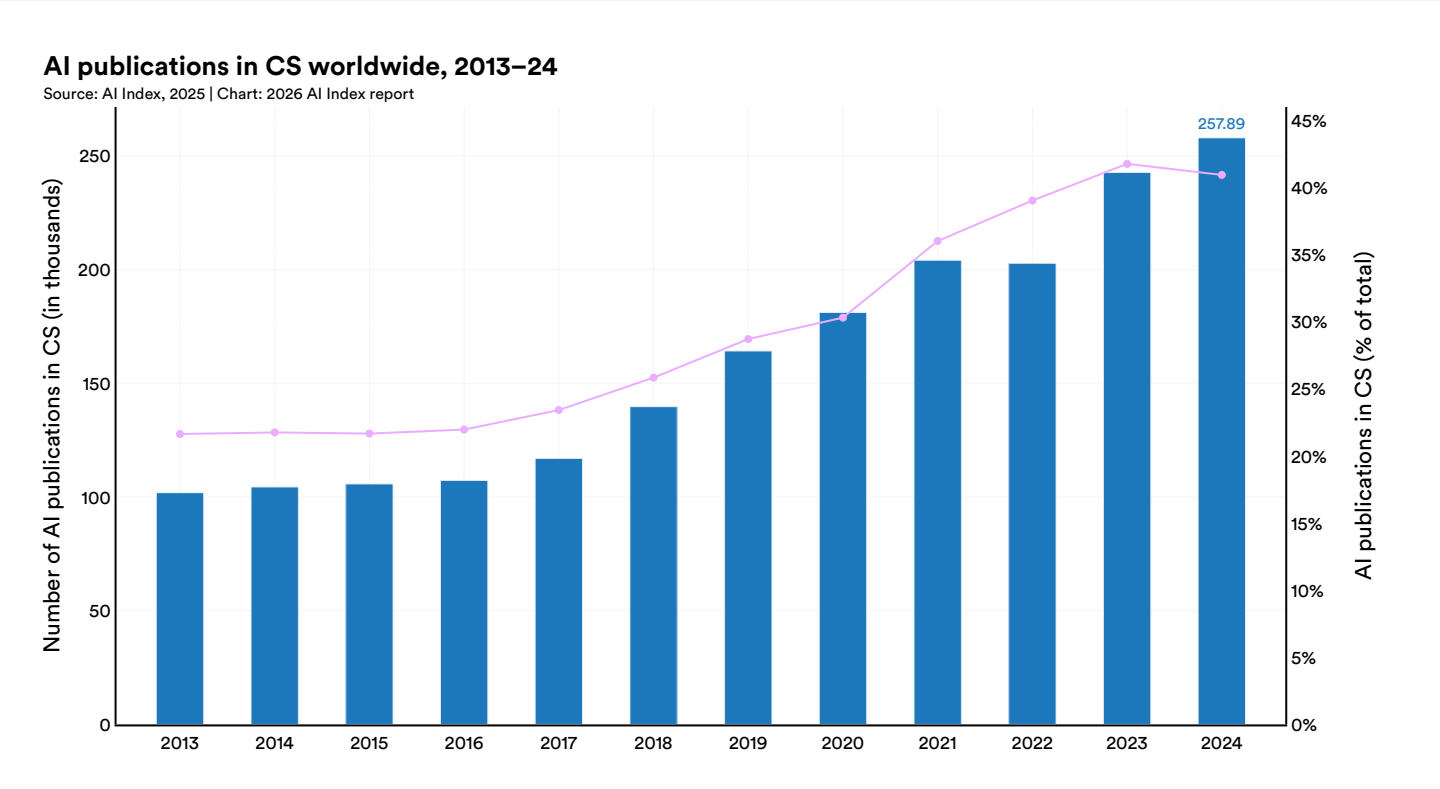


Figure 1.6.1

24 OpenAlex is a fully open catalog of scholarly metadata, including scientific papers, authors, institutions, and more. The AI Index used OpenAlex as a bibliographic database and automatically classified AI-related research using the latest version of the CSO Classifier. The CSO Classifier (v3.3) is an automated text classification system designed to categorize research papers in computer science using a comprehensive ontology of 15,000 topics and 166,000 relationships, including emerging fields like GenAI, large language models (LLMs), and prompt engineering. It processes metadata (such as title and abstract) through three modules: a syntactic module for exact topic matches, a semantic module leveraging word embeddings to infer related topics, and a post-processing module that refines results by filtering outliers and adding relevant higher-level areas.

By Venue

In 2024, journals accounted for the largest share of AI publications (62.8%), followed by conferences (23.8%) (Figure 1.6.2). Since 2013, both journal and conference publications have increased in absolute numbers, though their relative shares have shifted. The proportion of AI publications appearing in conferences has steadily declined from 36.6% in 2013 to its current level. The most recent year's results, however, may also reflect a lag in venue assignment, as papers often appear first in repositories[25] like arXiv before being formally published in a journal or conference.

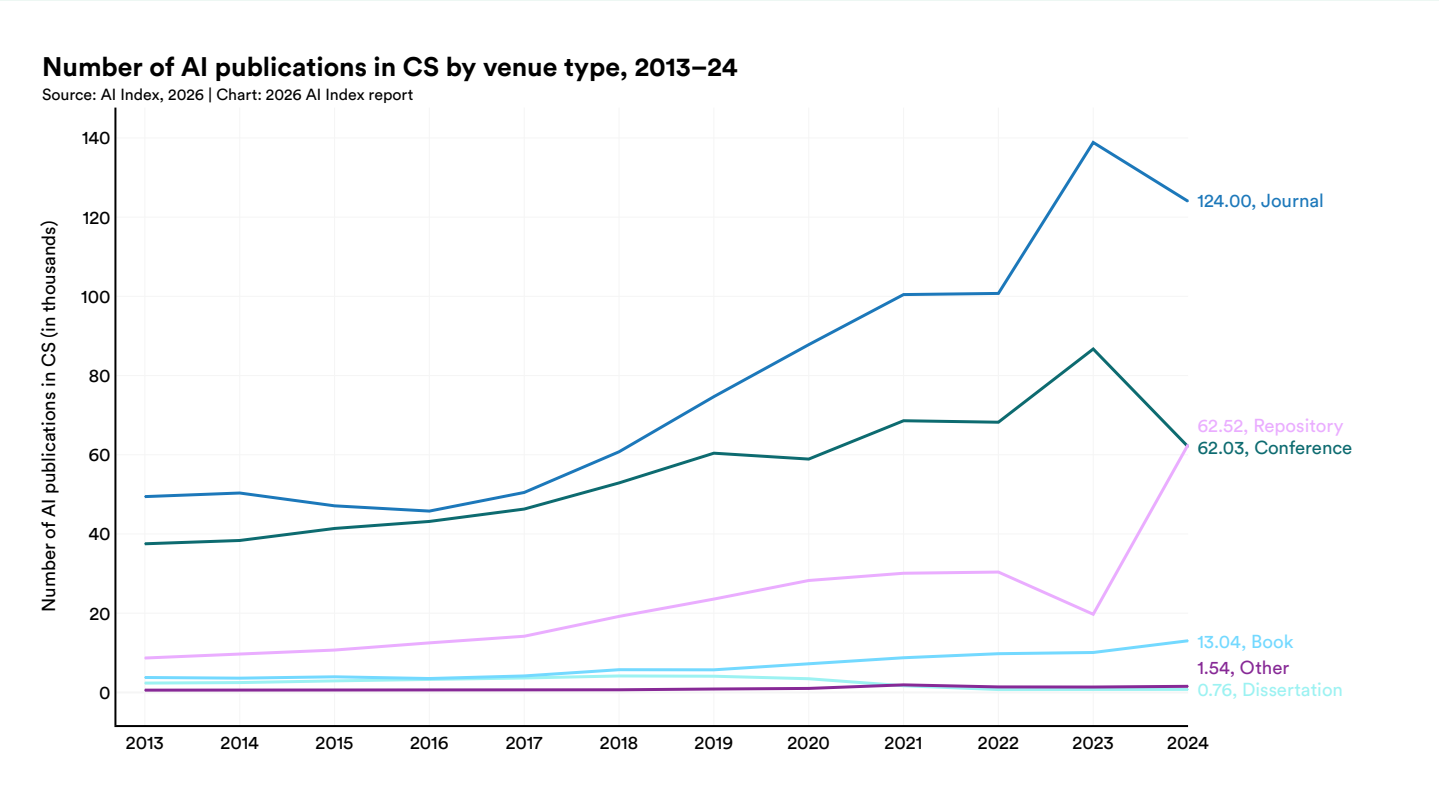


Figure 1.6.2

Conference Attendance

Publication venue patterns capture where AI research is formally published, while conference attendance offers a complementary view of research community engagement. Across the 16 major conferences tracked by the AI Index—AAAI, AAMAS, CVPR, EMNLP, FAccT, ICAPS, ICCV, ICLR, ICML, ICRA, IJCAI, IROS, KR, NeurIPS, UAI, and IUI—total attendance increased in 2024 from the previous year (Figure 1.6.3). The largest conferences, including NeurIPS, CVPR, and ICML, continued to draw the highest attendance, while smaller ones such as ICAPS, KR, and UAI maintained stable participation levels (Figures 1.6.4 and 1.6.5). This data should be interpreted with caution, as many conferences have recently switched to virtual or hybrid formats. Conference organizers report that measuring the exact attendance numbers at virtual conferences is difficult, as virtual conferences allow for higher attendance by researchers from around the world. The AI Index reports total attendance figures, encompassing virtual, hybrid, and in-person participation.

25 In this context, 'repository' refers to preprint platforms such as arXiv, where researchers post papers prior to or independent of formal peer-reviewed publication in a journal or conference.

Attendance at select AI conferences, 2010–25

Source: AI Index, 2025 | Chart: 2026 AI Index report

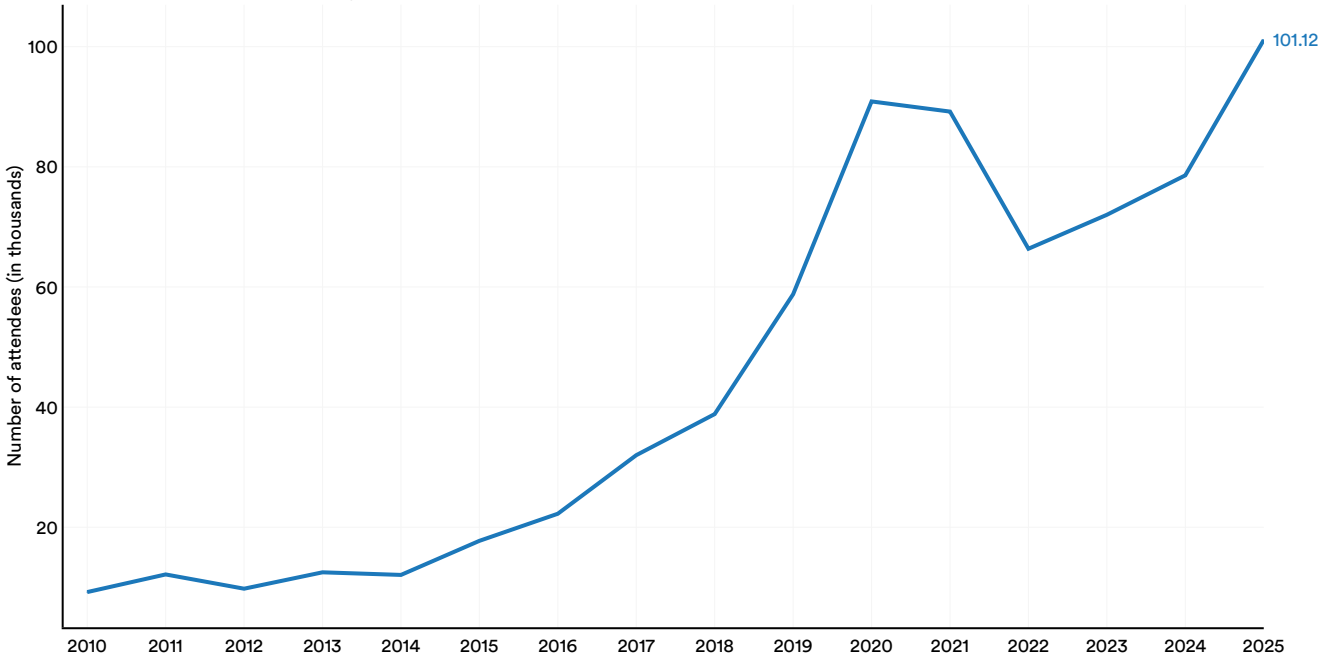


Figure 1.6.3

Attendance at larger conferences, 2010–25

Source: AI Index, 2025 | Chart: 2026 AI Index report

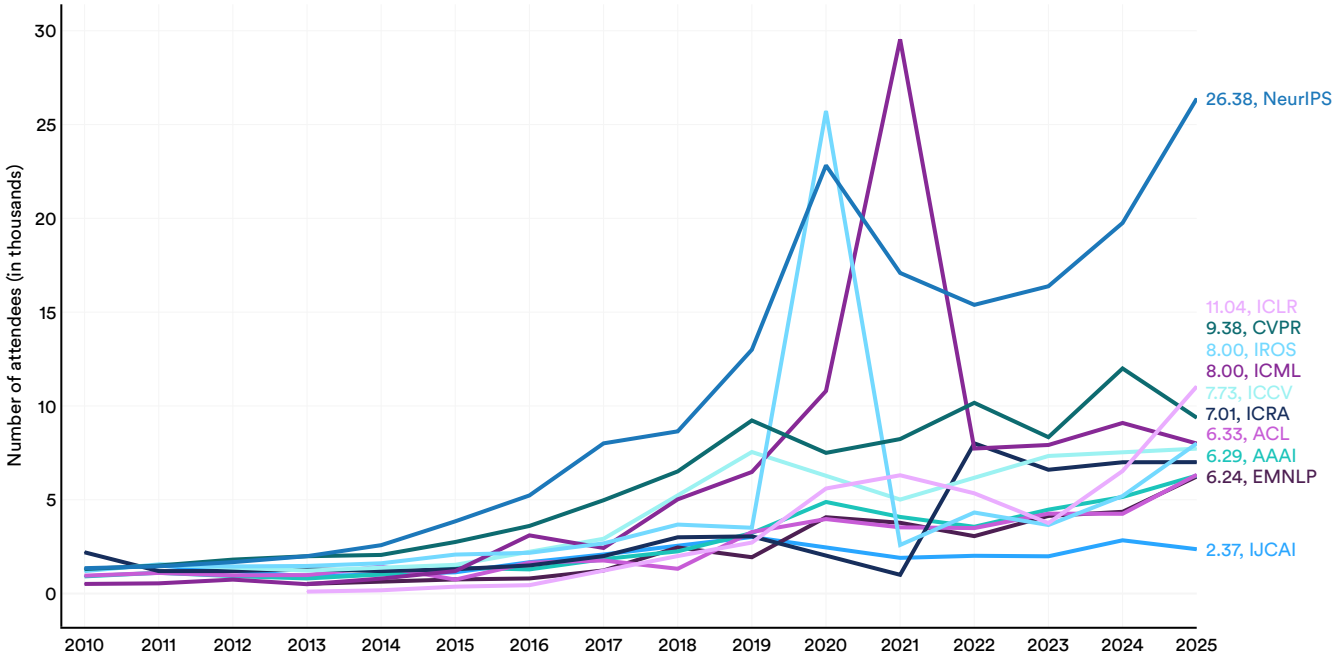


Figure 1.6.4[26]

26 The significant spike in ICML attendance in 2021 was likely due to the conference being held virtually that year.

Attendance at smaller conferences, 2010–25
Source: AI Index, 2025 | Chart: 2026 AI Index report

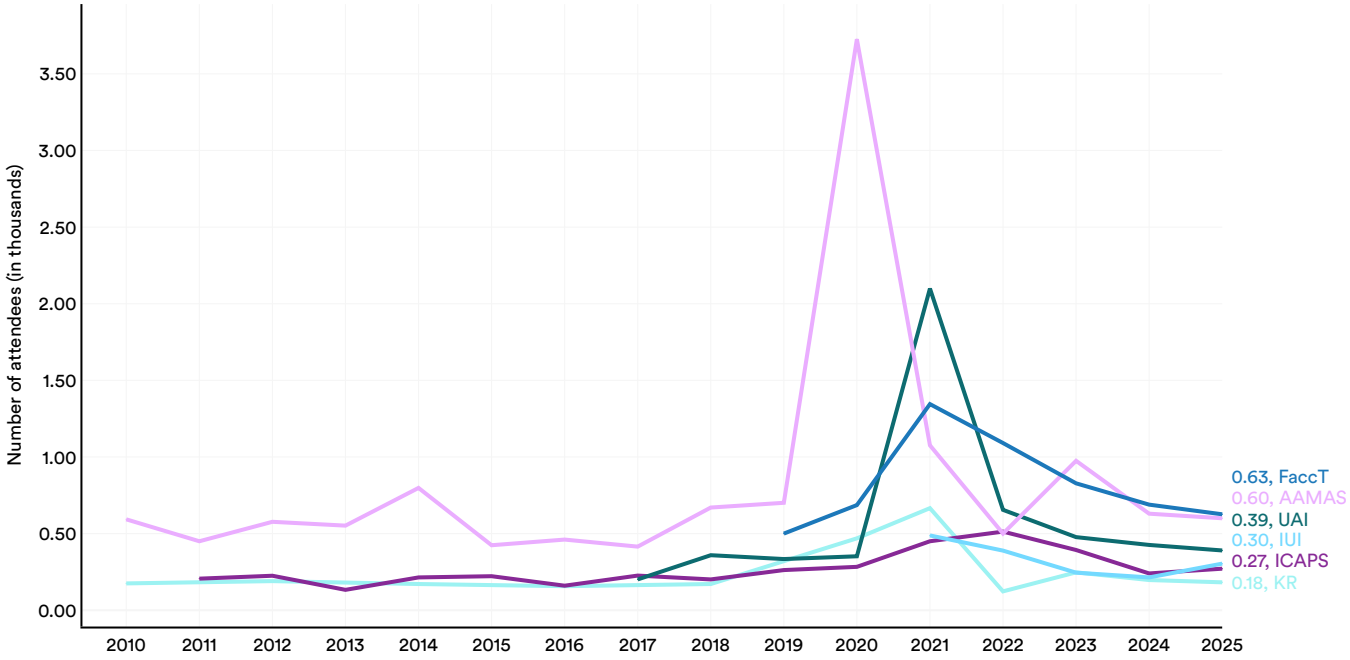


Figure 1.6.5[27]

By National Affiliation[28]

In 2024, China accounted for 17.8% of AI publications in 2024, compared to 11.1% from Europe and 7.6% from India (Figure 1.6.6[29]). Chinese AI publications also accounted for 20.6% of all AI citations in 2024, followed closely by Europe at 19.5% and the United States at 12.6% (Figure 1.6.7). The United States saw a decline of 3 percentage points in publication share, though its citation share remained relatively unchanged (12.6% in 2024 vs. 13.03% in 2023). The "unknown" share in publication data rose to 39.3% in 2024, a spike that likely reflects changes in metadata coverage. The geographic distribution of publications and citations adds context to the notable model trends discussed earlier in the chapter, where a relatively small number of countries account for a disproportionate share of activity.

27 IUI 2021 and 2022 were held exclusively virtually.

28 Regions in this chapter are classified according to the World Bank analytical grouping. The AI Index determines an author's country affiliation using the "countries" field from the authorship data. This field lists all the countries with which an author is affiliated, as retrieved from OpenAlex based on institutional affiliations. These affiliations can be explicitly stated in the paper or inferred from the author's most recent publications. When counting publications by country, the AI Index assigns one count to each country linked to the publication. For example, if a paper has three authors, two affiliated with institutions in the United States and one in China, the publication is counted once for the United States and once for China.

29 A publication may have an "unknown" country affiliation when the author's institutional affiliation is missing or incomplete. This issue arises due to various factors, including unstructured or omitted institution names, platform functional deficiencies, group authorship practices, unstandardized affiliation labeling, document type inconsistencies, or the author's limited publication record. The problem as it relates to OpenAlex is addressed in this paper; however, the issue of missing institutions pertains to other bibliographic databases as well.

AI publications in CS (% of total) by select geographic areas, 2013–24

Source: AI Index, 2026 | Chart: 2026 AI Index report

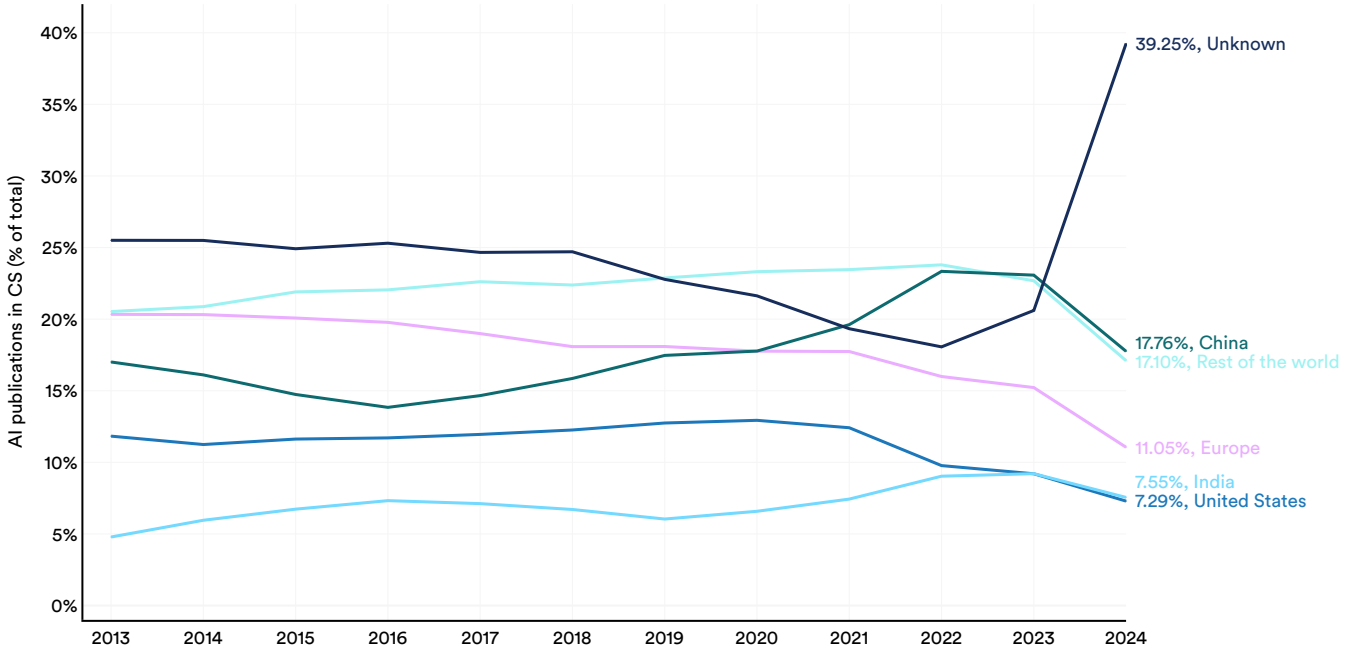


Figure 1.6.6[30]

AI publications in CS (% of total) by region, 2013–24

Source: AI Index, 2026 | Chart: 2026 AI Index report

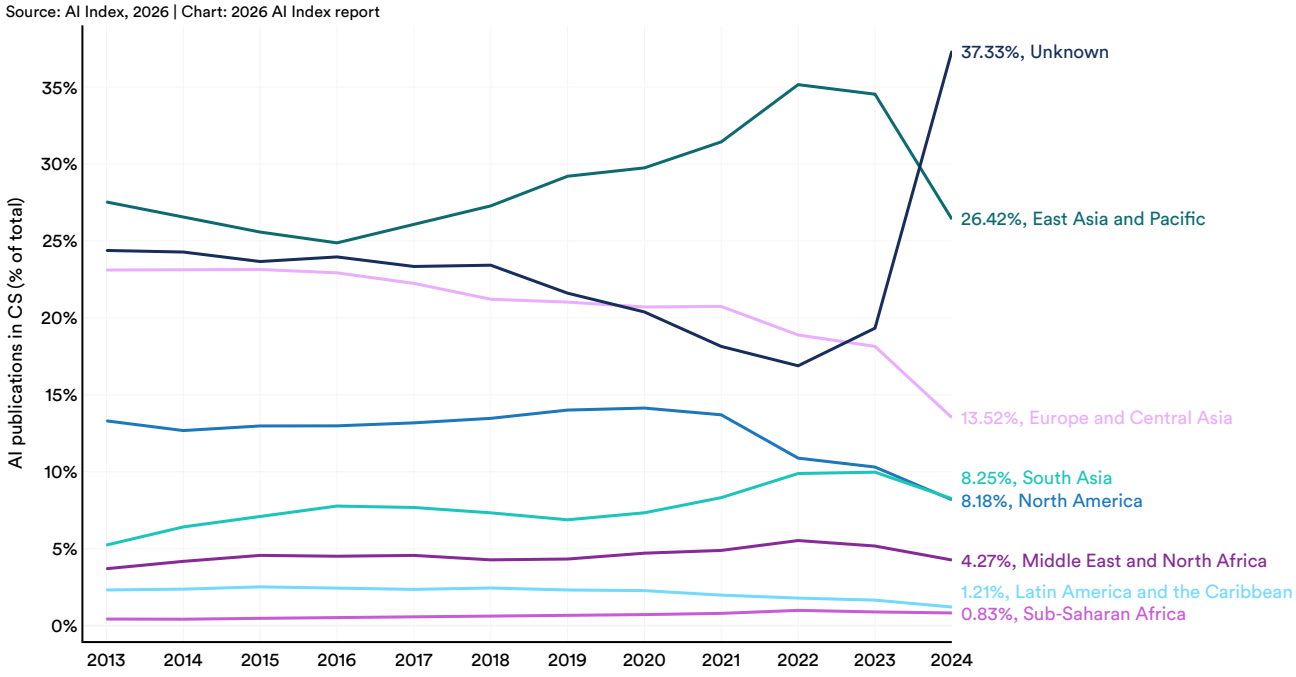


Figure 1.6.7

30 For the sake of brevity, the AI Index visualized results for a select group of countries. However, complete results for all countries will be available on the AI Index's Global Vibrancy Tool by the end of 2026. For immediate access to country-specific research and development data, please contact the AI Index team.

By Sector

Academia produced the majority of AI publications in 2024 (68.1%), followed by government institutions (12.4%), industry (11.5%), and nonprofit organizations (4.6%)(Figure 1.6.8). The sector breakdown does vary by region (Figure 1.6.9). In the United States, a higher share of AI publications came from industry (24.6%) compared to China (18%), where government institutions were more meaningful contributors (25.1%). Europe had the highest percentage of AI publications originating from academia (55.3%).

AI publications in CS (% of total) by sector, 2013–24

Source: AI Index, 2026 | Chart: 2026 AI Index report

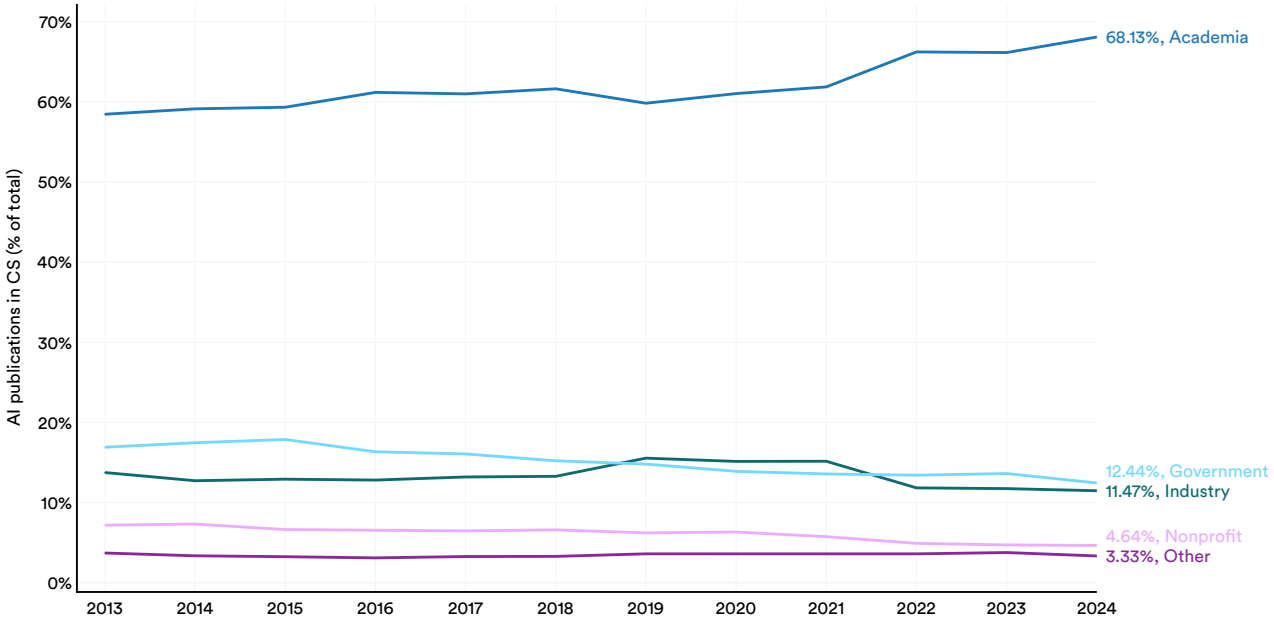


Figure 1.6.8[31]

AI publications in CS (% of total) by sector and geographic area, 2024

Source: AI Index, 2026 | Chart: 2026 AI Index report

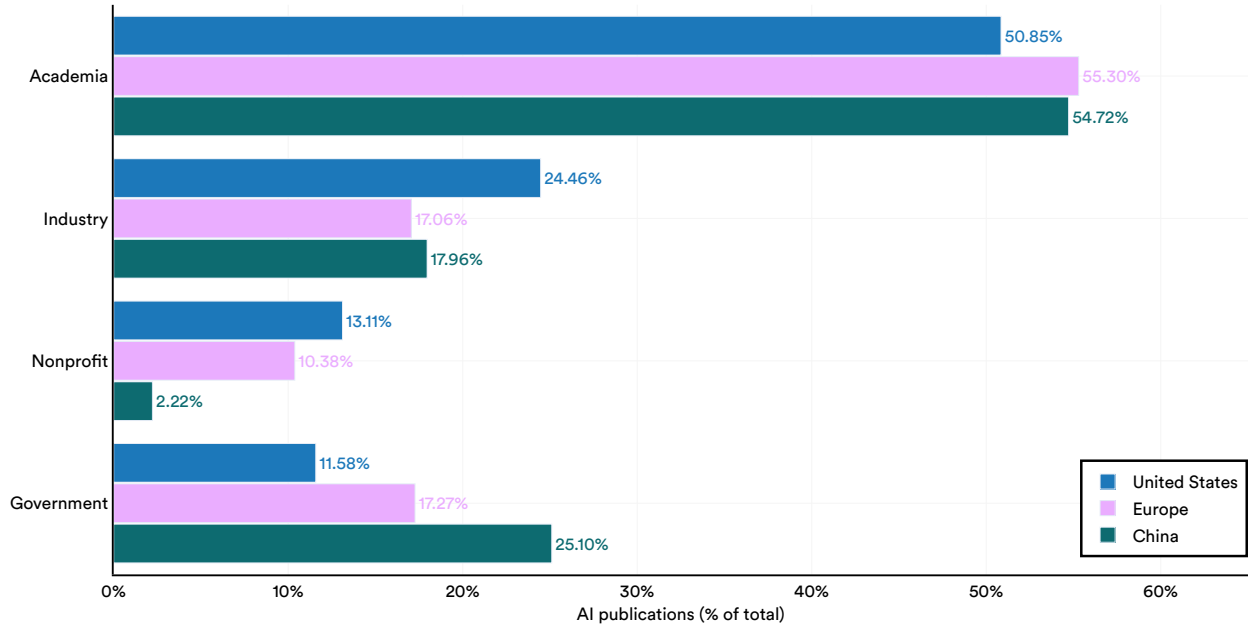


Figure 1.6.9

31 For Figure 1.6.8 and Figure 1.6.9, publications with unknown affiliations were excluded.

By Topic

AI research in 2024 remained concentrated in a small set of core topics, though the breadth of areas continued to expand. Similar to the previous year, the most prevalent research topic was machine learning (37%), followed by computer vision (22.4%), pattern recognition (11.2%), and natural language processing (10%) (Figure 1.6.10). Publications on generative AI continued to show sharp growth, extending the trend from previous years. It is also worth noting that the AI Index topic classifier can assign multiple topic labels to a single publication, so topic totals can be seen as overlapping categories rather than mutually exclusive.

Number of AI publications by select top topics, 2013–24

Source: AI Index, 2026 | Chart: 2026 AI Index report

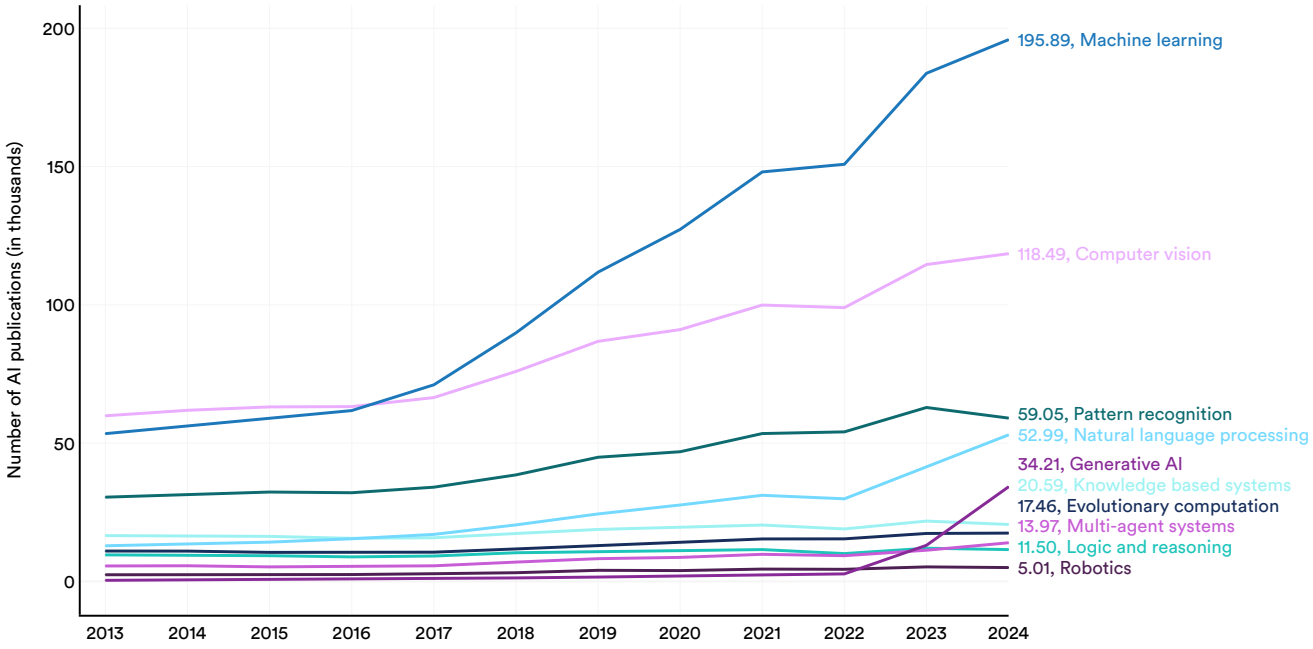


Figure 1.6.10[32]

Top 100 Publications

The AI Index identified the 100 most-cited AI publications from 2021 to 2024 using citation data from OpenAlex.[33] Due to citation lag, this set can shift as citations accumulate over time.[34] The publication volume data above captures the scale of research activity, while the top 100 offers a more selective view on which work is gaining the most recognition and influence.

By National Affiliation

The geographic distribution of the top 100 has shifted over time (Figure 1.6.11). The United States still ranks highest in top-cited publications each year, though its share has gradually declined from 64 in 2021 to 46 in 2024. China's share increased to 41 in 2024, from 34 in 2023, and Australia increased to 14 highly cited publications, up from 2 in 2023 and 6 in 2021.

32 The AI Index categorized papers using its own topic classifier. It is possible for a single publication to be assigned multiple topic labels.

33 The full methodological guide can be accessed in the Appendix, along with the list of the top 100 articles.

34 A publication can have multiple authors from different countries or organizations. If a paper includes authors from multiple countries, each country is credited once. As a result, some of the totals in this section exceed 100.

Number of highly cited publications in top 100 by select geographic areas, 2021–24

Source: AI Index, 2026 | Chart: 2026 AI Index report

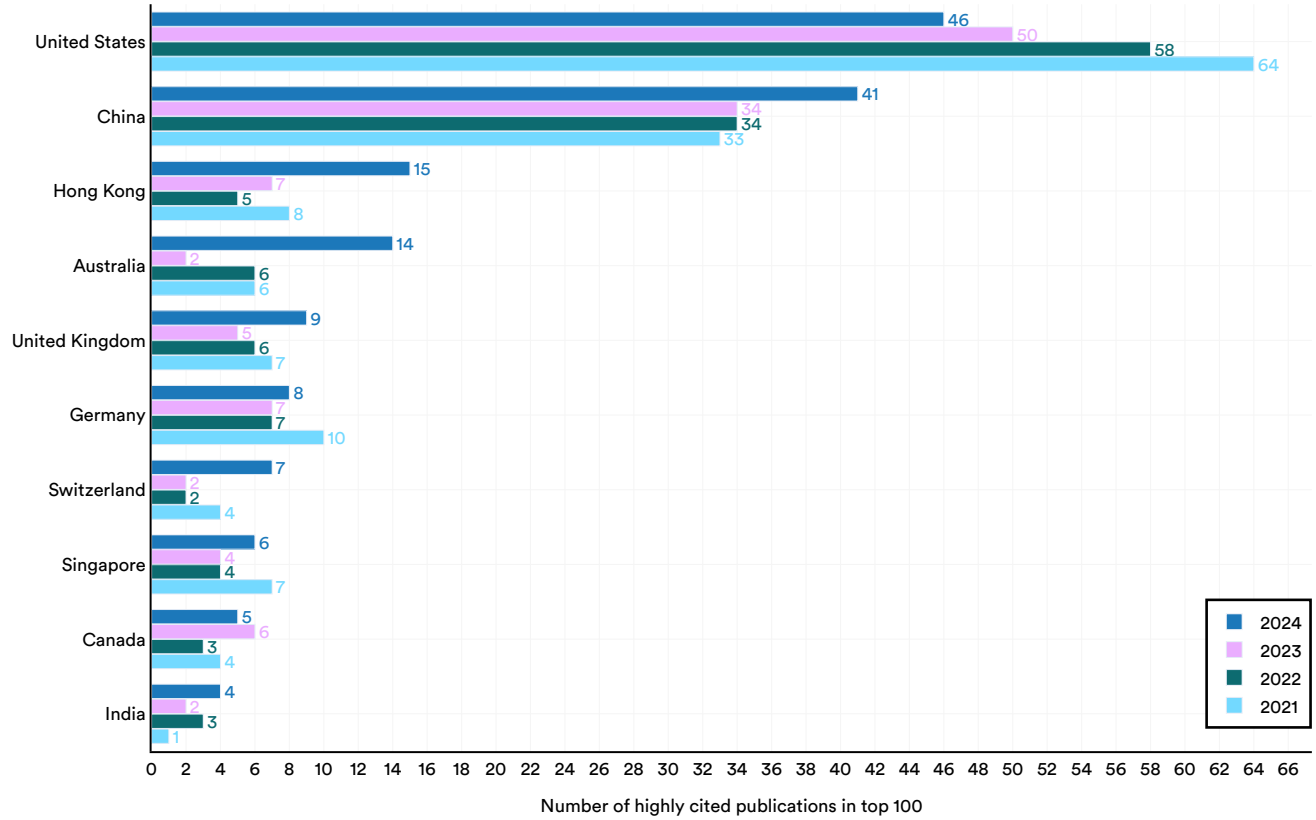


Figure 1.6.11

By Sector and Organization

The sector composition of the top 100 remained consistent, with academia producing the most top-cited publications year over year (Figure 1.6.12). Industry contributions declined sharply from 17 in 2021 and 19 in 2022 to six in 2024, even as industry's share of notable model releases has continued to grow (Section 1.1). The organization distribution varies by year, though output remains concentrated among a small set of institutions (Figure 1.6.13). In 2024, Stanford University and Google led with seven publications each, and the Chinese Academy of Sciences and Microsoft followed closely, with each contributing five.

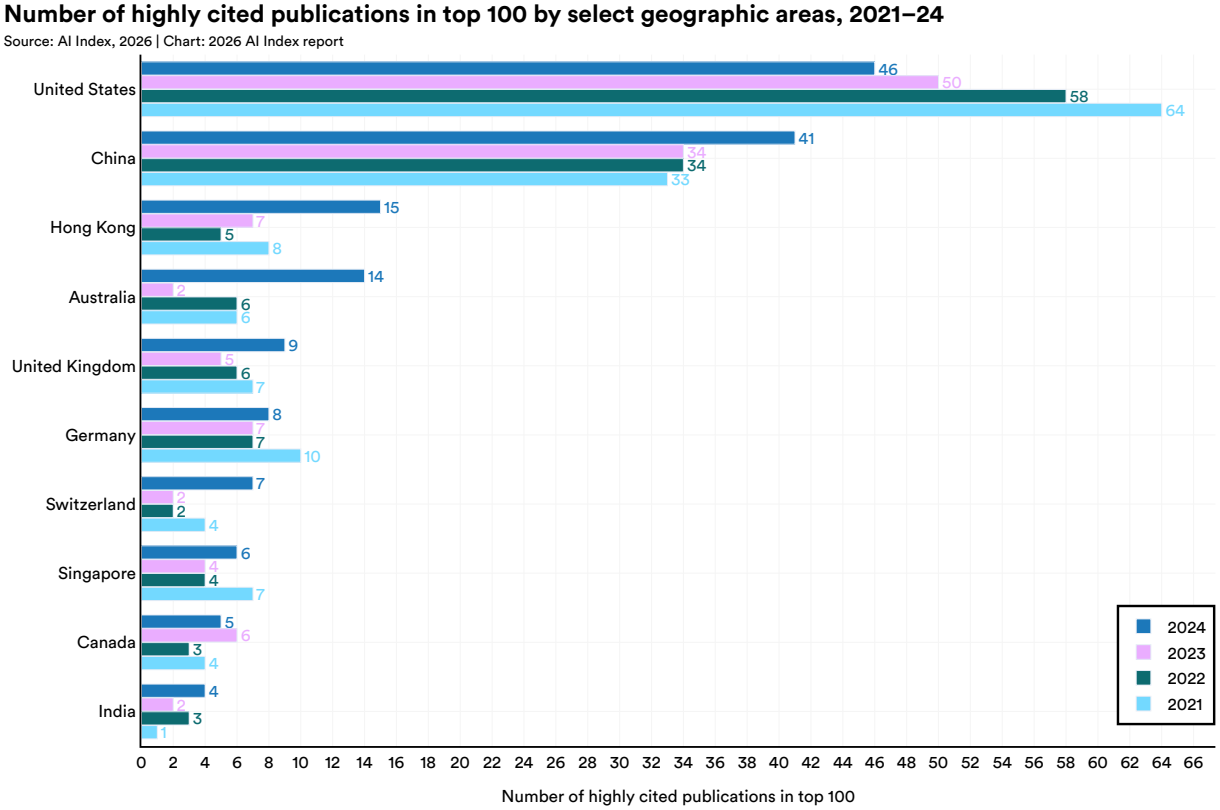


Figure 1.6.12[35]

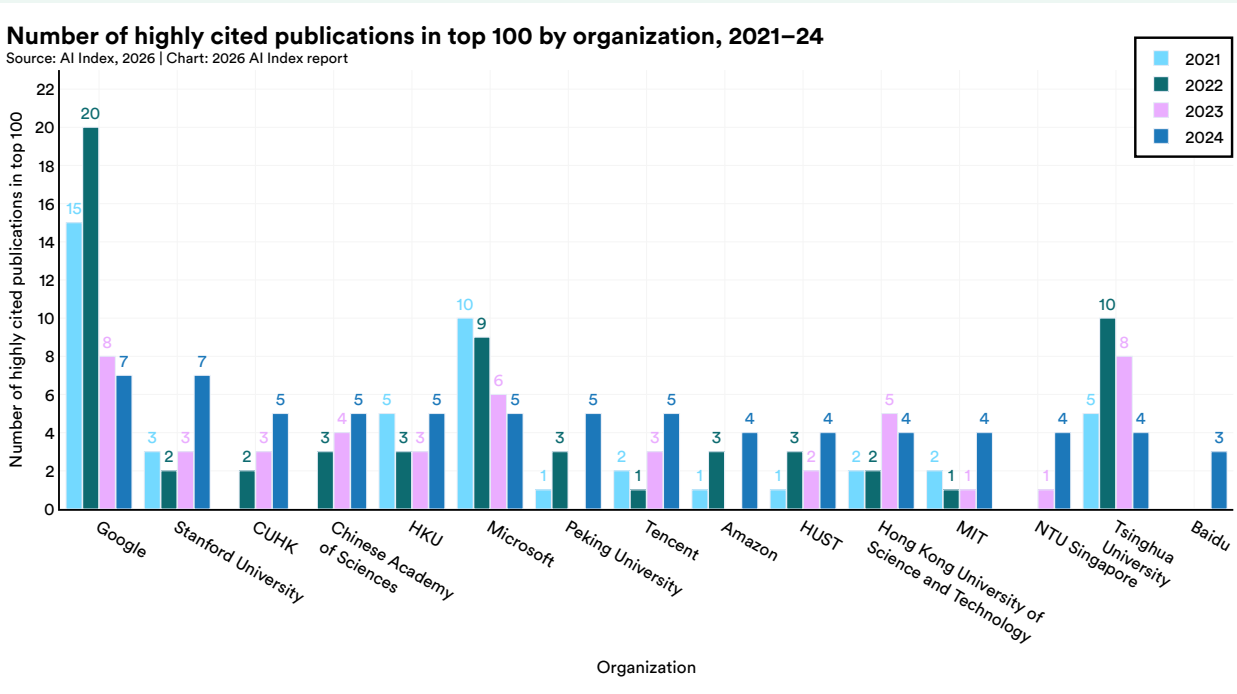


Figure 1.6.13[36]

35 The "other" category includes sectors and intersector collaborations that are not industry and academia, or industry-academia collaborations (e.g., industry and government, academia and nonprofit). Some institutions lack data for 2021 because they did not have papers included in the top 100 that year. Since papers can have multiple authors from different institutions, the total institutional tags in Figure 1.6.13 may exceed 100. Also, because two of the papers had authors with an unknown sectoral affiliation in 2022, the total sum of publications in Figure 1.6.12 is 98.

36 Universities are abbreviated as follows: CUHK = The Chinese University of Hong Kong; HKU = The University of Hong Kong; HUST = Huazhong University of Science and Technology; MIT = Massachusetts Institute of Technology; NTU Singapore = Nanyang Technological University, Singapore.

1.7 Patents

While publications track research outputs, patents offer insight into applied innovation and commercial development. This section examines trends in global AI patents over time. Patents can provide another lens for tracking innovation across organizations and geographic areas, particularly in applied AI contexts. Similar to publications data, there are notable delays before AI patent data becomes available, with 2024 being the most recent year accessible. The analysis draws from patent-level bibliographic records in PATSTAT Global, a comprehensive database provided by the European Patent Office (EPO).[37]

Global Trends

Globally, the number of granted AI patents has grown exponentially, from 3,866 in 2010 to 131,121 in 2024 (Figure 1.7.1). Between 2023 and 2024, patent grants rose by 8.2%. China accounts for the majority, at 74.2% of the global total (Figures 1.7.2 and 1.7.3). The United States is the next major contributor at 12.1% (15,290 patents), followed by Europe (3%) and India (0.4%). Over the past decade, the United States' share has declined steadily from a peak of 42.8% in 2015, while China's share has risen from under 20% to its current level. Patents and publications reflect different stages in the R&D pipeline, so China's lead in both, while not directly correlated, is consistent with the country's growing research presence described earlier.

Other regional leaders emerge when patent activity is normalized by population size (Figure 1.7.4). In 2024, South Korea had the highest number of granted AI patents on a per capita basis (14.3%), followed by Luxembourg (12.3%) and China (7.0%).

Number of AI patents granted worldwide, 2010–24

Source: AI Index, 2026 | Chart: 2026 AI Index report

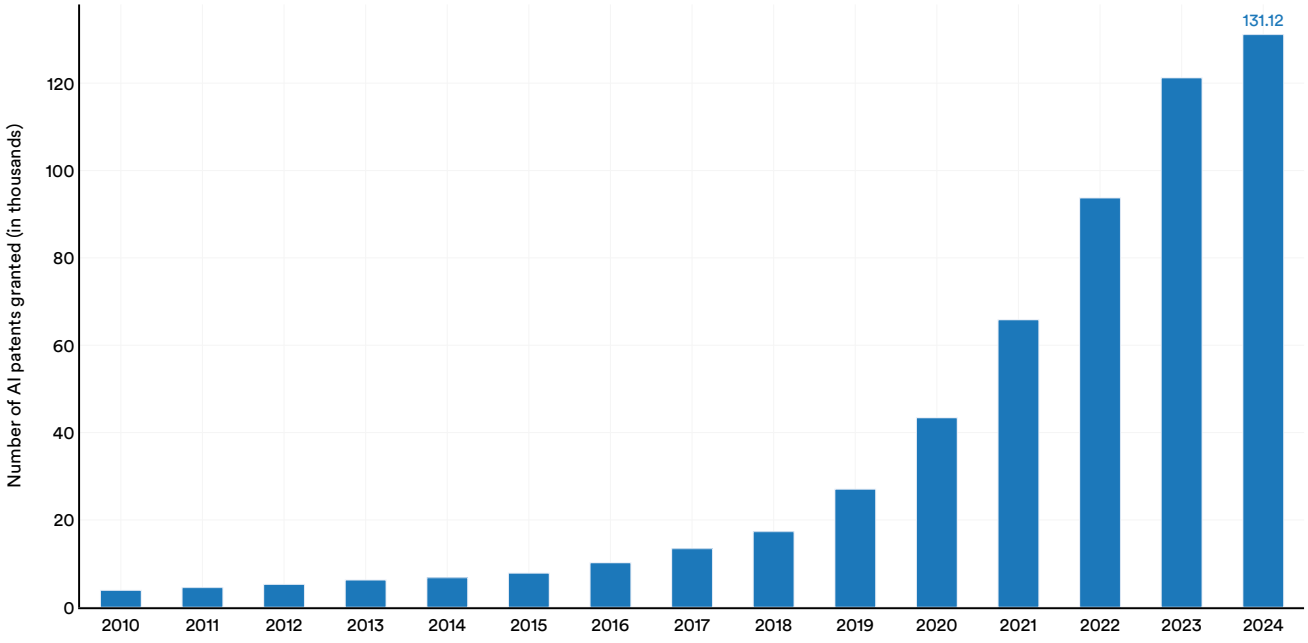


Figure 1.7.1[38]

37 More details on the methodology behind this section's patent analysis can be found in the Appendix.

38 Patent standards and laws vary across countries and regions, so these charts should be interpreted with caution. More detailed country-level patent information will be released in a subsequent edition of the AI Index's Global Vibrancy Tool.

Granted AI patents (% of world total) by select geographic areas, 2010–24

Source: AI Index, 2026 | Chart: 2026 AI Index report

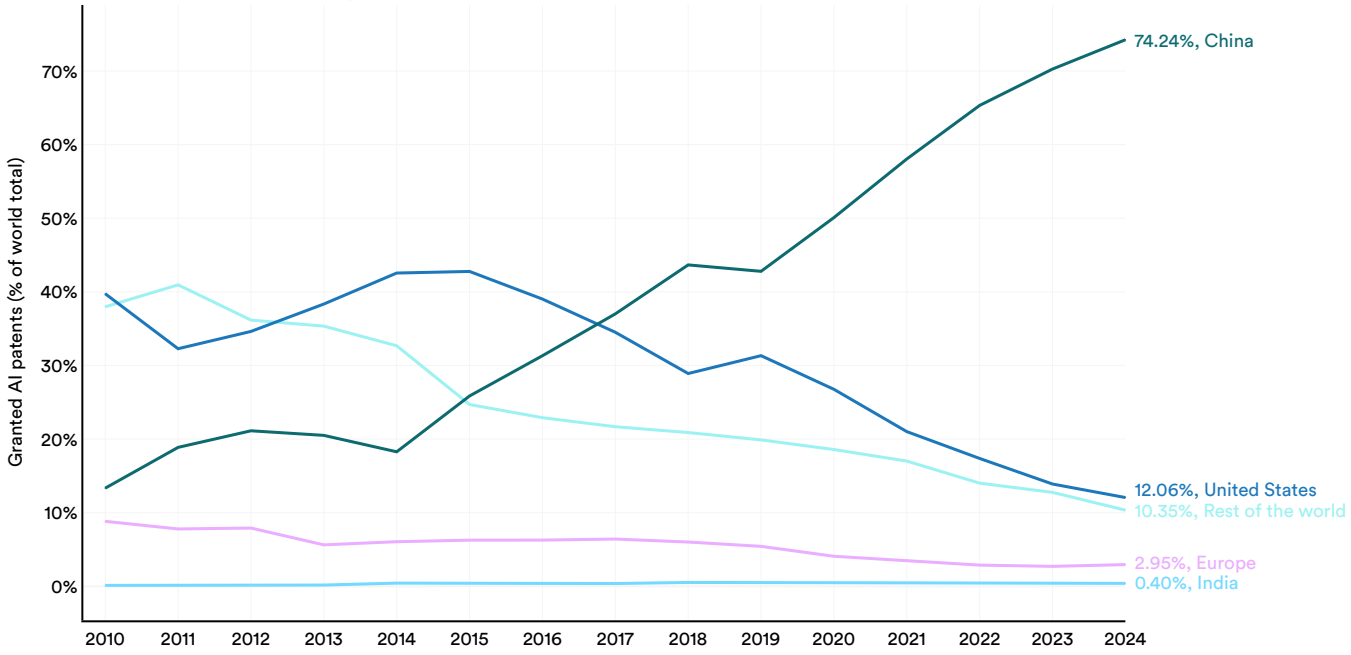


Figure 1.7.2

Number of AI patents granted by select geographic areas, 2010–24

Source: AI Index, 2026 | Chart: 2026 AI Index report

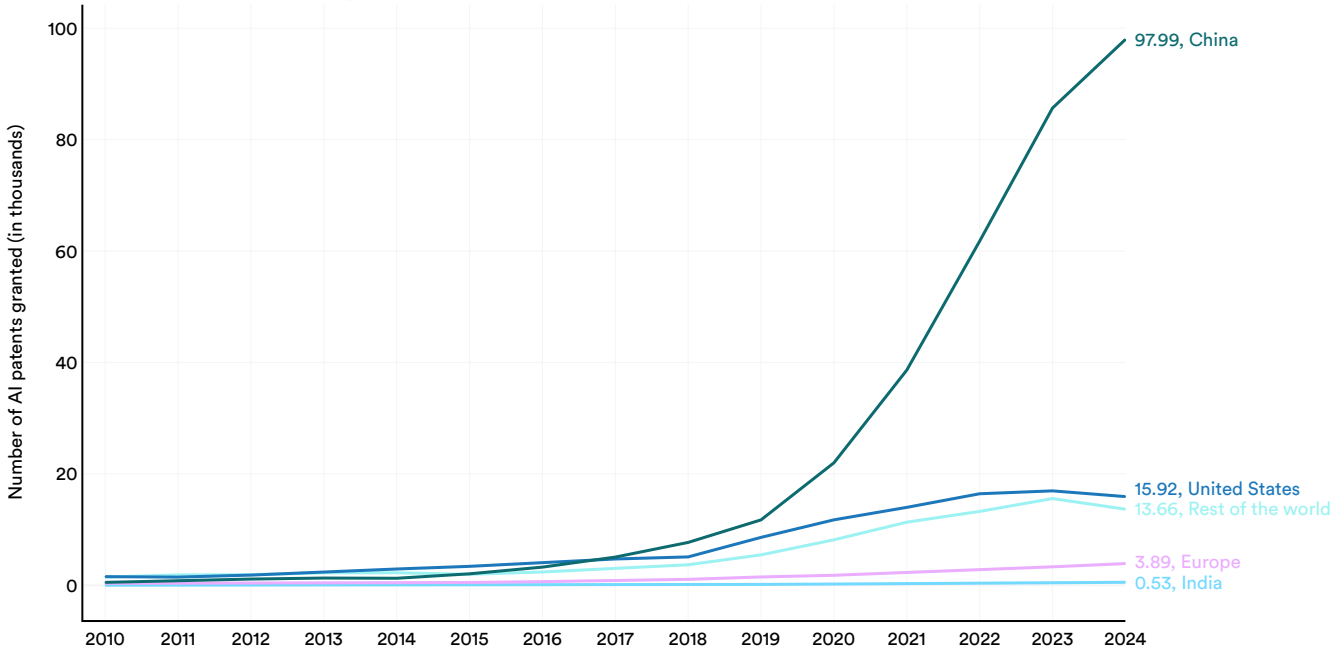


Figure 1.7.3

Granted AI patents per 100,000 inhabitants by country, 2024

Source: AI Index, 2026 | Chart: 2026 AI Index report

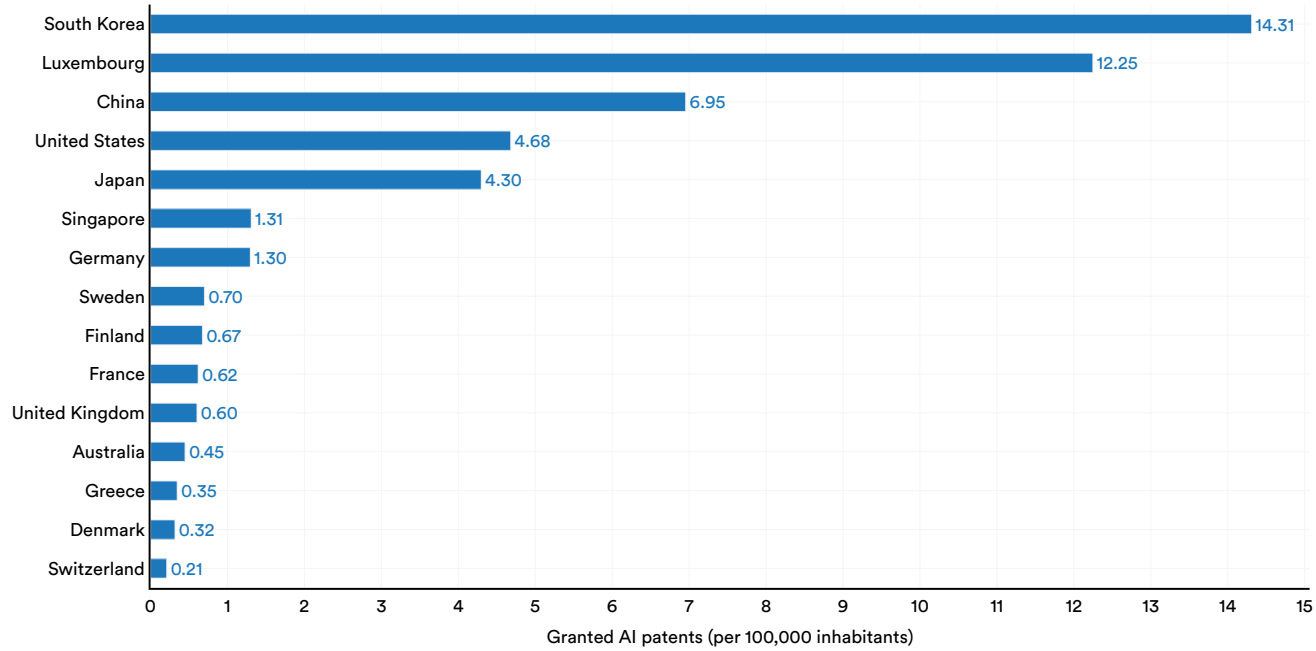
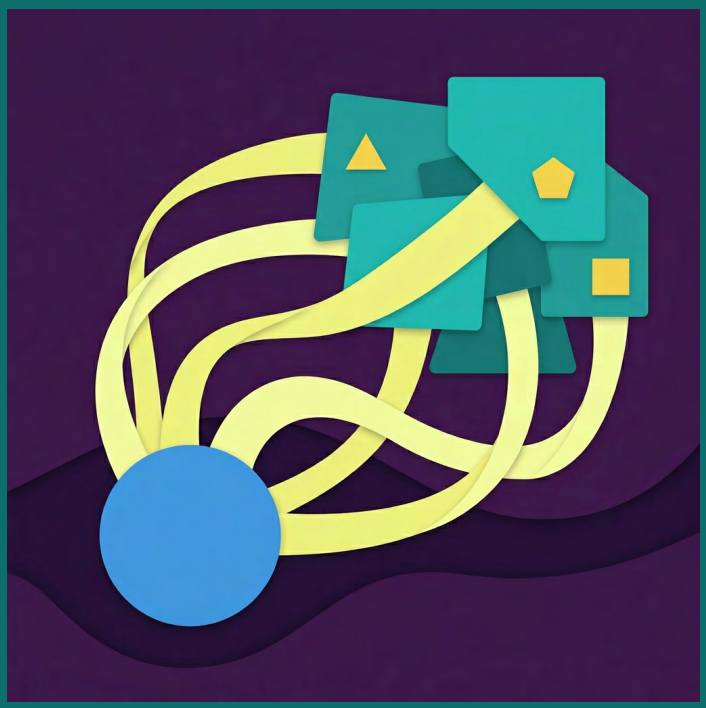


Figure 1.7.4

Forward Citations Flow

When newly filed patents reference earlier ones, those references are called forward citations. These are often used as a proxy for influence, since they indicate how often an invention informs later work. By this measure, the United States accounts for over half of all AI patent forward citations, a signal of downstream influence that contrasts with its 12.1% share of patent volume (Figure 1.7.5). China ranks second despite producing the largest volume of patents by a wide margin. The relationship between forward citations and technological impact is not straightforward and has been called into question (Higham et al., 2021). There is also a strong home bias across all countries, with most citations occurring domestically, a well-documented pattern in patent citation geography (Jaffe et al., 1993; Cotterlaz et al. 2025; and Verluise et al., 2025).That said, the cross-border flows are not symmetric. Chinese patents are cited frequently in U.S. filing, while U.S. patents appear far less often in Chinese ones.



Global distribution of forward citations to AI patents by geographic area, 2010–24
Source: AI Index, 2026 | Chart: 2026 AI Index report

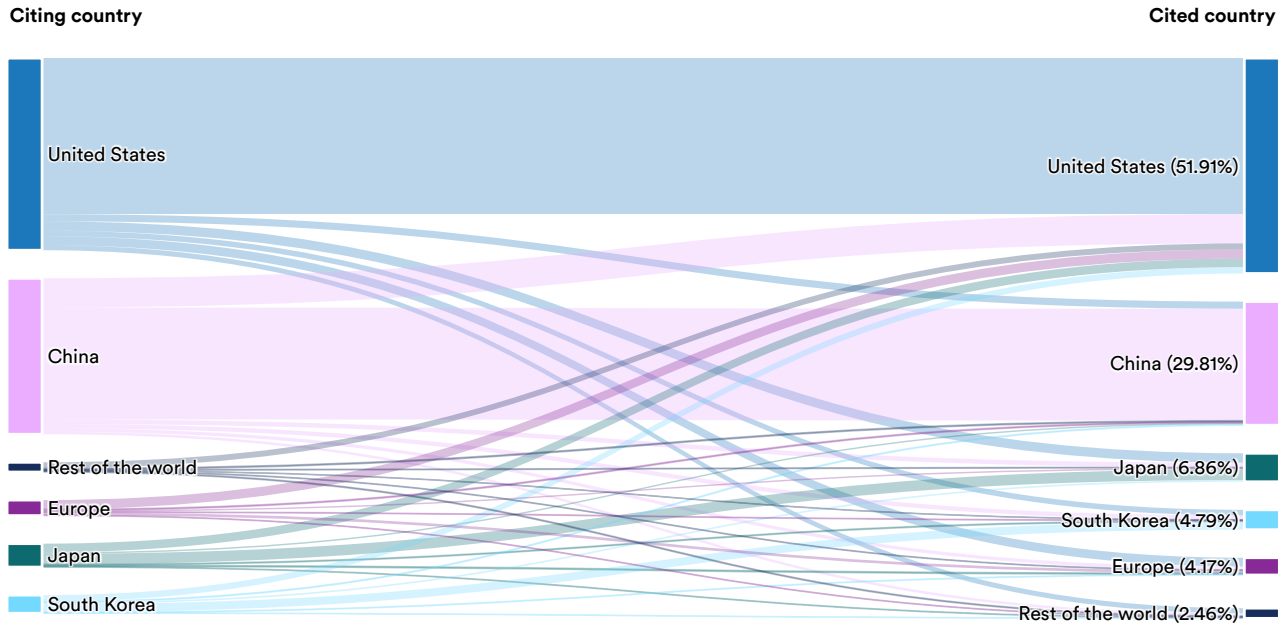


Figure 1.7.5[39]

Speed of Knowledge Diffusion

Patent citation lag—the time between a patent's publication and its first forward citation—can be used to measure how quickly knowledge diffuses within a discipline. For AI patents, most receive their first citation within two to three years, reflecting a relatively fast diffusion. The speed varies by country (Figure 1.7.6). U.S. patents tend to be cited sooner and more consistently over time, with only 19% remaining uncited compared to 32% to 44% in other geographic areas. Japan's patents show early but narrower influence, and those from China and South Korea experience slower initial citation but, after about six years, citation activity stabilizes across all regions. The patterns are consistent with the forward citation data above, but differences in citation norms and home bias may also play a role.

39 Each data point in the figure reflects forward citations to AI patents, grouped at the patent family level to represent unique inventions rather than individual filings. Values are expressed as shares of all AI patent forward citations for patents granted between 2010 and 2024.

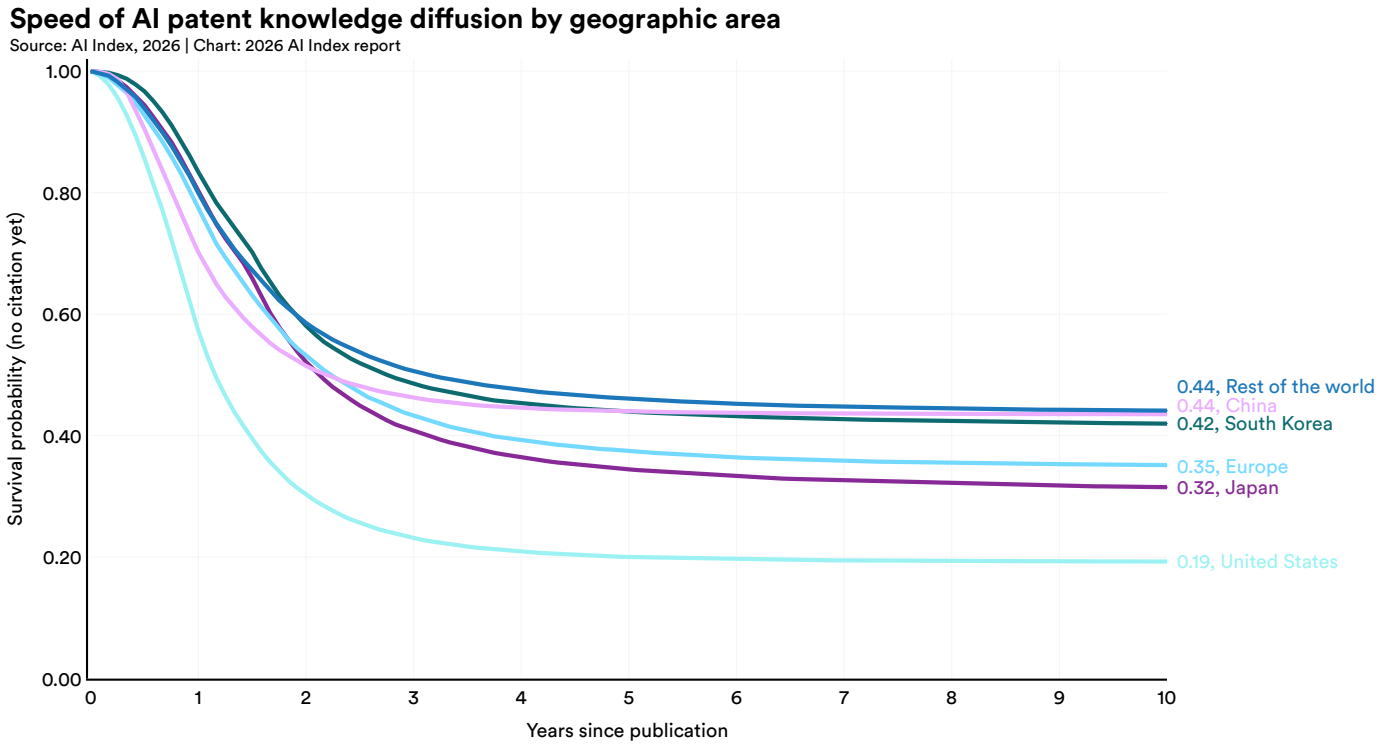


Figure 1.7.6[40]

Technological Proximity

Technological proximity[41] measures whether countries are converging on similar types of AI innovation or pursuing distinct paths. Using a method proposed by Bar et al. (2012), the analysis[42] compares how closely each country's AI patent portfolio aligns with the two largest reference points, the United States and China (Figure 1.7.7). Overlap is scored on a scale from 0 (no similarity) to 1 (identical). Most countries cluster in the upper right, meaning their AI patents cover similar technological areas to both the U.S. and China, with a stronger lean toward the U.S. portfolio. India and Australia, for example, have patent portfolios that show close to 80% overlap with both. Denmark is the least similar to either reference point, showing only a 45% overlap with China and a 52% overlap with the United States. This is because Denmark's AI patents are concentrated in energy and wind-related technology categories (patent codes Y02E, F03D, F05B) rather than core computing and data-processing categories (G06F, G06N, G06K) that dominate both the U.S. and China. While most countries' AI innovation portfolios are structured similarly, national industrial strengths tend to influence where AI is applied.

40 This figure plots the probability of *not* being cited, so curves with sharper drops indicate shorter lags.

41 Also known as the min-complement proximity measure.

42 Technological classes are identified by International Patent Classification (IPC) and Cooperative Patent Classification (CPC) codes.

AI patent portfolios' technological proximity to the United States and China, 2010–24
Source: AI Index, 2026 | Chart: 2026 AI Index report

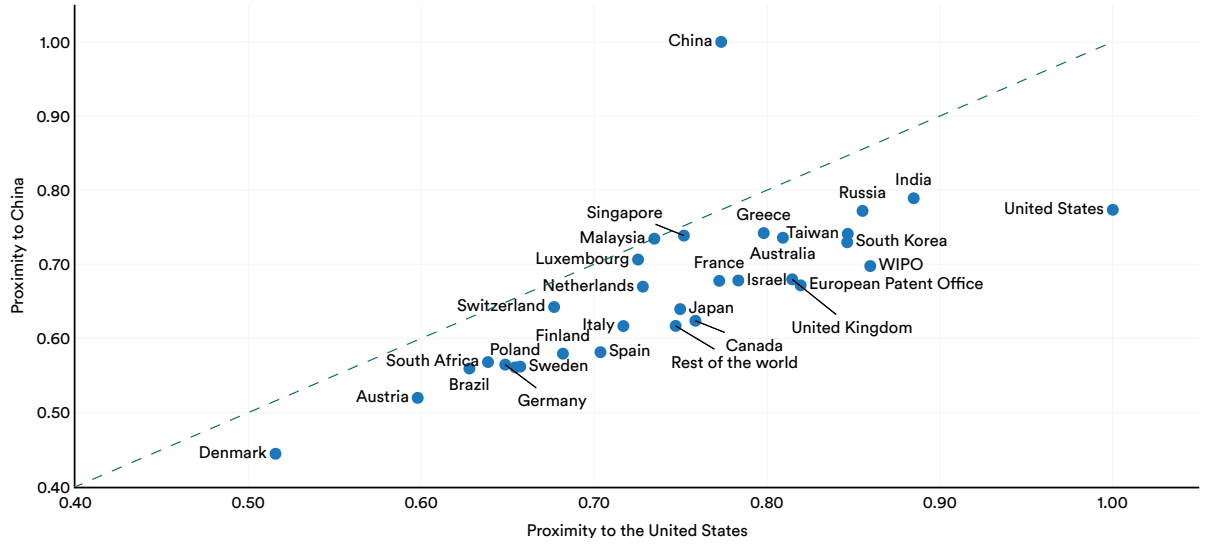


Figure 1.7.7

HIGHLIGHT:

AI Patent Examples

1 Patent CN111431996A: Resource configuration method and device, equipment and medium, 2022, China

A machine-learning prediction model determines how to allocate computing resources across multiple services in a cluster. The system learns from historical and real-time signals—such as traffic volumes and CPU, memory, and network usage—to infer the right resource configuration. This enables automated, dynamic scaling decisions without relying on manual rules.

2 Patent US11436777B1: Machine learning-based hazard visualization system, 2022, United States

The system trains machine learning models to forecast hazard attributes (time, path, severity) for specific locations and identify infrastructure in geospatial imagery. It combines model outputs to annotate maps, showing where hazards intersect with critical assets. The system also supports causal inference—for example, identifying infrastructure repeatedly affected by hazards. These capabilities rely on learned prediction and image-recognition models rather than deterministic mapping logic.

3 Patent US2023239456A1: Display system with ML-based stereoscopic view synthesis over a wide field of view, 2025, United States

This head-mounted display uses machine-learning techniques—including depth estimation and reconstruction—to create perspective-correct stereoscopic images from external cameras. Neural models handle real-time vision challenges like disocclusion, artifact reduction, and sharpening by inferring scene geometry and filling gaps where camera viewpoints fail to align with the user's eyes. ML is a core component of the VR/AR passthrough rendering pipeline.

1.8 AI Authors and Inventors

The publications and patents discussed above reflect research and development outputs. Using Zeki data, the AI Index examined the geographic distribution and mobility patterns of the authors and inventors behind this work over time. This section covers a narrower slice of AI talent activity than the broader labor market indicators discussed in Chapter 4 (Economy). Zeki identifies talent outside of China based on observable AI outputs such as research, data depositories, and new models. The dataset covers 2010 to 2025 across a group of countries in North America, Europe, Asia, Latin America, and the Middle East.[43]

Geographic Distribution

In 2025, the largest share of identified AI authors and inventors came from the United States (220,520), followed by India (50,460) and Germany (48,520) (Figure 1.8.1). The United Kingdom (34,370), Canada (31,450), and France (18,820) formed a second tier with Australia, the Netherlands, Italy, Brazil, Switzerland, and others making up the broader distribution of contributors. Looking at the data on a per capita basis surfaces countries that have relatively high levels of AI activity that are not visible when looking at total volume, as seen in the per capita patent data in Section 1.7. Switzerland led with 110.5 AI authors and inventors per 100,000 inhabitants, followed closely by Singapore (109.5) (Figure 1.8.2). Countries with smaller populations, such as Finland (77.6), the Netherlands (77.6), and Denmark (66.3), rank above larger nations including Germany (58.1) and the United Kingdom (49.6).

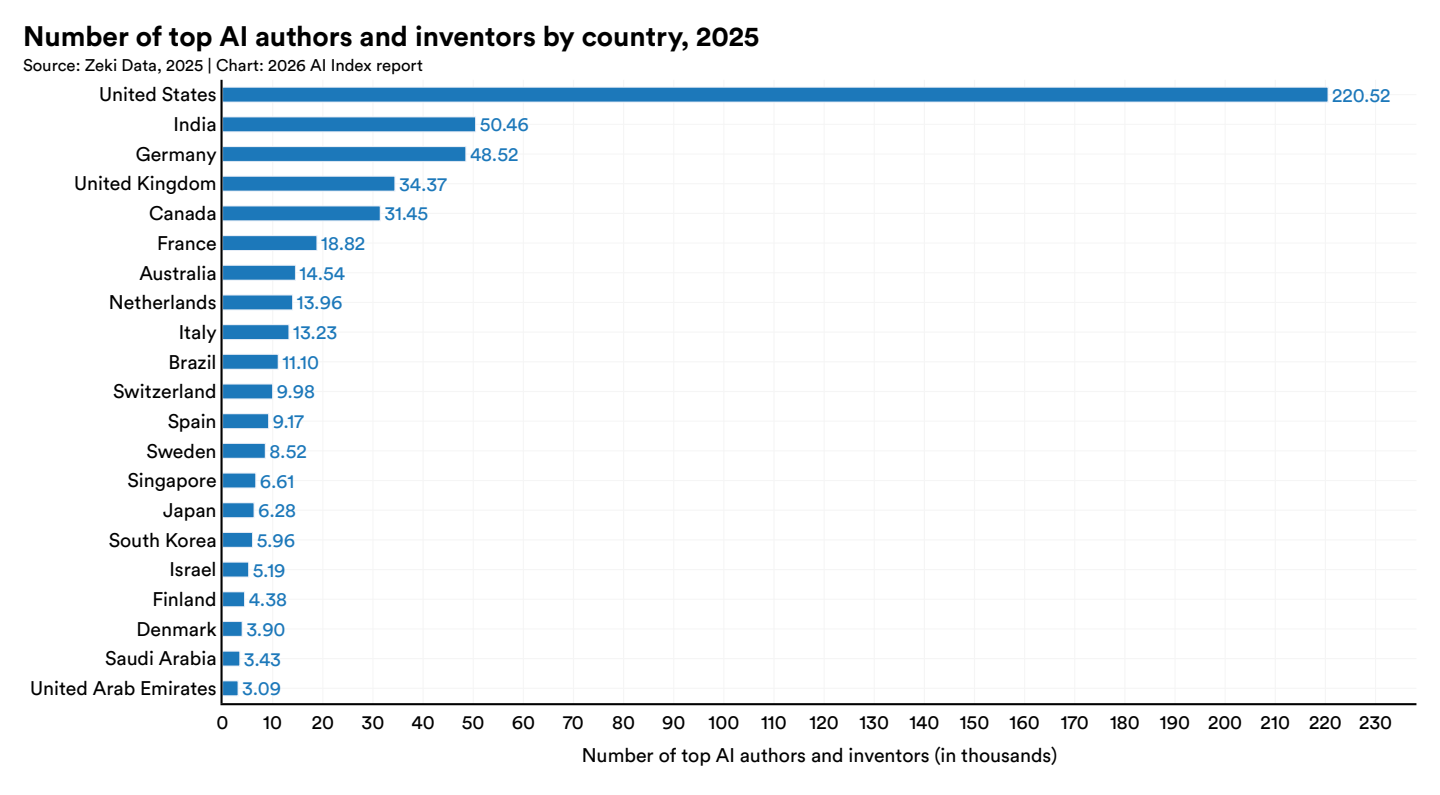


Figure 1.8.1

43 For more details, refer to the Appendix.

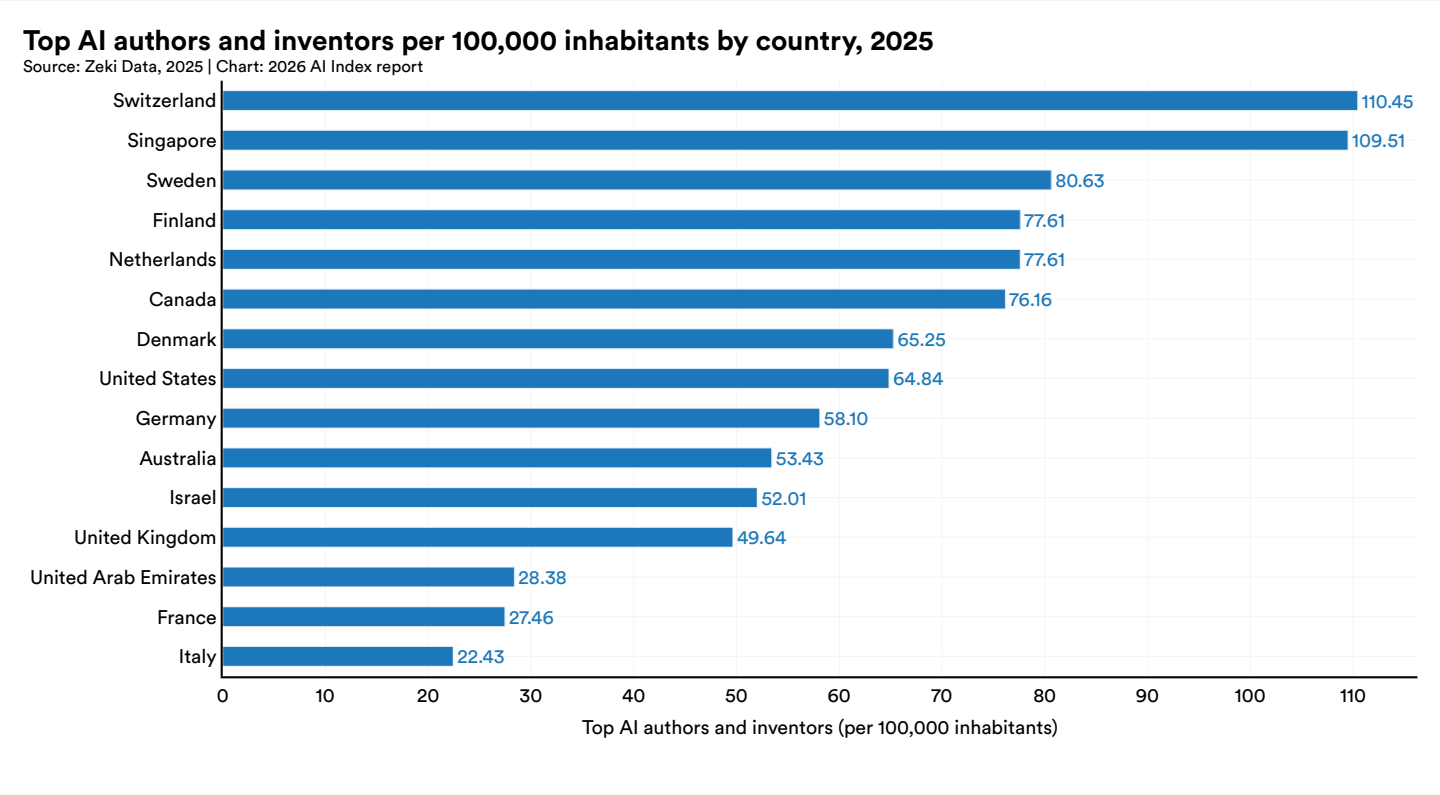


Figure 1.8.2

By Education Level

The educational profile of top AI authors and inventors varies by country, though in most of the countries, PhD holders and those with master's degrees together account for the majority in 2025 (Figure 1.8.3). The United Kingdom (51.1%) and Australia (50.5%) have the highest share of PhD holders, followed by Switzerland (43.6%), South Korea (42.5%), and the United States (42%). India and Brazil show a more varied distribution, with comparatively lower shares of PhD holders and a wider spread across other degree levels.

Percentage of top AI authors and inventors by education level and country, 2010–25
Source: Zeki Data, 2025 | Chart: 2026 AI Index report

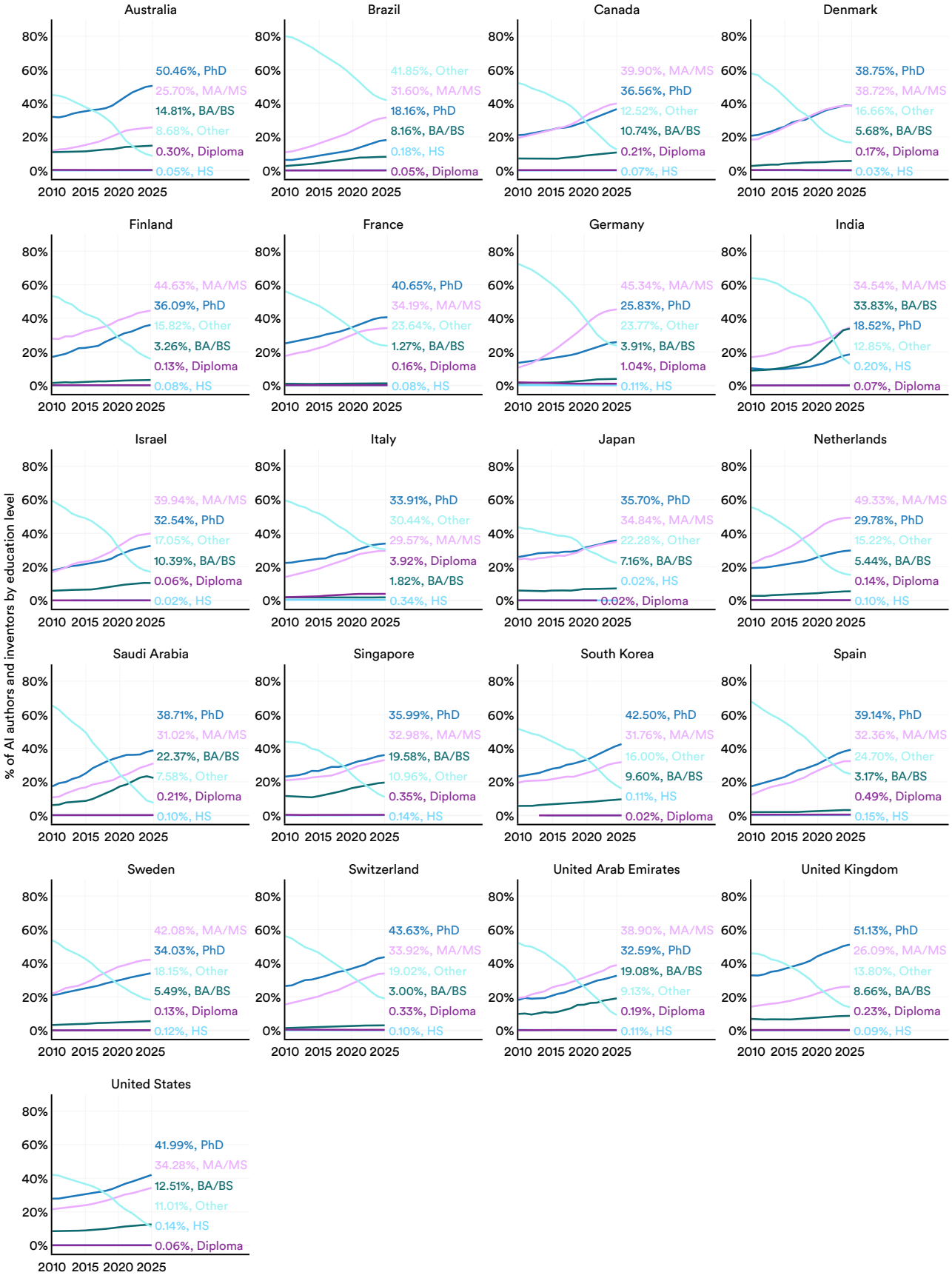


Figure 1.8.3

By Gender

The gender gap among AI authors and inventors is visible across all countries, with men making up the majority in all cases, though the size of the gap varies (Figure 1.8.4). In Brazil, South Korea, and Japan, more than 80% of identified AI talent is male. Female representation is somewhat higher in Saudi Arabia (32.3%), Australia (30.1%), Canada (29.6%), and Italy (29.5%), but no country comes close to parity. More significantly, in almost every country, the male-female ratio has remained flat from 2010 to 2025. Even with the growth in AI talent overall, there has been no meaningful progress on gender balance. Chapter 7 (Education) describes a similar pattern in AI-related degree attainment, where women remain underrepresented across all levels.

Top AI authors and inventors (% of total) by gender, 2010–25
Source: Zeki Data, 2025 | Chart: 2026 AI Index report

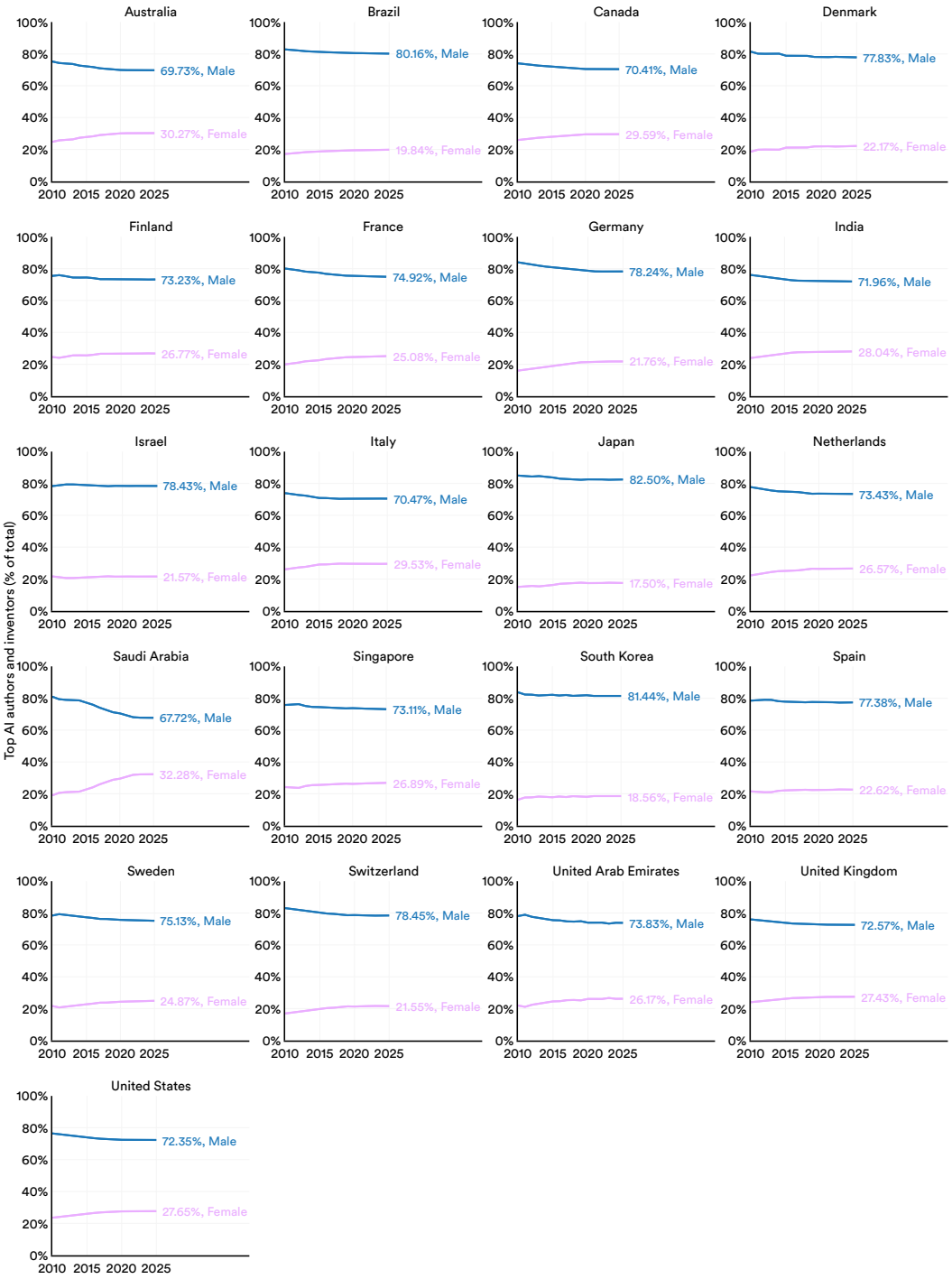


Figure 1.8.4

By Specialization

AI authors and inventors are distributed across a range of specialization areas, though each country shows its own emphasis (Figure 1.8.5). Healthcare and bioinformatics, computer vision and image processing, and software engineering are among the most common areas globally, accounting for 10% or more of the pool in several countries. A few country-level patterns connect to findings discussed earlier in this chapter. South Korea, for example, has the highest share of talent in hardware, VLSI, and IoT (20%), consistent with its role in the semiconductor supply chain described in Section 1.3. Brazil has the highest share of software engineering talent (18%), while Saudi Arabia leads in security, privacy, and cryptography (15%).

Within-country distribution of top AI authors and inventors across specialization areas

Source: Zeki Data, 2025 | Chart: 2026 AI Index report

Area of specialization	Australia	Brazil	Canada	Denmark	Finland	France	Germany	India	Israel	Italy	Japan	Netherlands	Saudi Arabia	Singapore	South Korea	Spain	Sweden	Switzerland	United Arab Emirates	United Kingdom	United States
AI applications (general and industrial)	7%	9%	6%	9%	9%	6%	10%	7%	4%	8%	5%	7%	8%	6%	4%	7%	8%	6%	10%	6%	5%
Computer vision and image processing	11%	10%	12%	8%	11%	12%	10%	13%	12%	12%	14%	10%	9%	12%	15%	12%	9%	11%	11%	11%	10%
Data science and big data	5%	6%	4%	5%	5%	4%	5%	4%	4%	4%	4%	5%	4%	4%	4%	5%	5%	5%	4%	4%	6%
Ethics and social impact	1%	2%	1%	2%	2%	1%	1%	1%	1%	2%	1%	2%	2%	1%	1%	2%	2%	1%	2%	2%	2%
HCI and user interaction	8%	6%	7%	13%	10%	5%	7%	6%	5%	6%	8%	10%	7%	6%	5%	6%	9%	5%	7%	8%	7%
Hardware, VLSI, and IoT	9%	8%	10%	8%	11%	12%	10%	12%	9%	11%	14%	9%	10%	15%	20%	10%	10%	10%	9%	9%	11%
Healthcare and bioinformatics	13%	6%	14%	14%	9%	10%	11%	10%	13%	13%	9%	14%	11%	8%	8%	11%	10%	14%	9%	14%	13%
Machine learning and deep learning algorithms	8%	10%	9%	7%	6%	10%	8%	8%	14%	8%	8%	8%	7%	10%	8%	8%	7%	9%	8%	9%	9%
NLP and speech processing	6%	6%	6%	6%	7%	8%	7%	8%	7%	6%	8%	7%	7%	8%	7%	8%	6%	6%	7%	7%	7%
Physical sciences and environment	7%	6%	6%	5%	4%	5%	5%	3%	5%	4%	5%	4%	4%	4%	5%	5%	4%	5%	4%	5%	5%
Robotics and control systems	6%	6%	7%	6%	4%	6%	8%	6%	5%	8%	9%	6%	4%	8%	7%	7%	7%	8%	7%	6%	7%
Security, privacy, and cryptography	8%	6%	7%	5%	6%	7%	6%	13%	9%	7%	6%	6%	15%	8%	6%	6%	6%	7%	12%	7%	7%
Software engineering and networks	10%	18%	11%	11%	14%	13%	11%	10%	10%	11%	9%	11%	12%	9%	10%	14%	16%	11%	10%	10%	11%
Other AI research	1%	0%	0%	1%	2%	1%	1%	0%	0%	1%	0%	1%	0%	0%	0%	1%	2%	0%	0%	1%	1%

Figure 1.8.5

Mobility

Mobility is measured through net flow, which is the difference between the number of AI authors and inventors who move to or out of their respective countries (Figure 1.8.6). The United States has remained net positive since 2020, meaning it attracts more talent than it loses, though the magnitude has declined from a peak of 324.6 in 2022 to 26.0 in 2025. Most other countries operate on a smaller scale. Saudi Arabia (3.1) and Denmark (2.1) were among the few with positive net flow in 2025.

Canada, which showed strong inflow around 2020, declined to -7.1 by 2025. Germany also showed negative net flow at -2.4, while India had the largest net outflows at -16.9 in 2025. These flows are relevant in the context of other factors, including immigration policy and geographic distribution of investment and employment, discussed further in Chapter 4's section on labor markets.

Net flow of top AI authors and inventors by country, 2010–25
Source: Zeki Data, 2025 | Chart: 2026 AI Index report

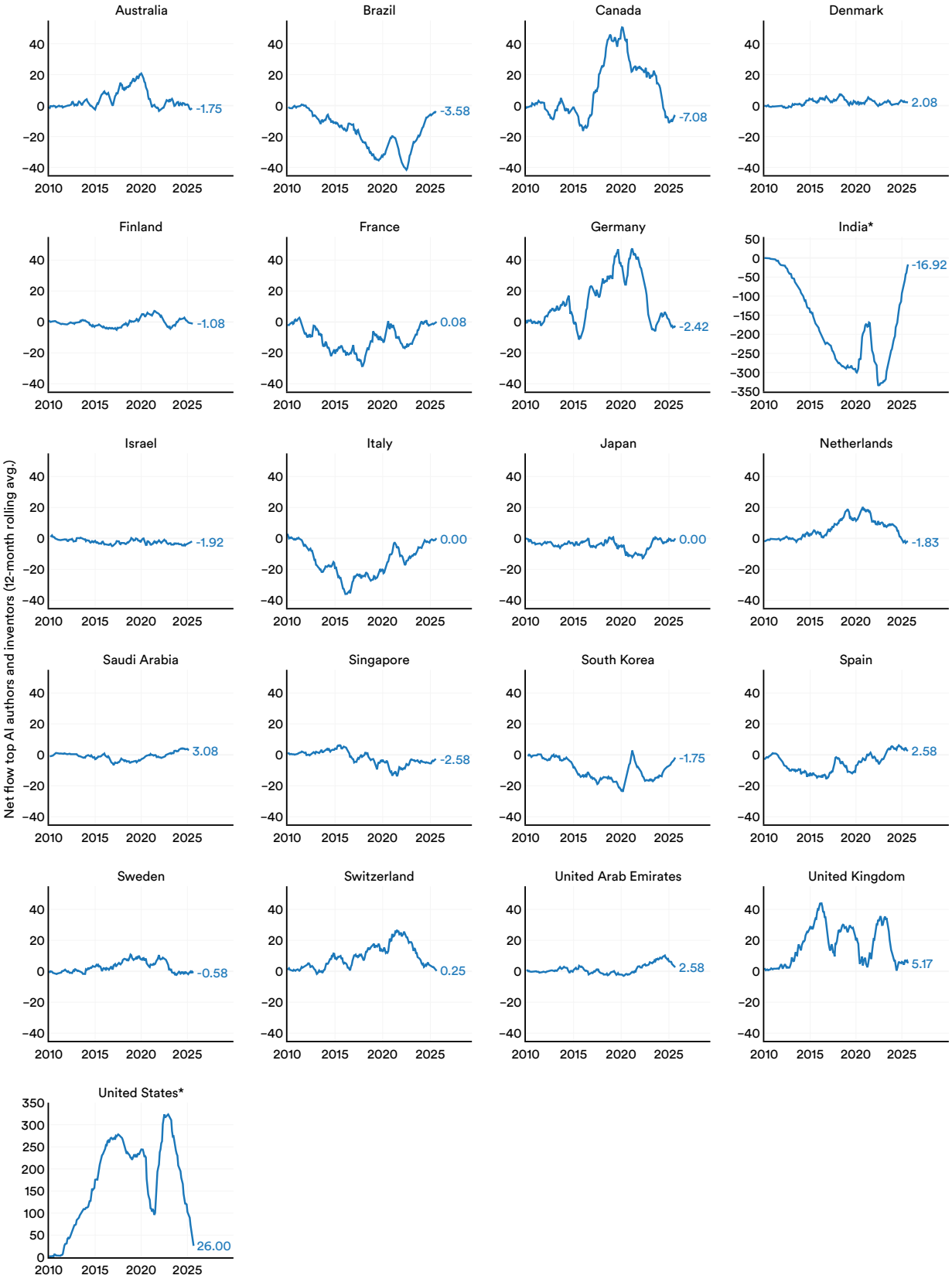


Figure 1.8.6[44]

44 Asterisks indicate that a country's y-axis label is scaled differently than the y-axis label for the other countries.