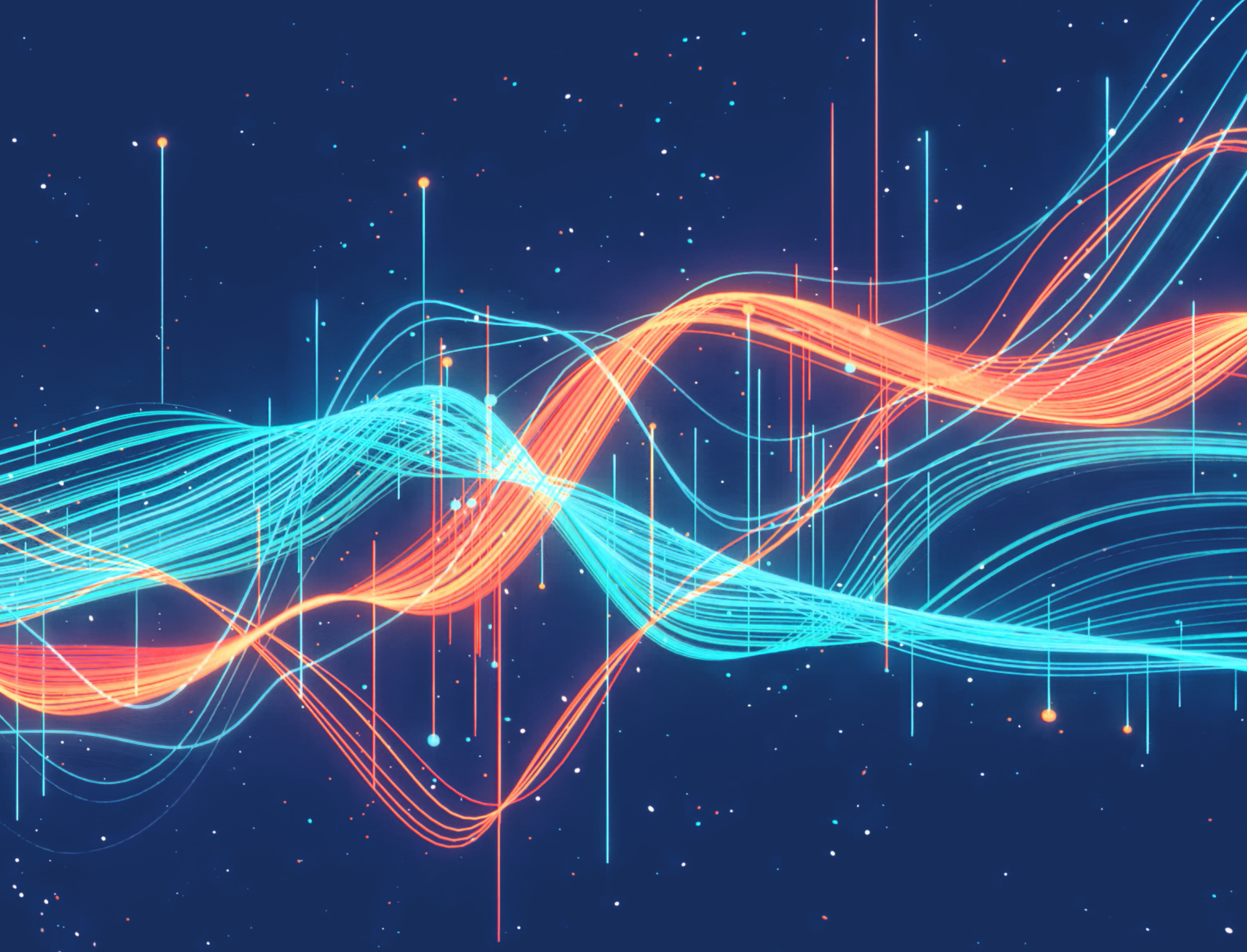
CHAPTER 1:
Research and Development



Chapter 1: Research and Development

ACCESS THE PUBLIC DATA

CHAPTER 1:
Research and Development

Overview

This chapter explores trends in AI research and development, beginning with an analysis of AI publications, patents, and notable AI systems. These topics are examined through the lens of the countries, organizations, and sectors producing them. The chapter also covers AI model training costs, AI conference attendance, and open-source AI software. New additions this year include profiles of the evolving AI hardware ecosystem, an assessment of AI training's energy requirements and environmental impact, and a temporal analysis of model inference costs.

CHAPTER 1:
Research and Development

Chapter Highlights

1. Industry continues to make significant investments in AI and leads in notable AI model development, while academia leads in highly cited research. Industry's lead in notable model development, highlighted in the two previous AI Index reports, has only grown more pronounced, with nearly 90% of notable models in 2024 (compared to 60% in 2023) originating from industry. Academia has remained the single leading institutional producer of highly cited (top 100) publications over the past three years.

2. China leads in AI research publication totals, while the United States leads in highly influential research. In 2023, China produced more AI publications (23.2%) and citations (22.6%) than any other country. Over the past three years, U.S. institutions have contributed the most top-100-cited AI publications.

3. AI publication totals continue to grow and increasingly dominate computer science. Between 2013 and 2023, the total number of AI publications in venues related to computer science and other scientific disciplines nearly tripled, increasing from approximately 102,000 to over 242,000. Proportionally, AI's share of computer science publications has risen from 21.6% in 2013 to 41.8% in 2023.

4. The United States continues to be the leading source of notable AI models. In 2024, U.S.-based institutions produced 40 notable AI models, significantly surpassing China's 15 and Europe's combined total of three. In the past decade, more notable machine learning models have originated from the United States than any other country.

5. AI models get increasingly bigger, more computationally demanding, and more energy intensive. New research finds that the training compute for notable AI models doubles approximately every five months, dataset sizes for training LLMs every eight months, and the power required for training annually. Large-scale industry investment continues to drive model scaling and performance gains.

Chapter Highlights (cont'd)

6. AI models become increasingly affordable to use. The cost of querying an AI model that scores the equivalent of GPT-3.5 (64.8) on MMLU, a popular benchmark for assessing language model performance, dropped from $20.00 per million tokens in November 2022 to just $0.07 per million tokens by October 2024 (Gemini-1.5-Flash-8B)—a more than 280-fold reduction in approximately 18 months. Depending on the task, LLM inference prices have fallen anywhere from 9 to 900 times per year.

7. AI patenting is on the rise. Between 2010 and 2023, the number of AI patents has grown steadily and significantly, ballooning from 3,833 to 122,511. In just the last year, the number of AI patents has risen 29.6%. As of 2023, China leads in total AI patents, accounting for 69.7% of all grants, while South Korea and Luxembourg stand out as top AI patent producers on a per capita basis.

8. AI hardware gets faster, cheaper, and more energy efficient. New research suggests that machine learning hardware performance, measured in 16-bit floating-point operations, has grown 43% annually, doubling every 1.9 years. Price performance has improved, with costs dropping 30% per year, while energy efficiency has increased by 40% annually.

9. Carbon emissions from AI training are steadily increasing. Training early AI models, such as AlexNet (2012), had modest amounts of carbon emissions at 0.01 tons. More recent models have significantly higher emissions for training: GPT-3 (2020) at 588 tons, GPT-4 (2023) at 5,184 tons, and Llama 3.1 405B (2024) at 8,930 tons. For perspective, the average American emits 18 tons of carbon per year.

1.1 Publications

The figures below show the global count of English-language AI publications from 2010 to 2023, categorized by affiliation type, publication type, and region. New to this year's report, the AI Index includes a section analyzing trends among the 100 most-cited AI publications, which can offer insights into particularly high-impact research. This year, the AI Index analyzed AI publication trends using the OpenAlex database. As a result, the numbers in this year's report differ slightly from those in previous editions.[1] Given that there is a significant lag in the collection of publication metadata, and that in some cases it takes until the middle of any given year to fully capture the previous year's publications, in this year's report, the AI Index team elected to examine publication trends only through 2023.

Overview

The following section reports on trends in the total number of English-language AI publications.

Total Number of AI Publications

Figure 1.1.1 displays the global count of AI publications. These are the publications with a computer science (CS) label in the OpenAlex catalog that were classified by the AI Index as being related to AI.[2] Between 2013 and 2023, the total number of AI

Number of AI publications in CS worldwide, 2013–23
Source: AI Index, 2025 | Chart: 2025 AI Index report

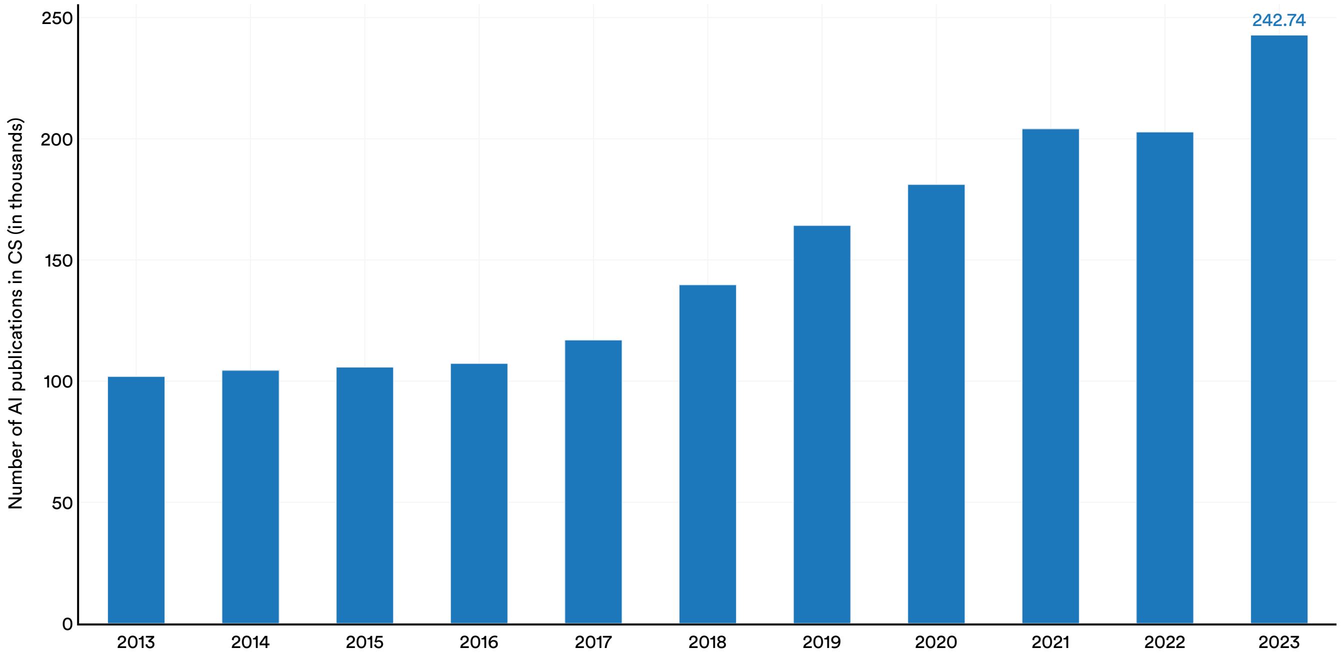


Figure 1.1.1

[1] 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. In previous years, the Index relied on third-party providers with different underlying data sources and classification methods. As a result, this year's findings differ slightly from those included in previous reports. Additionally, the AI Index applied the classifier only to papers that OpenAlex categorized under the broad field of computer science. This approach may have led to an undercount of AI-related publications by excluding research from fields like social sciences that employ AI methodologies but fall outside the computer science–designated classification.

[2] 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, 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.

publications more than doubled, rising from approximately 102,000 in 2013 to more than 242,000 in 2023. The increase over the last year was a meaningful 19.7%. Many fields within computer science, from hardware and software engineering to human-computer interaction, are now contributing to AI. As a result, the observed growth reflects a broader and increased interest in AI across the discipline.

AI publications in CS (% of total) worldwide, 2013–23

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

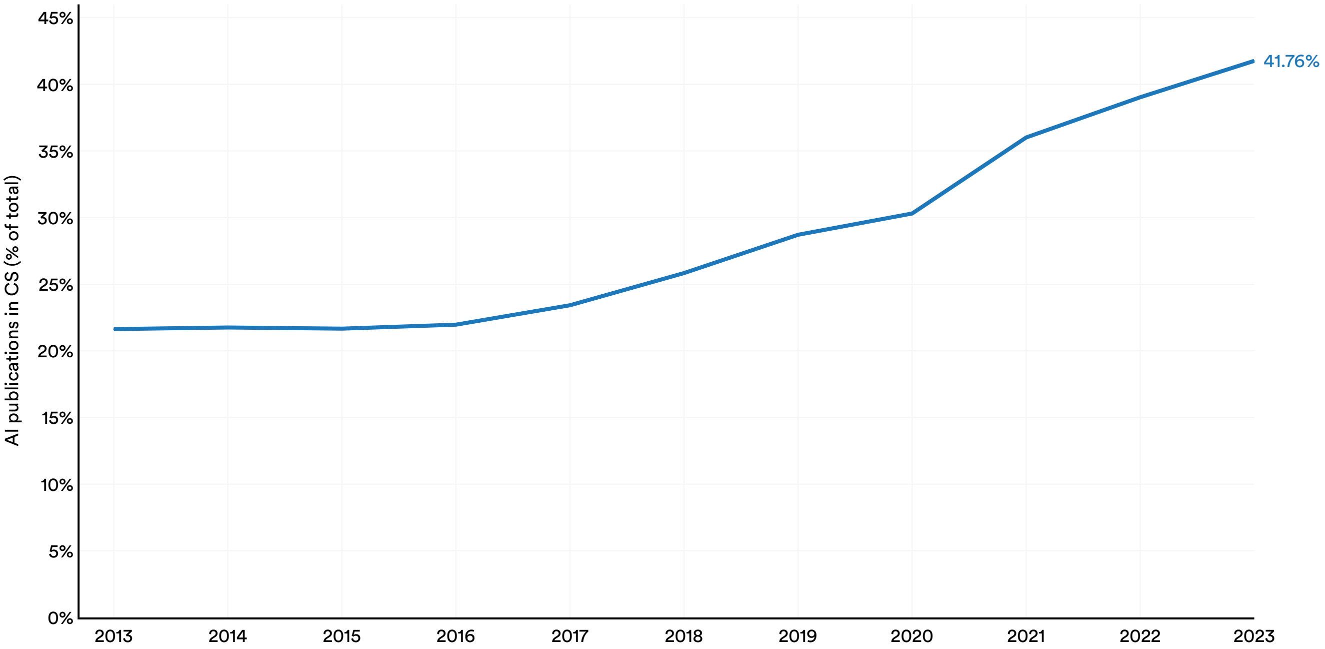


Figure 1.1.2

Figure 1.1.2 shows the proportion of computer science publications in the OpenAlex corpus classified as AI-related. Figure 1.1.2 features the same data included in Figure 1.1.1 but in a proportional form. The share of AI publications has grown significantly, almost doubling from 2013 to 2023.

By Venue

AI researchers publish their work across various venues. Figure 1.1.3 visualizes the total number of AI publications by venue type. In 2023, journals accounted for the largest share of AI publications (41.8%), followed by conferences (34.3%). Even though the total number of journal and conference publications has increased since 2013, the share of AI publications in journals and conferences has steadily declined, from 52.6% and 36.4% in 2013 to 41.8% and 34.3%, respectively, in 2023. Conversely, AI publications in repositories like arXiv have seen a growing share.

AI publications in CS (% of total) worldwide, 2013–23

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

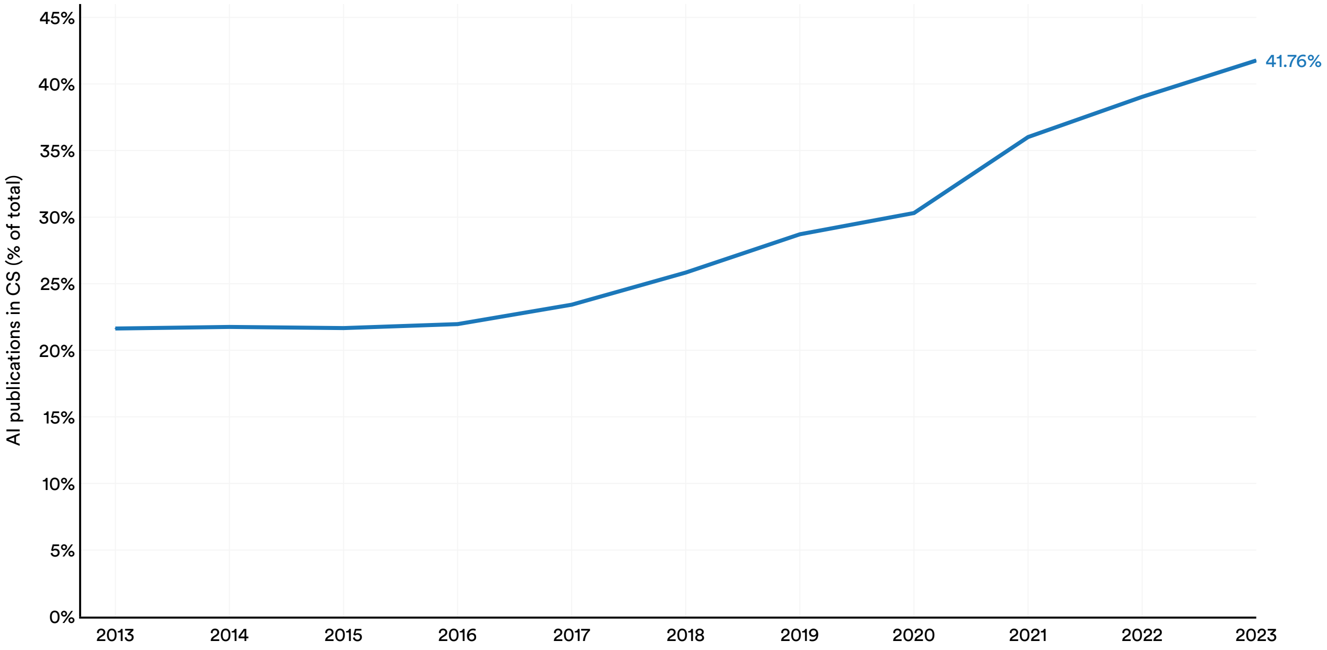


Figure 1.1.3

By National Affiliation

Figure 1.1.4 visualizes AI publications over time by region.[3] In 2023, East Asia and the Pacific led AI research output, accounting for 34.5% of all AI publications, followed by Europe and Central Asia (18.2%) and North America (10.3%).[4]

While Figure 1.1.4 examines the geographic distribution of AI publications, identifying which regions produce the most research, Figure 1.1.5 focuses on citations, measuring the share of total AI publication citations attributed to work originating from each region. As of 2023, AI publications from East Asia and the Pacific accounted for the largest share of AI article citations at 37.1% (Figure 1.1.5). In 2017, citation shares from East Asia and the Pacific and North America were roughly equal, but since then, North American and European citation shares have declined, while East Asia and the Pacific's share has risen sharply.

AI publications in CS (% of total) by region, 2013–23

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

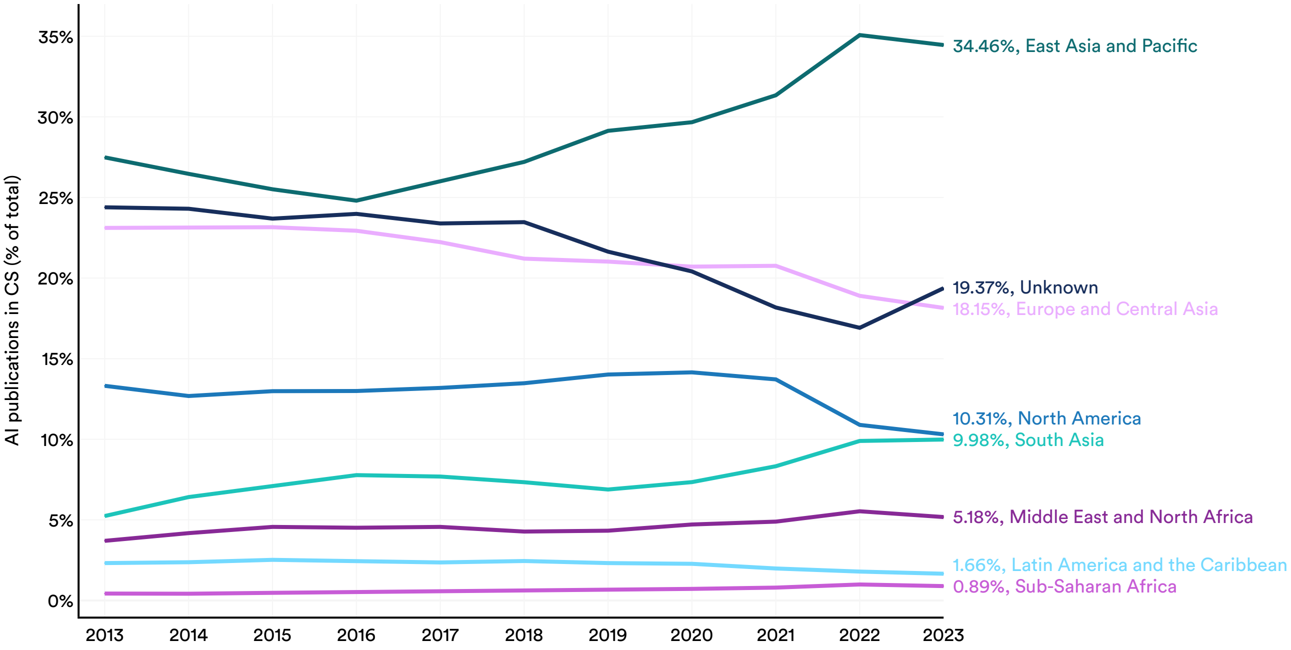


Figure 1.1.4

[3] Regions in this chapter are classified according to the World Bank analytical grouping. The AI Index determines the country affiliation of authors using the "countries" field from the authorship data. This field lists all the countries an author is affiliated with, 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 U.S. and one in China, the publication is counted once for the U.S. and once for China.

[4] 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 publication citations in CS (% of total) by region, 2013–23

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

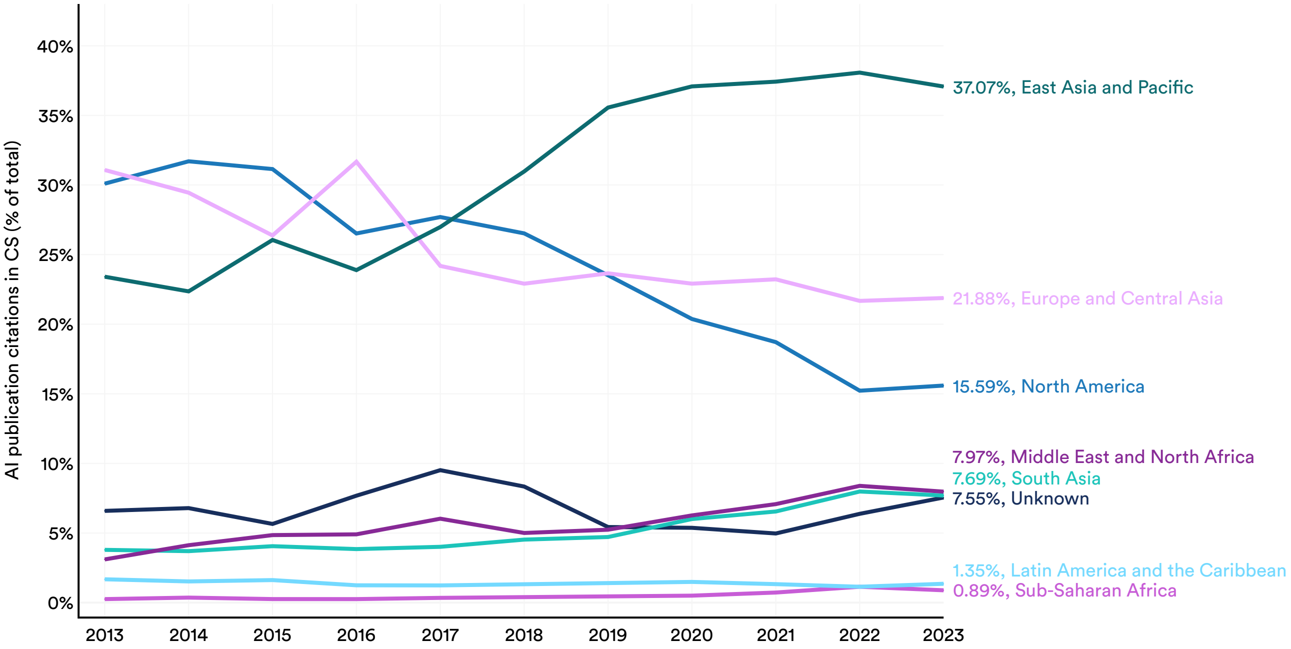


Figure 1.1.5

In 2023, China was the global leader in AI article publications, accounting for 23.2% of the total, compared to 15.2% from Europe and 9.2% from India (Figure 1.1.6).[5] Since 2016, China's share has steadily increased, while the proportion attributed to Europe has declined. AI publications attributed to the United States remained relatively stable until 2021 but have shown a slight decline since then.

AI publications in CS (% of total) by select geographic areas, 2013–23

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

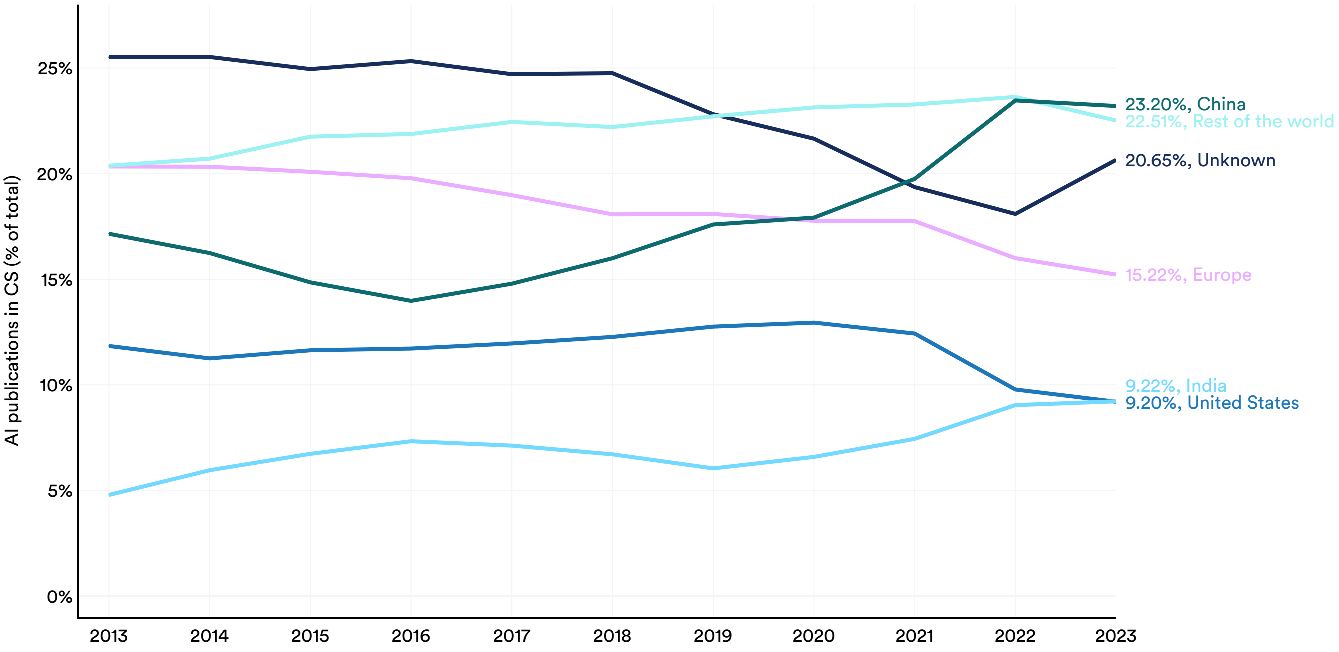


Figure 1.1.6[6]

[5] For the "Europe" designation in this and other chapters of the report, the AI Index follows the list of countries defined by the United Nations Statistics Division.

[6] To maintain concision, the AI Index visualized results for a select group of countries. However, full results for all countries will be available on the AI Index's Global Vibrancy Tool, which is set to be updated in summer 2025. For immediate access to country-specific research and development data, please contact the AI Index team.

In 2023, Chinese AI publications accounted for 22.6% of all AI citations, followed by Europe at 20.9% and the United States at 13.0% (Figure 1.1.7). As with total AI publications, the late 2010s marked a turning point when China surpassed Europe and the U.S. as the leading source of AI publication citations.

AI publication citations in CS (% of total) by select geographic areas, 2013–23

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

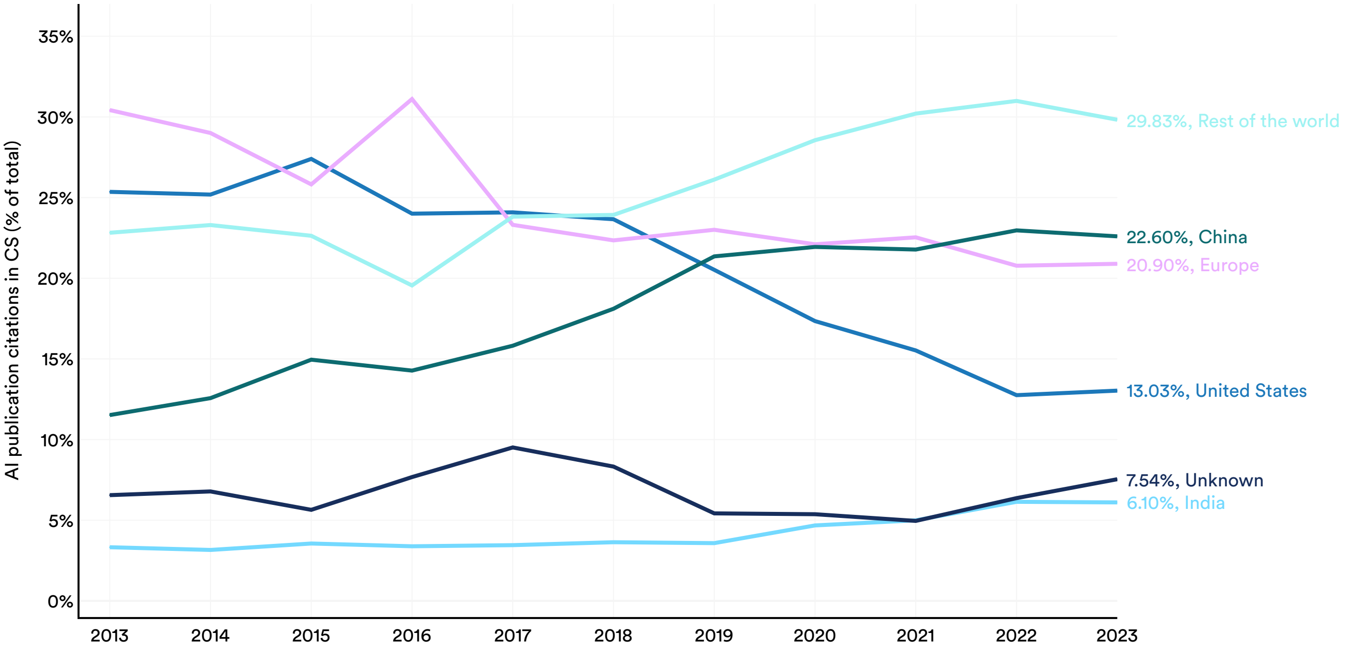


Figure 1.1.7

By Sector

Academic institutions remain the primary source of AI publications worldwide (Figure 1.1.8). In 2013, they accounted for 85.9% of all AI publications, a figure that remained high, at 84.9%, in 2023. Industry contributed 7.1% of AI publications in 2023, followed by government institutions at 4.9% and nonprofit organizations at 1.7%.

AI publications in CS (% of total) by sector, 2013–23

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

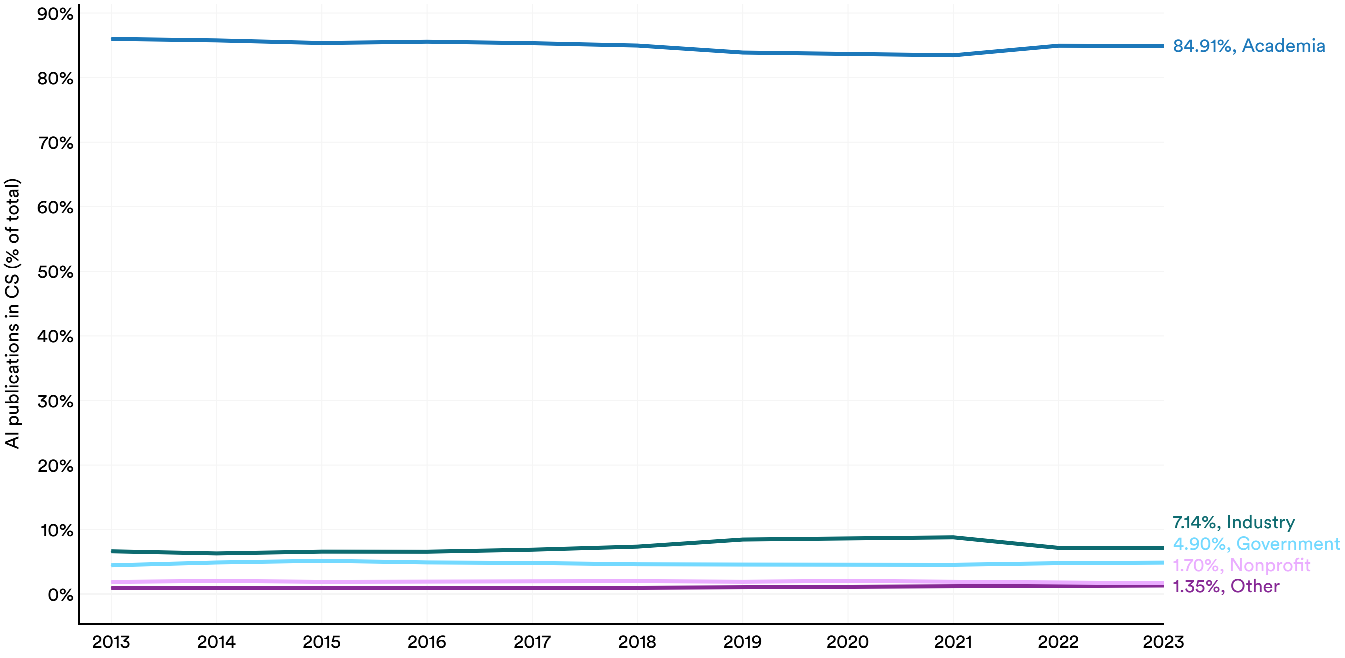


Figure 1.1.8[7]

[7] For Figures 1.1.8 and 1.1.9, publications with unknown affiliations were excluded from the final visualization.

AI publications emerge from various sectors in differing proportions across geographic regions. In the United States, a higher share of AI publications (16.5%) comes from industry compared to China (8.0%) (Figure 1.1.9). Among major geographic areas, China has the highest percentage of AI publications originating from the education sector (84.5%).

AI publications in CS (% of total) by sector and select geographic areas, 2023

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

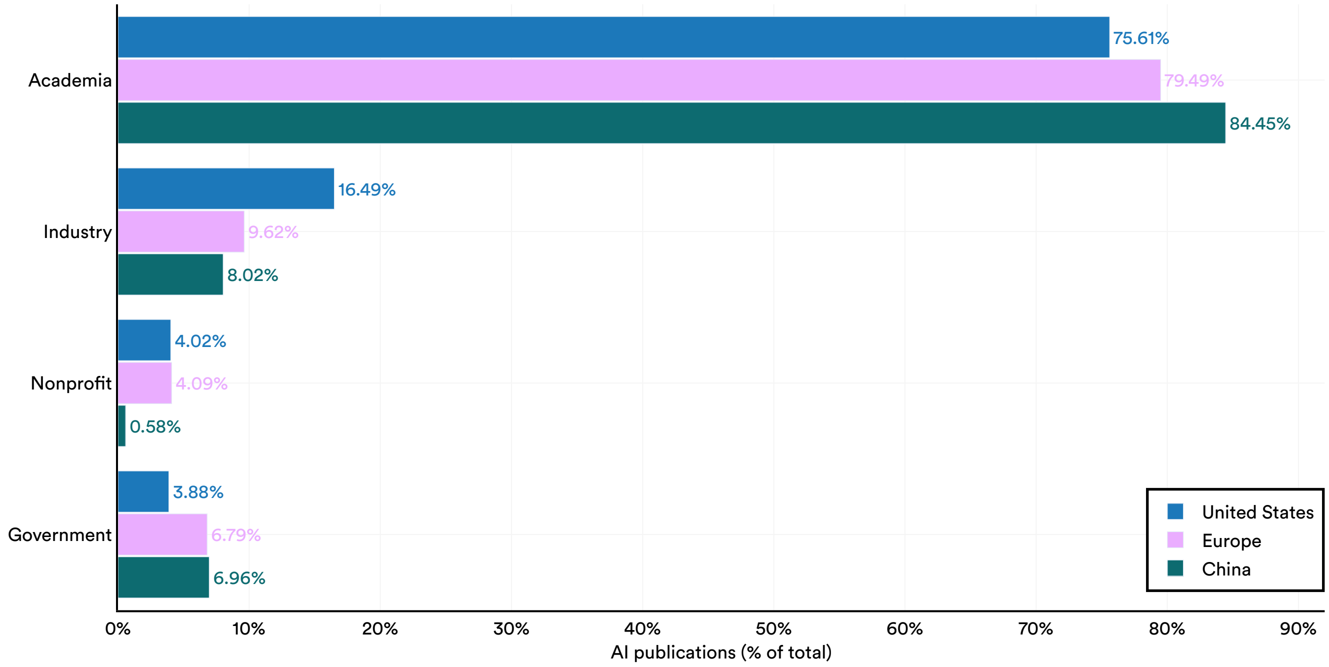


Figure 1.1.9

By Topic

Machine learning was the most prevalent research topic in AI publications in 2023, comprising 75.7% of publications, followed by computer vision (47.2%), pattern recognition (25.9%) and natural language processing (17.1%) (Figure 1.1.10). Over the past year, there has been a sharp increase in publications on generative AI.

Number of AI publications by select top topics, 2013–23

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

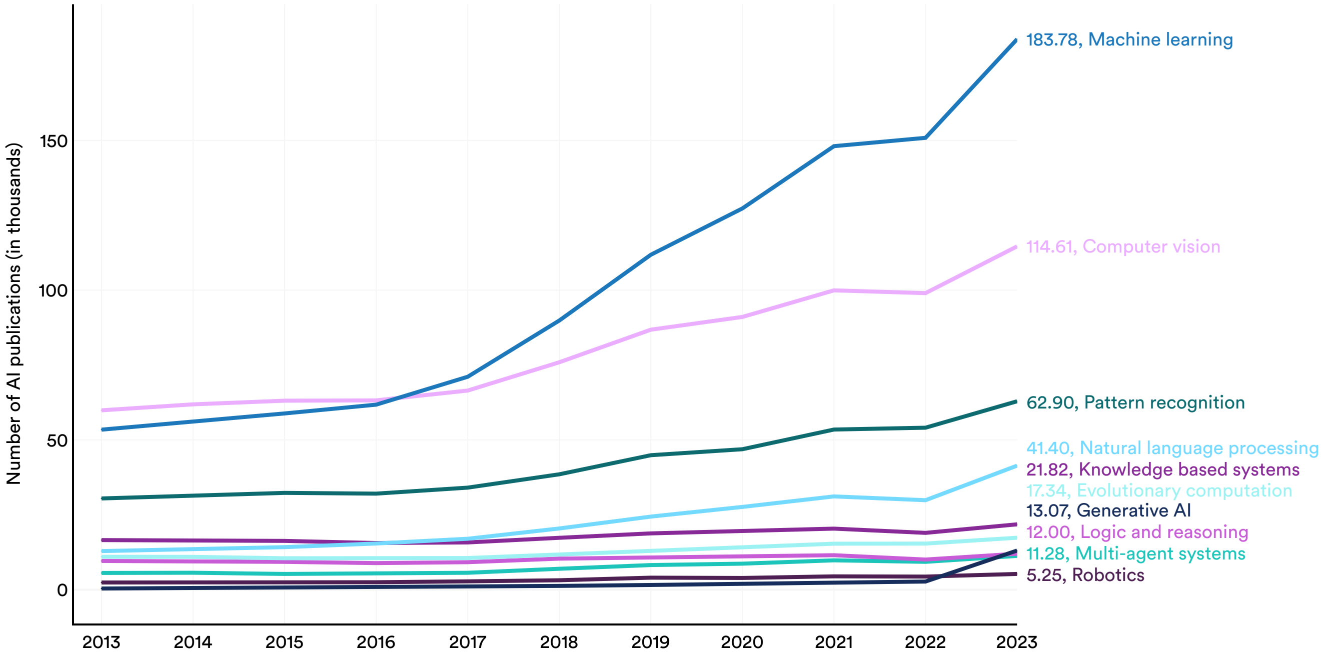


Figure 1.1.10[8]

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

Top 100 Publications

While tracking total AI publications provides a broad view of research activity, focusing on the most-cited papers offers a perspective of the field's most influential work. This analysis sheds light on where some of the most groundbreaking and influential AI research is emerging. This year, the AI Index identified the 100 most-cited AI publications in 2021, 2022, and 2023, using citation data from OpenAlex. This analysis was further supplemented with insights from Google Scholar and Semantic Scholar.[9] Some of the most highly cited AI publications in 2023 included OpenAI's GPT-4 technical report, Meta's Llama 2 technical report, and Google's PaLM-E technical report. It is important to note that due to citation lag, the most-cited papers in this year's report may change in future editions.

By National Affiliation

Figure 1.1.11 illustrates the geographic distribution of the top 100 most-cited AI publications by year. From 2021 to 2023, the U.S. consistently had the highest number of top-cited publications, with 64 in 2021, 59 in 2022, and 50 in 2023.[10] In each of these years, China ranked second. Since 2021, the U.S. share of top AI publications has gradually declined.

Number of highly cited publications in top 100 by select geographic areas, 2021–23

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

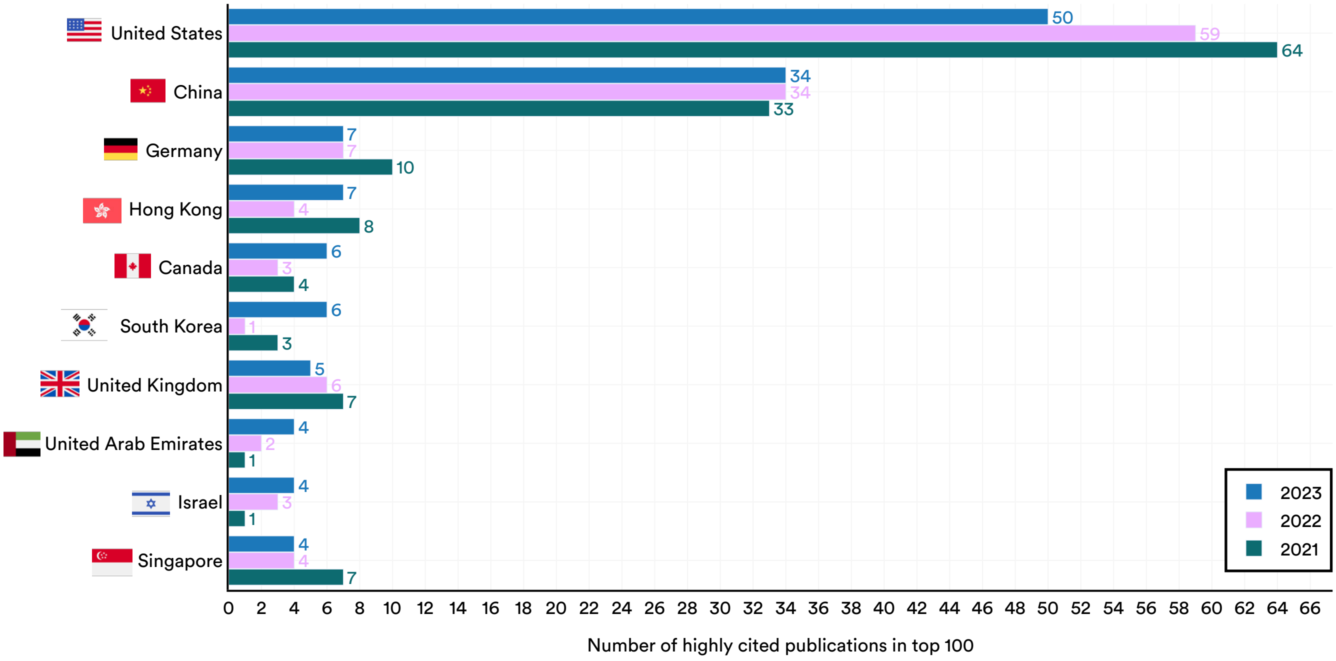


Figure 1.1.11

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

[10] A publication can have multiple authors from different countries or organizations. For example, if a paper includes authors from multiple countries, each country is credited once. As a result, the totals in this section's figures exceed 100.

By Sector

Academia consistently produces the most top-cited AI publications, with 42 in 2023, 27 in 2022, and 34 in 2021 (Figure 1.1.12). Notably, there was a sharp decline in industry contributions, with the number of top 100 publications dropping from 17 in 2021 and 19 in 2022 to just 7 in 2023. As AI research grows more competitive, many industrial AI labs are publishing less frequently or disclosing fewer details about their research in their publications.

Number of highly cited publications in top 100 by sector, 2021–23

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

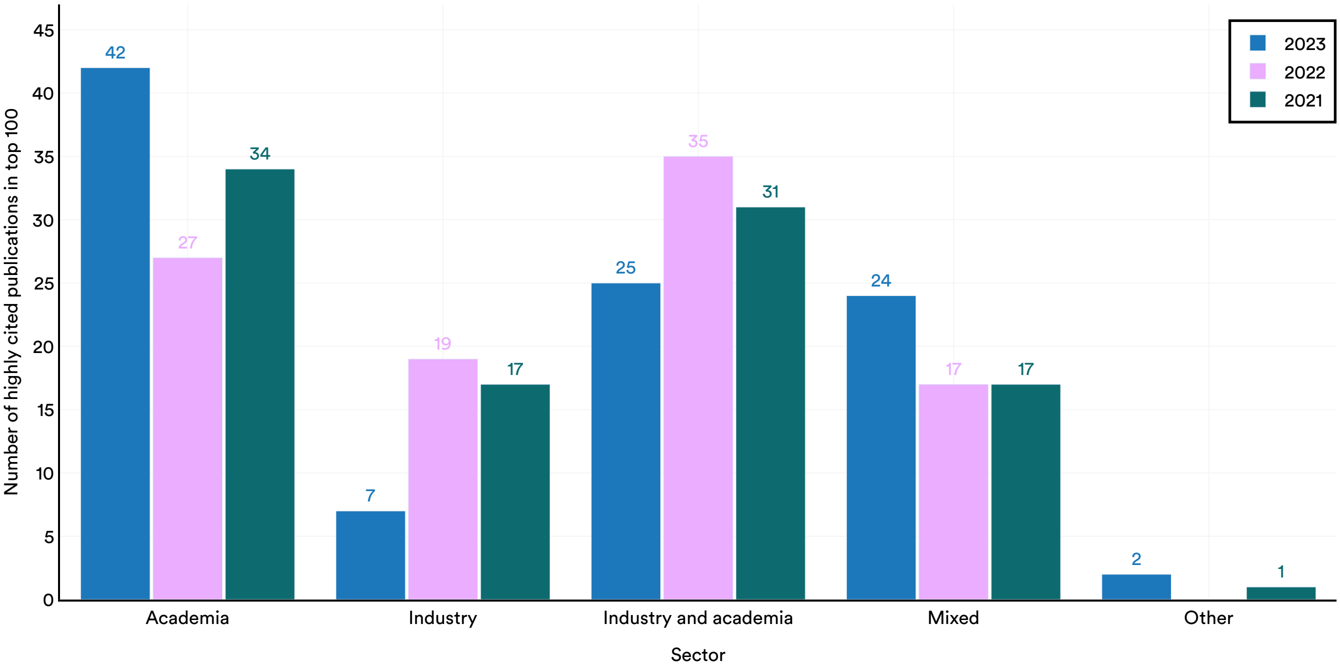


Figure 1.1.12[11]

[11] The "mixed" designation includes all intersector collaborations that are not industry and academia (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 1.1.12 may exceed 100. Also, because two of the papers had authors with an unknown sectoral affiliation, the total sum of publications in Figure 1.1.12 is 98.

By Organization

Figure 1.1.13 highlights the organizations that produced the top 100 most-cited AI publications from 2021 to 2023. Some organizations may have empty bars on the chart if they lacked a top 100 publication in a given year. Additionally, Figure 1.1.13 highlights only the top 10 institutions, though many others contribute significant research.

Google led each year, but it tied with Tsinghua University in 2023, when both contributed eight publications to the top 100. In 2023, Carnegie Mellon University was the highest-ranked U.S. academic institution.

Number of highly cited publications in top 100 by organization, 2021–23

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

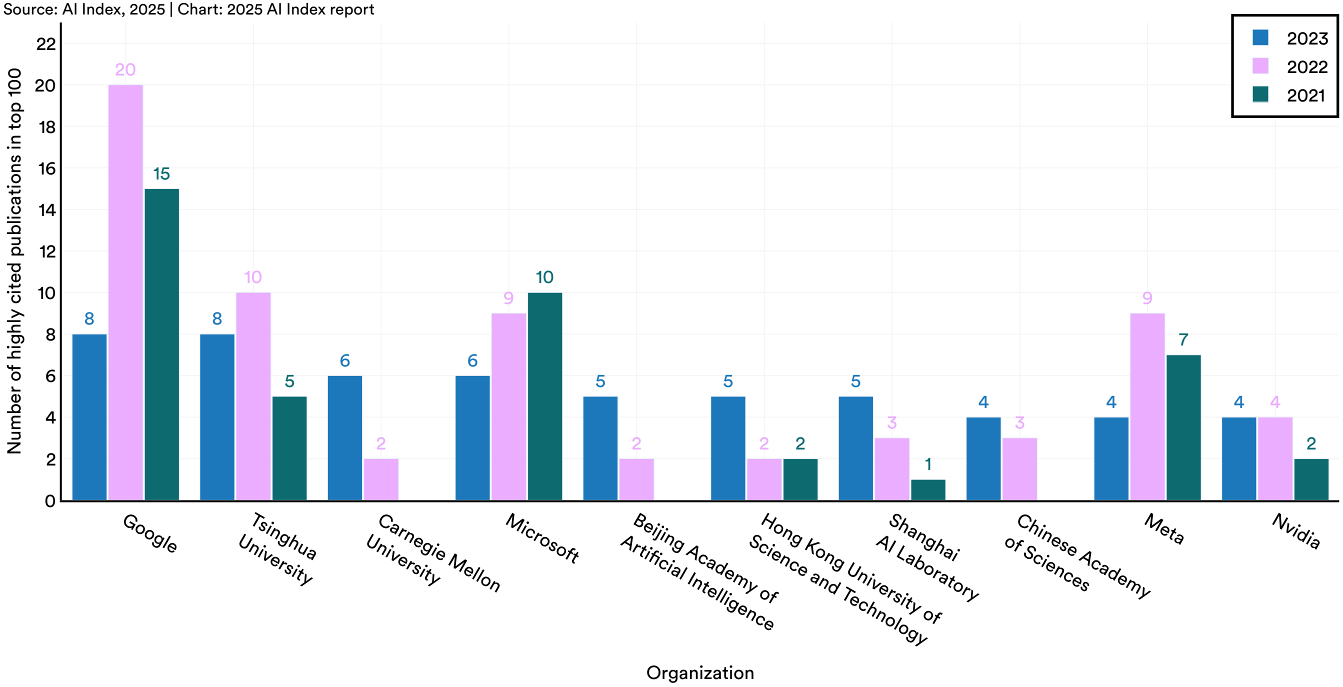


Figure 1.1.13

This section examines trends over time in global AI patents, which can reveal important insights into the evolution of innovation, research, and development within AI. Additionally, analyzing AI patents can reveal how these advances are distributed globally. Similar to the publications data, there is a noticeable delay in AI patent data availability, with 2023 being the most recent year for which data is accessible. The data in this section is sourced from patent-level bibliographic records in PATSTAT Global, a comprehensive database provided by the European Patent Office (EPO).[12]

1.2 Patents

Overview

Figure 1.2.1 examines the global growth in granted AI patents from 2010 to 2023. Over the past dozen years, the number of AI patents has grown steadily and significantly, increasing from 3,833 in 2010 to 122,511 in 2023. In the last year, the number of AI patents has risen 29.6%.

Number of AI patents granted worldwide, 2010–23

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

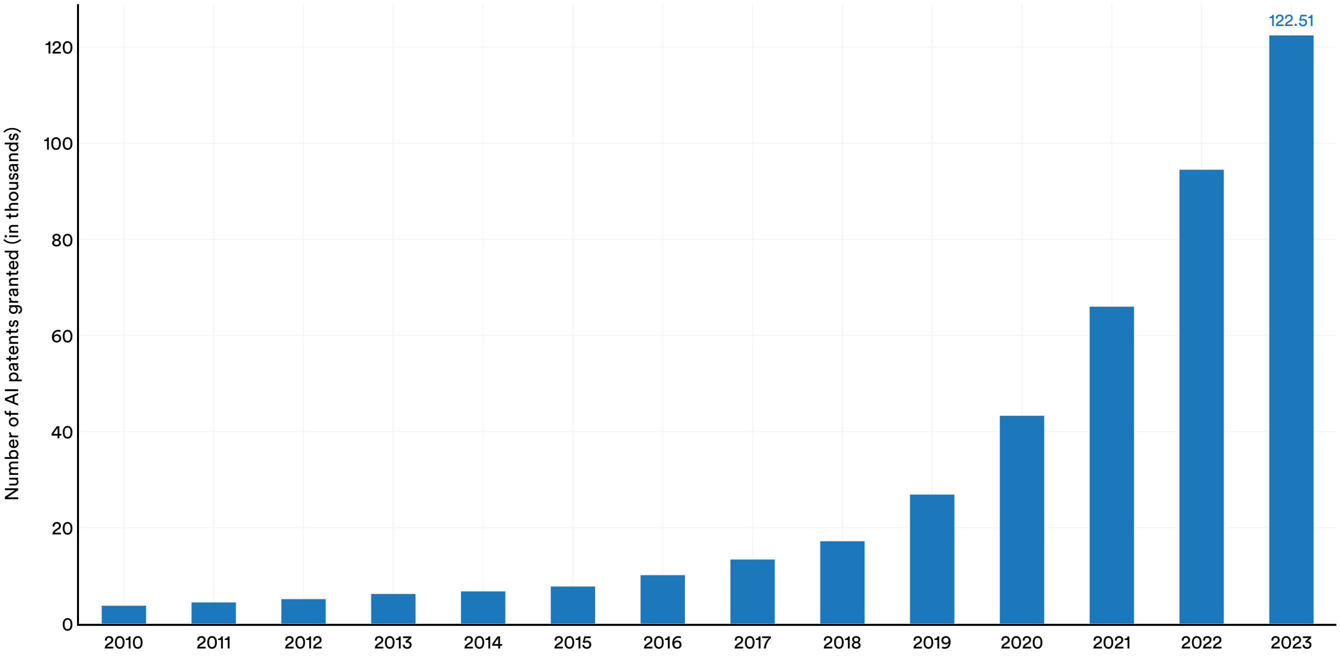


Figure 1.2.1

[12] More details on the methodology behind the patent analysis in this section can be found in the Appendix.

By National Affiliation

Figure 1.2.2 showcases the regional breakdown of granted AI patents, as in the number of patents filed in different regions across the world. As of 2023, the bulk of the world's granted AI patents (82.4%) originated from East Asia and the Pacific, with North America being the next largest contributor at 14.2%. Since 2010, the gap in AI patent grants between East Asia and the Pacific and North America has steadily widened.

Granted AI patents (% of world total) by region, 2010–23
Source: AI Index, 2025 | Chart: 2025 AI Index report

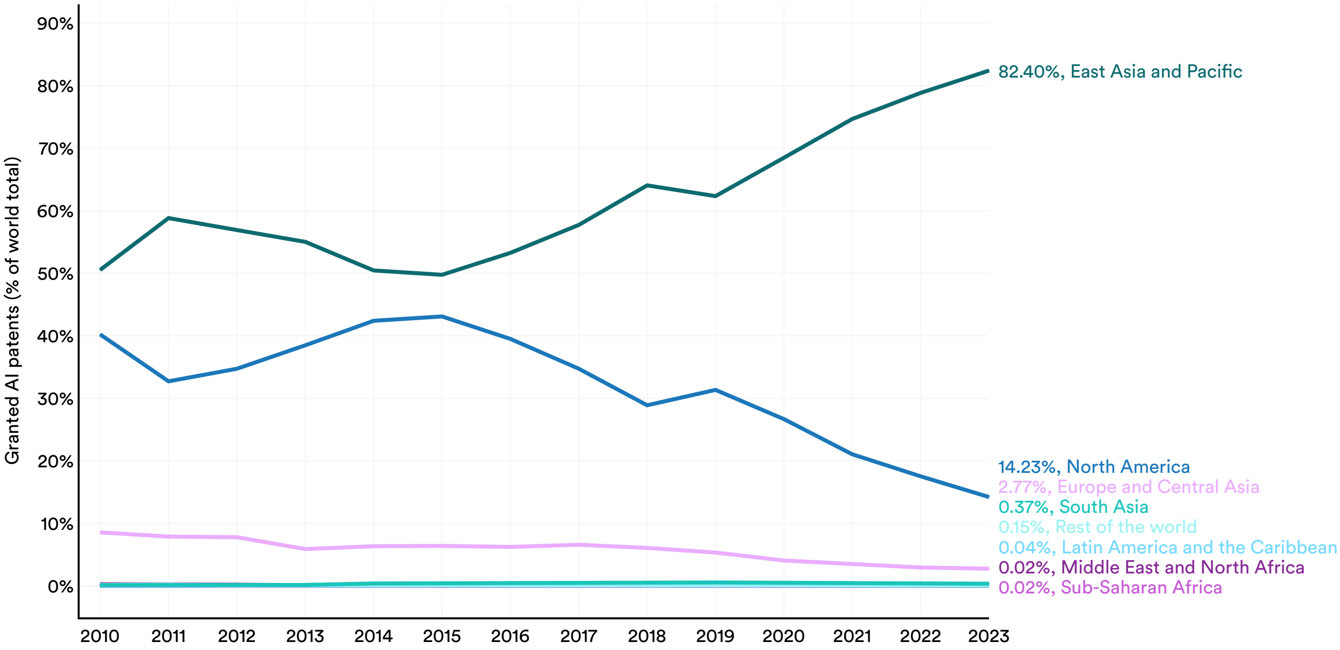


Figure 1.2.2[13]

[13] 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.

Disaggregated by geographic area, the majority of the world's granted AI patents are from China (69.7%) and the United States (14.2%) (Figure 1.2.3). The share of AI patents originating from the United States has declined from a peak of 42.8% in 2015.

Figure 1.2.3 and Figure 1.2.4 document which countries lead in AI patents per capita. In 2023, the country with the most granted AI patents per 100,000 inhabitants was South Korea (17.3), followed by Luxembourg (15.3) and China (6.1) (Figure 1.2.3). Figure 1.2.5 highlights the change in granted AI patents per capita from 2013 to 2023. Luxembourg, China and Sweden experienced the greatest increase in AI patenting per capita during that time period.

Granted AI patents (% of world total) by select geographic areas, 2010–23

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

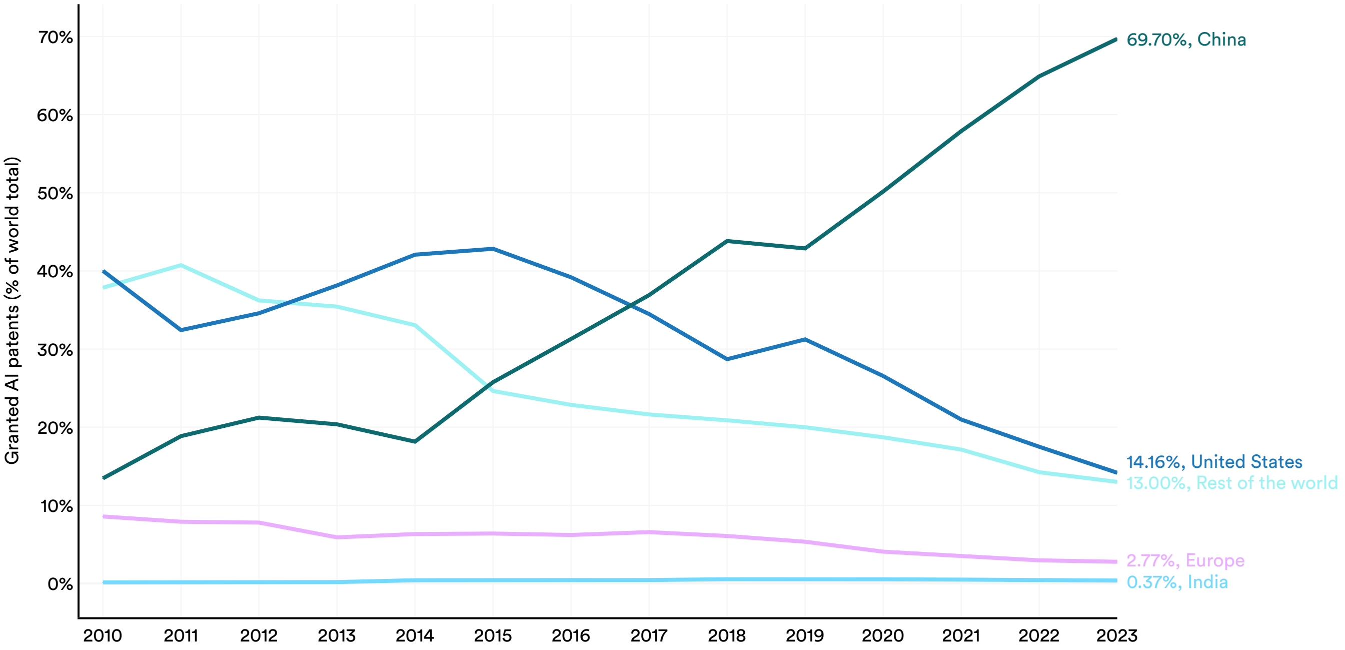


Figure 1.2.3

Granted AI patents per 100,000 inhabitants by country, 2023

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

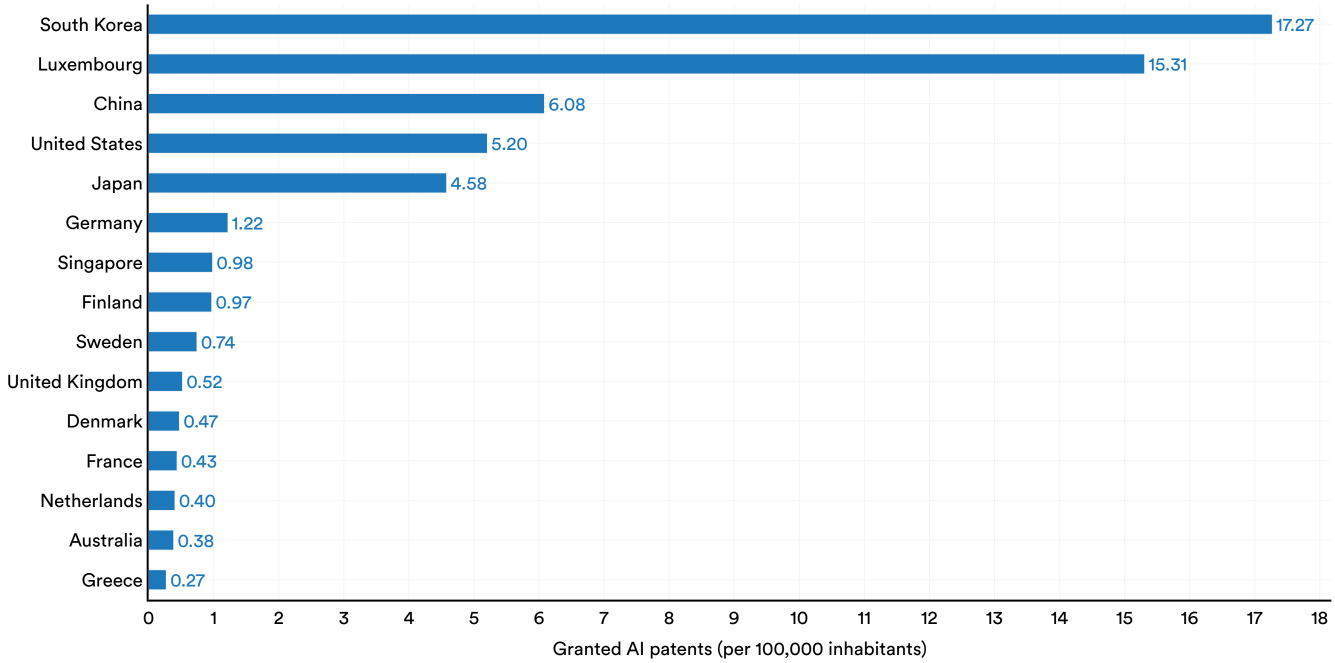


Figure 1.2.4

Percentage change of granted AI patents per 100,000 inhabitants by country, 2013 vs. 2023

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

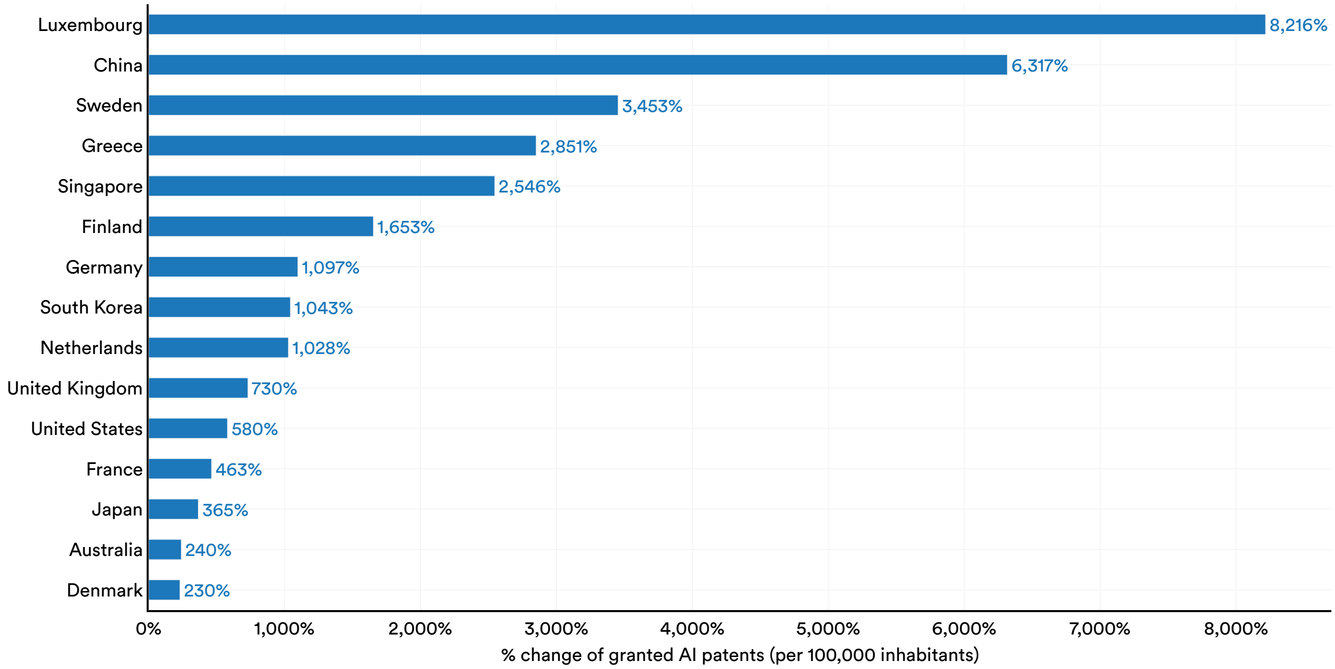


Figure 1.2.5

1.3 Notable AI Models

This section explores notable AI models. Epoch AI, an AI Index data provider, uses the term "notable machine learning models" to designate particularly influential models within the AI/machine learning ecosystem. Epoch maintains a database of 900 AI models released since the 1950s, selecting entries based on criteria such as state-of-the-art advancements, historical significance, or high citation rates. Since Epoch manually curates the data, some models considered notable by some may not be included. Analyzing these models provides a comprehensive overview of the machine learning landscape's evolution, both in recent years and over the past few decades. Some models may be missing from the dataset; however, the dataset can reveal trends in relative terms. Examples of notable AI models include GPT-4o, Claude 3.5, and AlphaGeometry.

Within this section, the AI Index explores trends in notable models from various perspectives, including country of origin, originating organization, gradient of model release, parameter count, and compute usage. The analysis concludes with an examination of machine learning training as well as inference costs.

By National Affiliation

To illustrate the evolving geopolitical landscape of AI, the AI Index shows the country of origin of notable models. Figure 1.3.1 displays the total number of notable AI models attributed to the location of researchers' affiliated institutions.[16] In 2024, the United States led with 40 notable AI models, followed by China with 15 and France with three. All major geographic groups, including the United States, China, the European Union, and the United Kingdom, reported releasing fewer notable models in 2024 than in the previous year (Figure 1.3.2). Since 2003, the United States has produced more models than other major countries such as the United Kingdom, China, and Canada (Figure 1.3.3).

It is difficult to pinpoint the exact cause of the decline in total model releases, but it may stem from a combination of factors: increasingly large training runs, the growing complexity of AI technology, and the heightened challenge of

Number of notable AI models by select geographic areas, 2024
Source: Epoch AI, 2025 | Chart: 2025 AI Index report

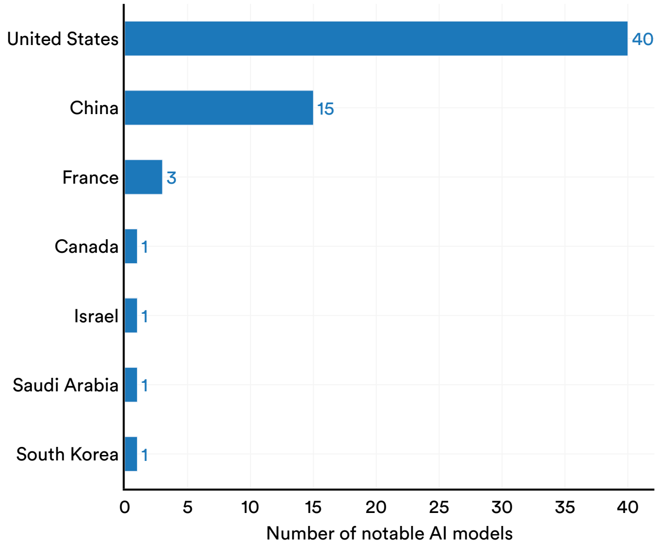


Figure 1.3.1[17]

Number of notable AI models by select geographic areas, 2003–24
Source: Epoch AI, 2025 | Chart: 2025 AI Index report

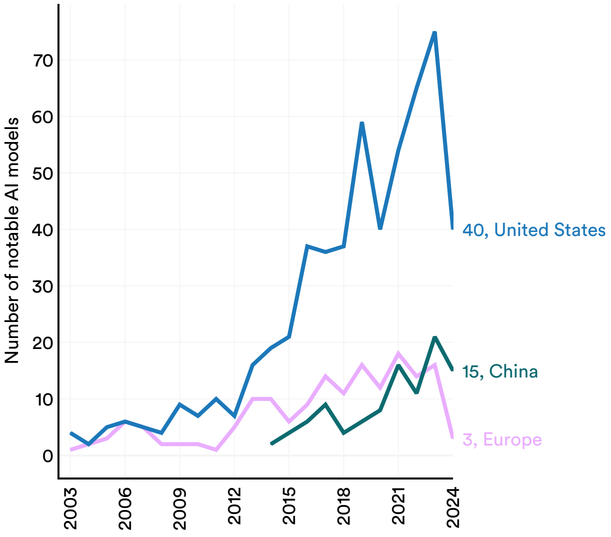


Figure 1.3.2

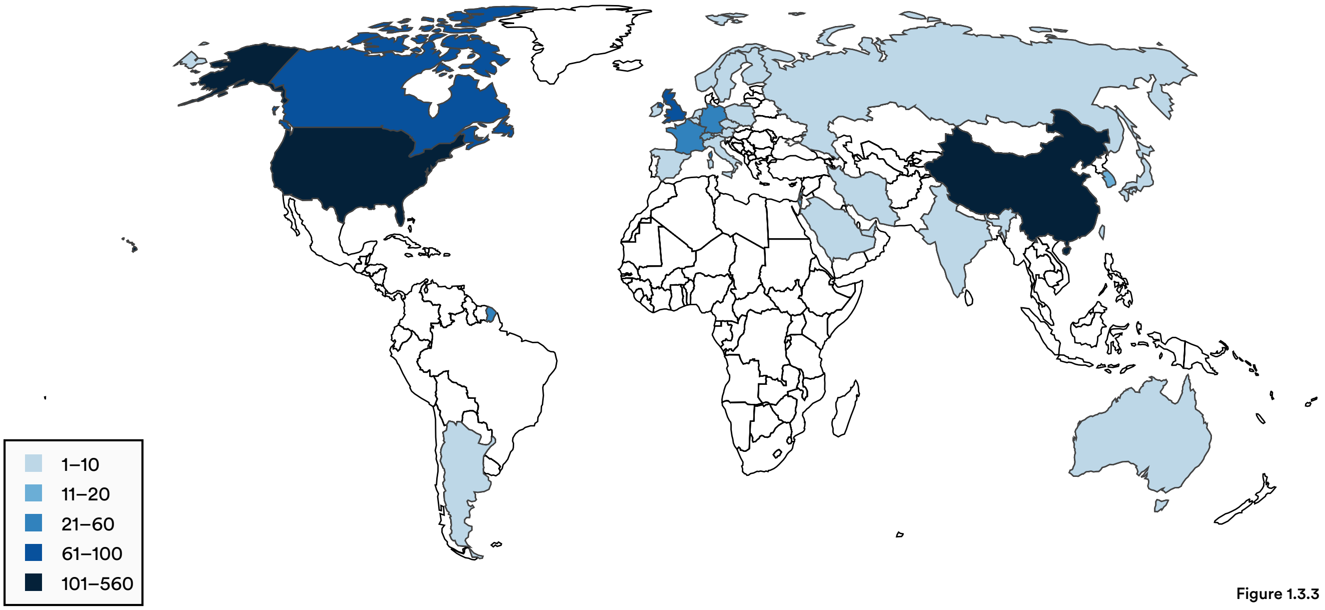
[14] "AI system" refers to a computer program or product based on AI, such as ChatGPT. "AI model" includes a collection of parameters whose values are learned during training, such as GPT-4

[15] 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 from a snapshot taken on March 17, 2025.

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

[17] 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 release.

developing new modeling approaches. Epoch AI's curation of notable models may overlook releases from certain countries that receive less coverage. The AI Index, in cooperation with Epoch, is committed to improving global representation in the AI model ecosystem. If readers believe that models from specific countries are missing, they are encouraged to contact the AI Index team, which will work to address the issue.



Number of notable AI models by geographic area, 2003–24 (sum)

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

Figure 1.3.3

By Sector

Figure 1.3.4 illustrates the sectoral origin of notable AI releases by the year the models were released. Epoch categorizes models based on their source: Industry includes companies such as Google, Meta, and OpenAI; academia covers universities like Tsinghua, MIT, and Oxford; government refers to state-affiliated research institutes like the UK's Alan Turing Institute for AI and Abu Dhabi's Technology Innovation Institute; and research collectives encompass nonprofit AI research organizations such as the Allen Institute for AI and the Fraunhofer Institute.

Until 2014, academia led in terms of releasing machine learning models. Since then, industry has taken the lead. According to Epoch AI, in 2024, industry produced 55 notable AI models. That same year, Epoch AI identified no notable AI models originating from academia (Figure 1.3.5).[18] Over time, industry-academia collaborations have contributed to a growing number of models. The proportion of notable AI models originating from industry has steadily increased over the past decade, growing to 90.2% in 2024.

[18] This figure should be interpreted with caution. A count of zero academic models does not mean that no notable models were produced by academic institutions in 2023, but rather that Epoch AI has not identified any as notable. Additionally, academic publications often take longer to gain recognition, as highly cited papers introducing significant architectures may take years to achieve prominence.

Number of notable AI models by sector, 2003–24

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

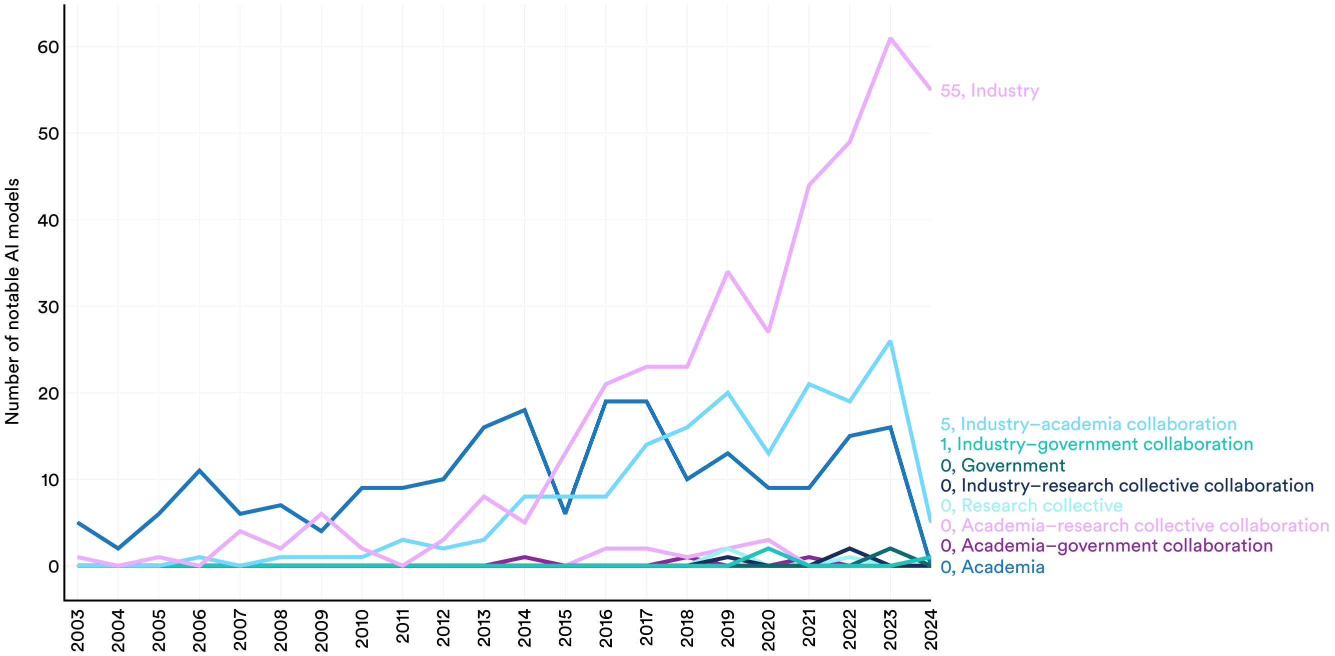


Figure 1.3.4

Notable AI models (% of total) by sector, 2003–24

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

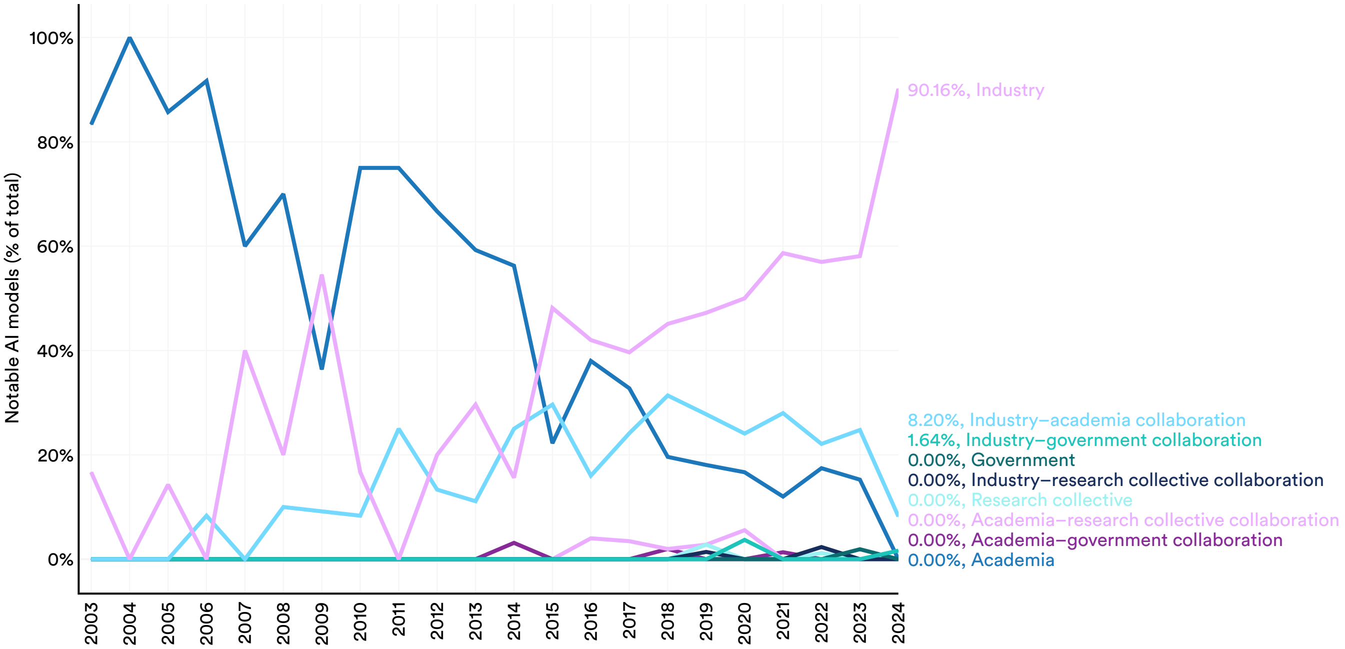


Figure 1.3.5

By Organization

Figure 1.3.6 and Figure 1.3.7 highlight the organizations leading in the production of notable machine learning models in 2024 and over the past decade. In 2024, the top contributors were OpenAI (7 models), Google (6), and Alibaba (4). Since 2014, Google has led with 186 notable models, followed by Meta (82) and Microsoft (39). Among academic institutions, Carnegie Mellon University (25), Stanford University (25), and Tsinghua University (22) have been the most prolific since 2014.

Number of notable AI models by organization, 2024
Source: Epoch AI, 2025 | Chart: 2025 AI Index report

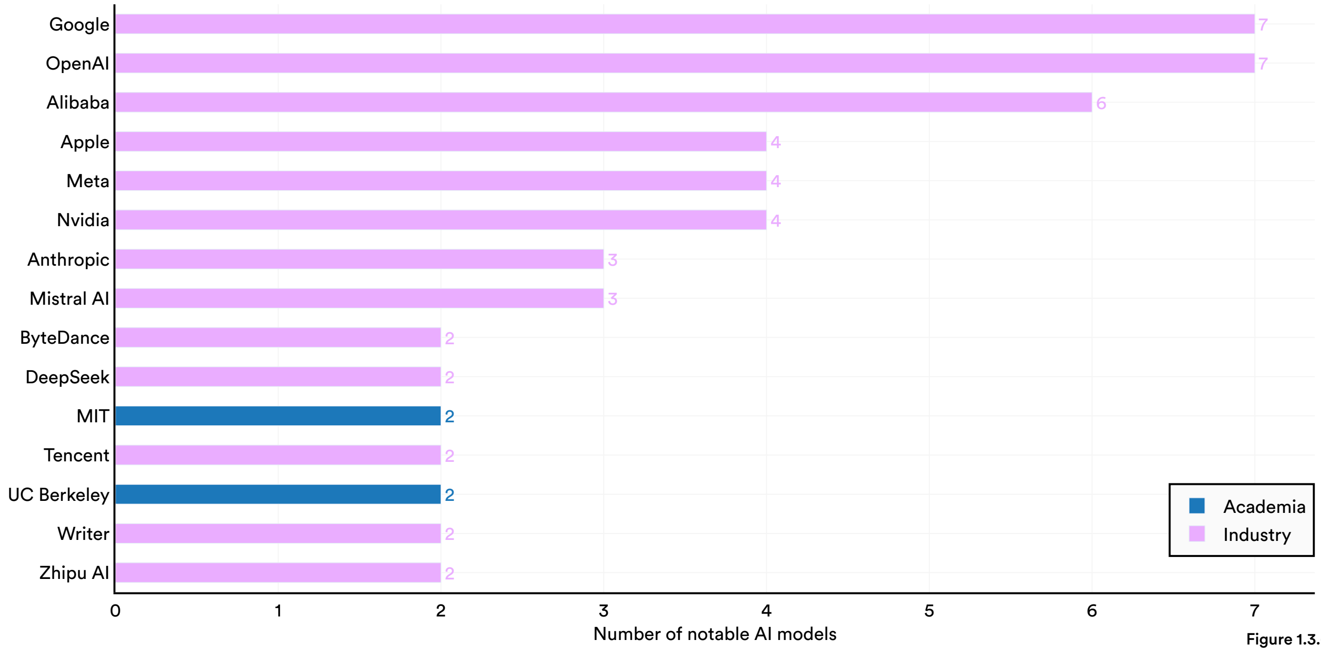


Figure 1.3.6[19]

Number of notable AI models by organization, 2014–24 (sum)
Source: Epoch AI, 2025 | Chart: 2025 AI Index report

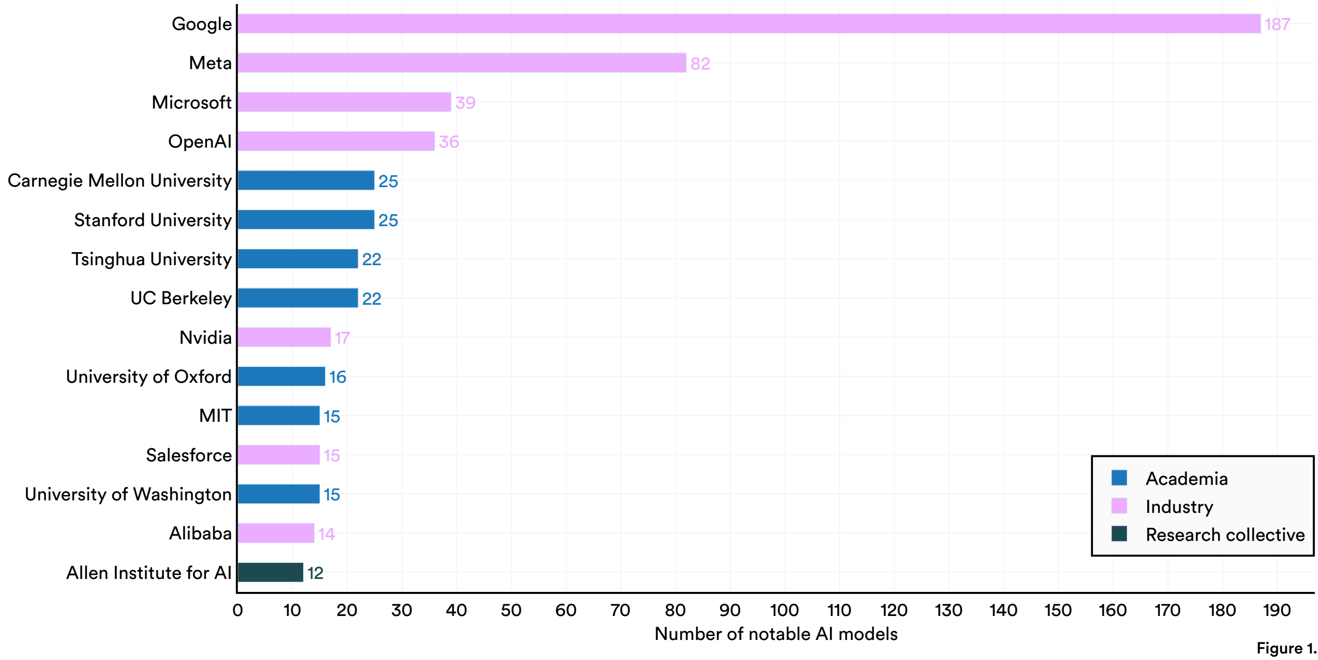


Figure 1.3.7

[19] In the organizational tally figures, research published by DeepMind is classified under Google.

Model Release

Machine learning models are released under various access types, each with varying levels of openness and usability. API access models, like OpenAI's o1, allow users to interact with models via queries without direct access to their underlying weights. Open weights (restricted use) models, like DeepSeek's-V3, provide access to their weights but impose limitations, such as prohibiting commercial use or redistribution. Hosted access (no API) models, like Gemini 2.0 Pro, refer to models available through a platform interface but without programmatic access. Open weights (unrestricted) models, like AlphaGeometry, are fully open, allowing free use, modification, and redistribution. Open weights (noncommercial) models, like Mistral Large 2, share their weights but restrict use to research or noncommercial purposes. Lastly, unreleased models, like ESM3 98B, remain proprietary, accessible only to their developers or select partners. The unknown designation refers to models that have unclear or undisclosed access types.

Figure 1.3.8 illustrates the different access types under which models have been released.[20] In 2024, API access was the most common release type, with 20 of 61 models made available this way, followed by open weights with restricted use and unreleased models.

Figure 1.3.9 visualizes machine learning model access types over time from a proportional perspective. In 2024, most AI models were released via API access (32.8%), which has seen a steady rise since 2020.

Number of notable AI models by access type, 2014–24

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

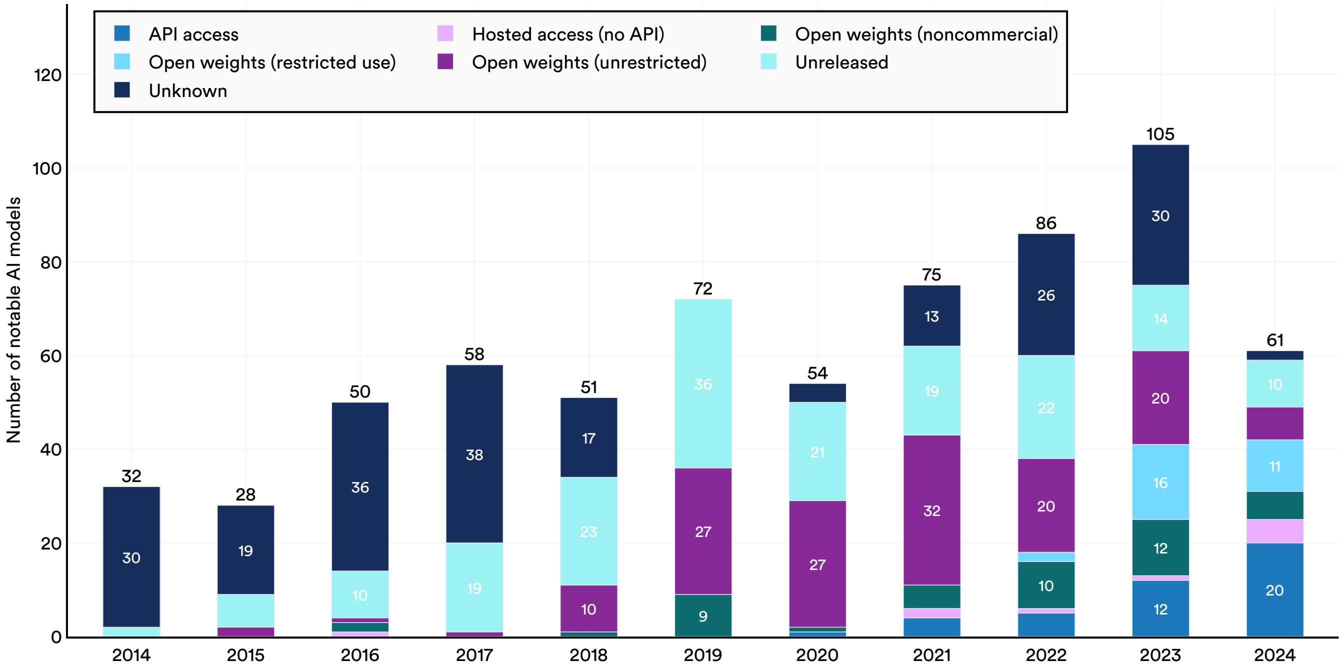


Figure 1.3.8[21]

[20] 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.

[21] Not all models in the Epoch database are categorized by access type, so the totals in Figures 1.3.8 through 1.3.10 may not fully align with those reported elsewhere in the chapter.

Notable AI models (% of total) by access type, 2014–24

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

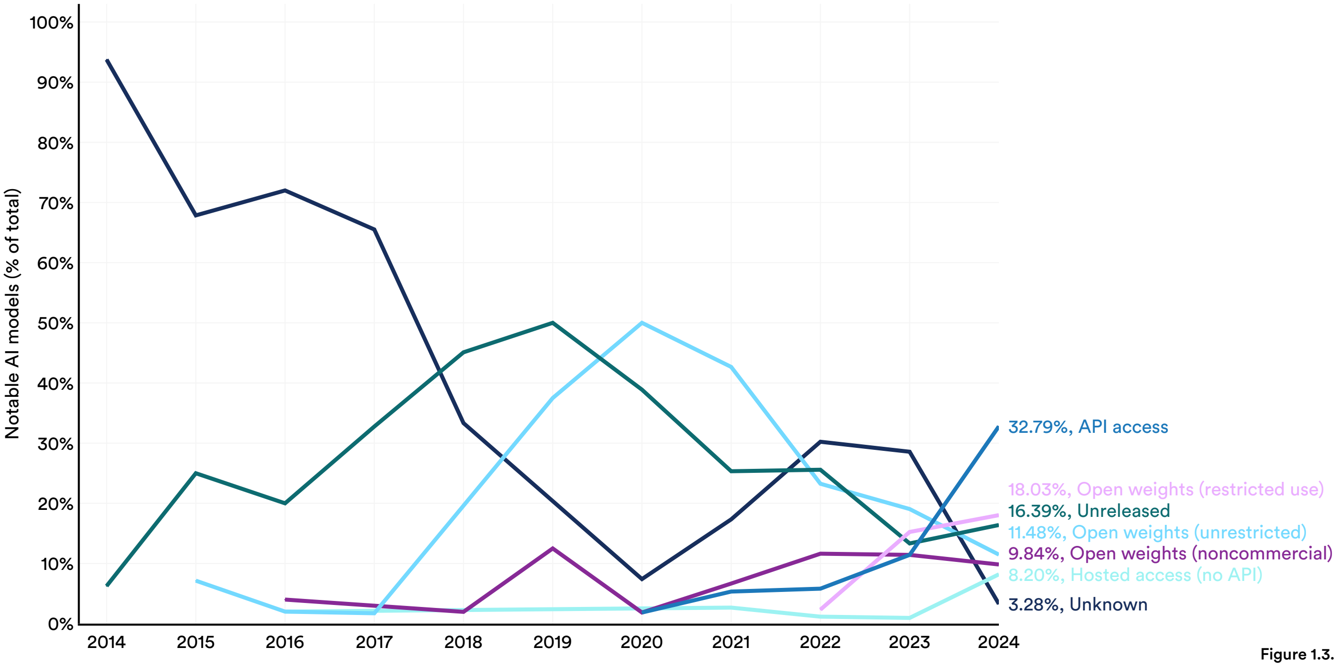


Figure 1.3.9

In traditional open-source software releases, all components, including the training code, are typically made available. However, this is often not the case with AI technologies, where even developers who release model weights may withhold the training code. Figure 1.3.10 categorizes notable AI models by the openness of their code release. In 2024, the majority—60.7%—were launched without corresponding training code.

Number of notable AI models by training code access type, 2014–24

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

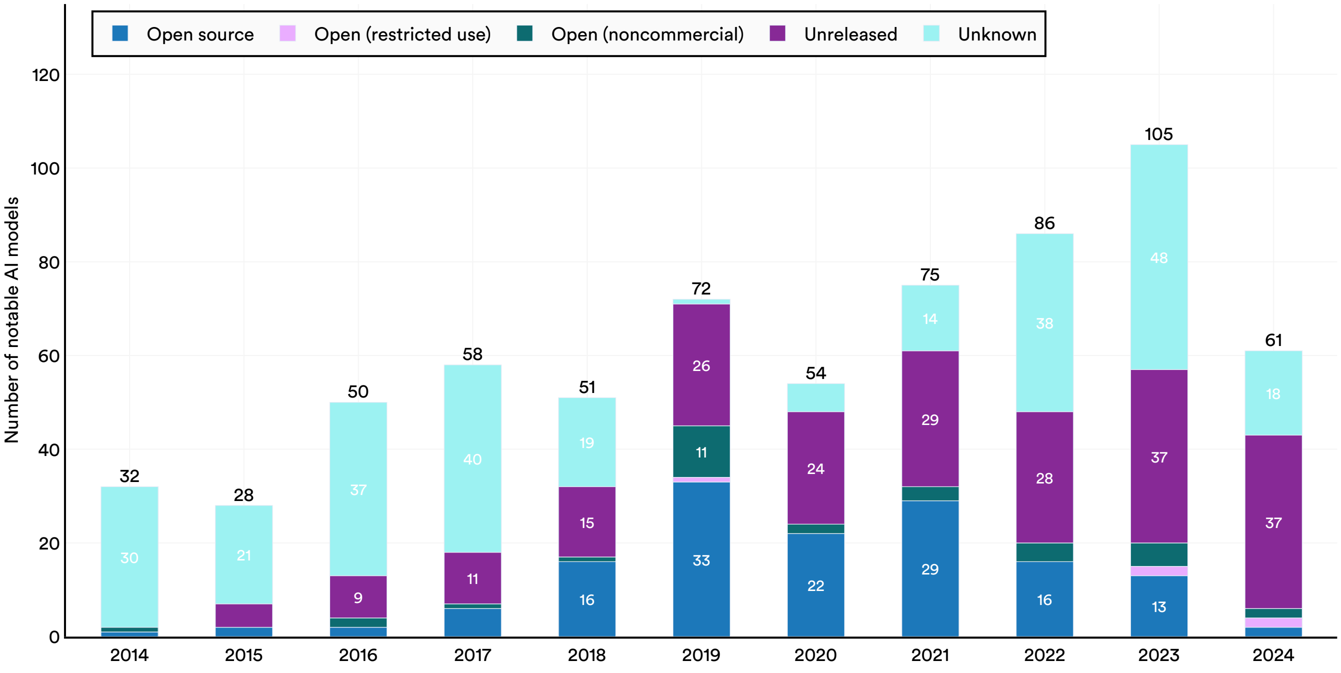


Figure 1.3.10

Parameter Trends

Parameters in machine learning models are numerical values learned during training that determine how a model interprets input data and makes predictions. Models with more parameters require more data to be trained, but they can take on more tasks and typically outperform models with fewer parameters.

Figure 1.3.11 demonstrates the parameter count of machine learning models in the Epoch dataset, categorized by the sector from which the models originate. Figure 1.3.12 visualizes the same data, but for a smaller selection of notable models. Parameter counts have risen sharply since the early 2010s, reflecting the growing complexity of their architecture, greater availability of data, improvements in hardware, and proven efficacy of larger models. High-parameter models are particularly notable in the industry sector, underscoring the substantial financial resources available to industry to cover the computational costs of training on vast volumes of data. Several of the figures below use a log scale to reflect the exponential growth in AI model parameters and compute in recent years.

Number of parameters of notable AI models by sector, 2003–24

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

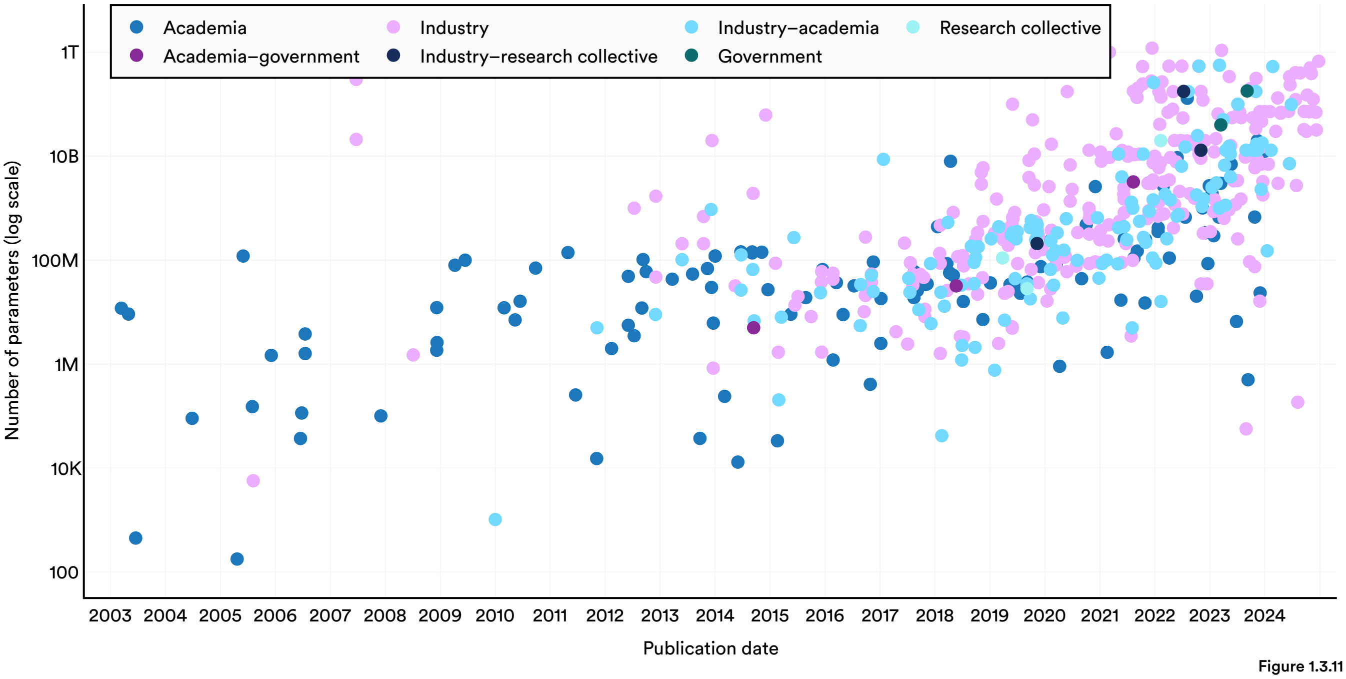


Figure 1.3.11

Number of parameters of select notable AI models by sector, 2012–24

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

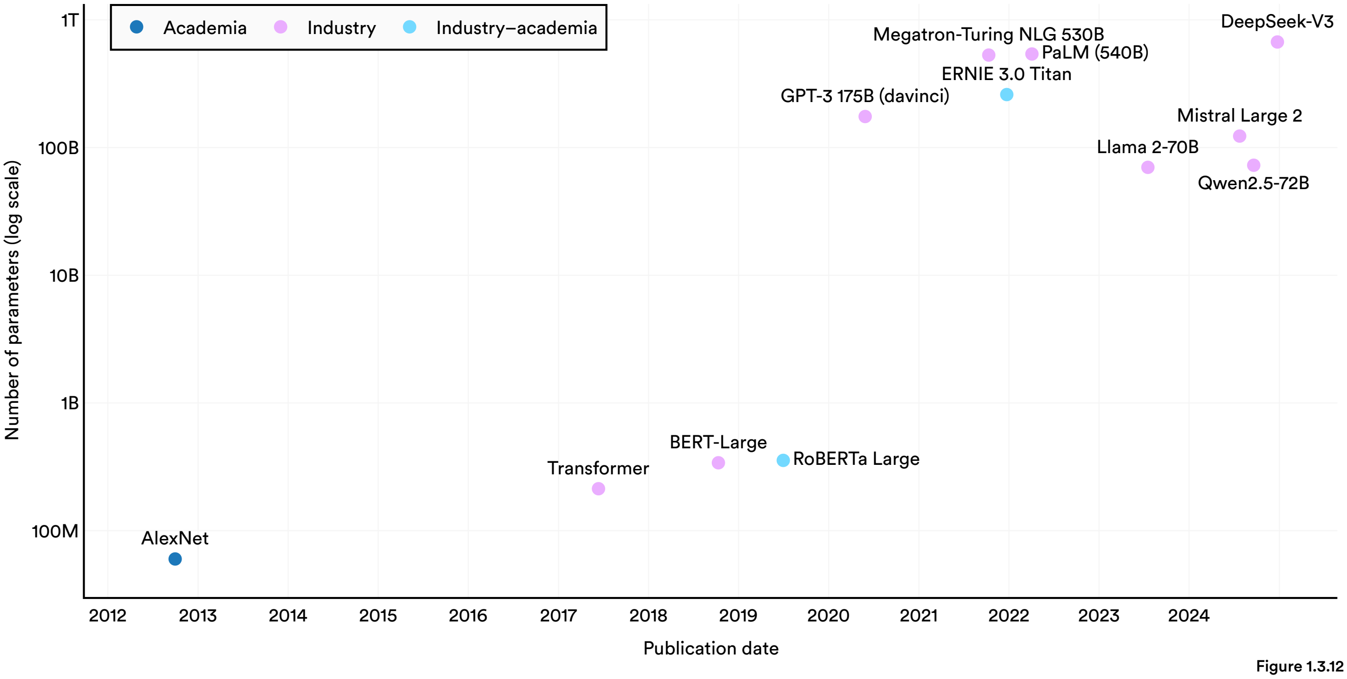


Figure 1.3.12

As model parameter counts have increased, so has the volume of data used to train AI systems. Figure 1.3.13 illustrates the growth in dataset sizes used to train notable machine learning models. The Transformer model, released in 2017 and widely credited with sparking the large language model revolution, was trained on approximately 2 billion tokens. By 2020, GPT-3 175B—one of the models underpinning the original ChatGPT—was trained on an estimated 374 billion tokens. In contrast, Meta's flagship LLM, Llama 3.3, released in the summer of 2024, was trained on roughly 15 trillion tokens. According to Epoch AI, LLM training datasets double in size approximately every eight months.

Training dataset size of notable AI models, 2010–24

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

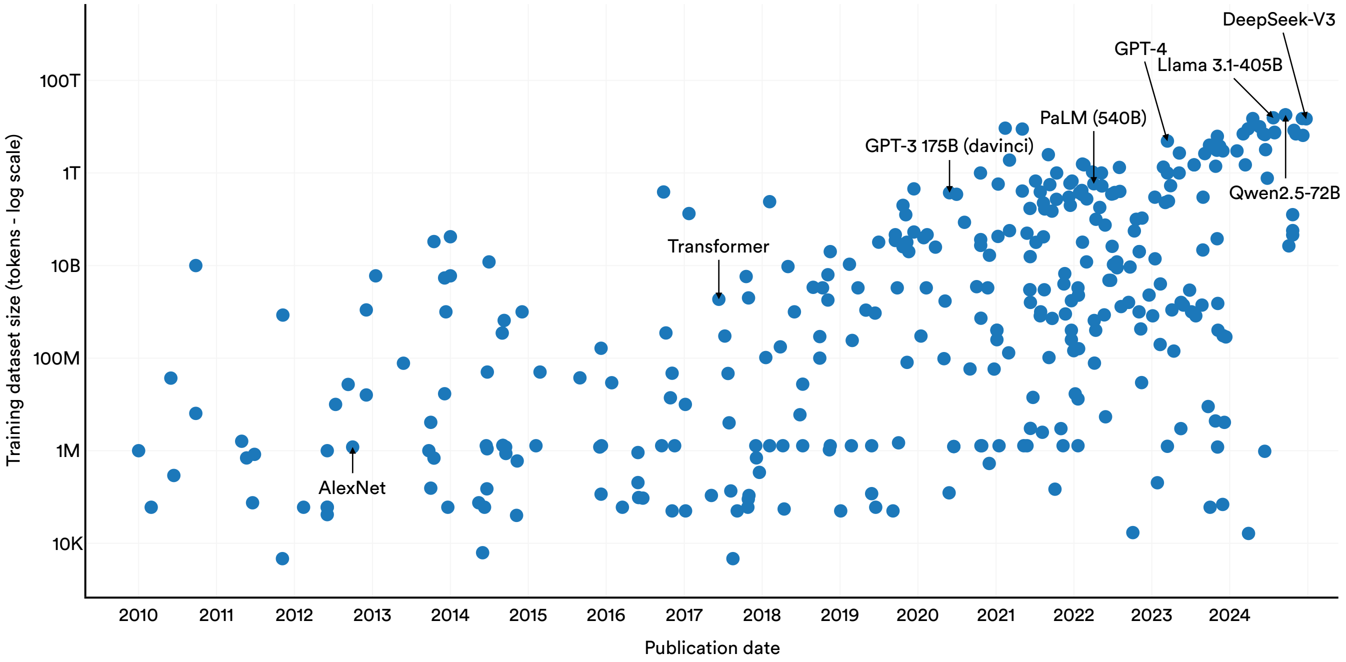


Figure 1.3.13

Training models on increasingly large datasets has led to significantly longer training times (Figure 1.3.14). Some state-of-the-art models, such as Llama 3.1-405B, required approximately 90 days to train—a typical window by today's standards. Google's Gemini 1.0 Ultra, released in late 2023, took around 100 days. This stands in stark contrast to AlexNet, one of the first models to leverage GPUs for enhanced performance, which trained in just five to six days in 2012. Notably, AlexNet was trained on far less advanced hardware.

Training length of notable AI models, 2010–24

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

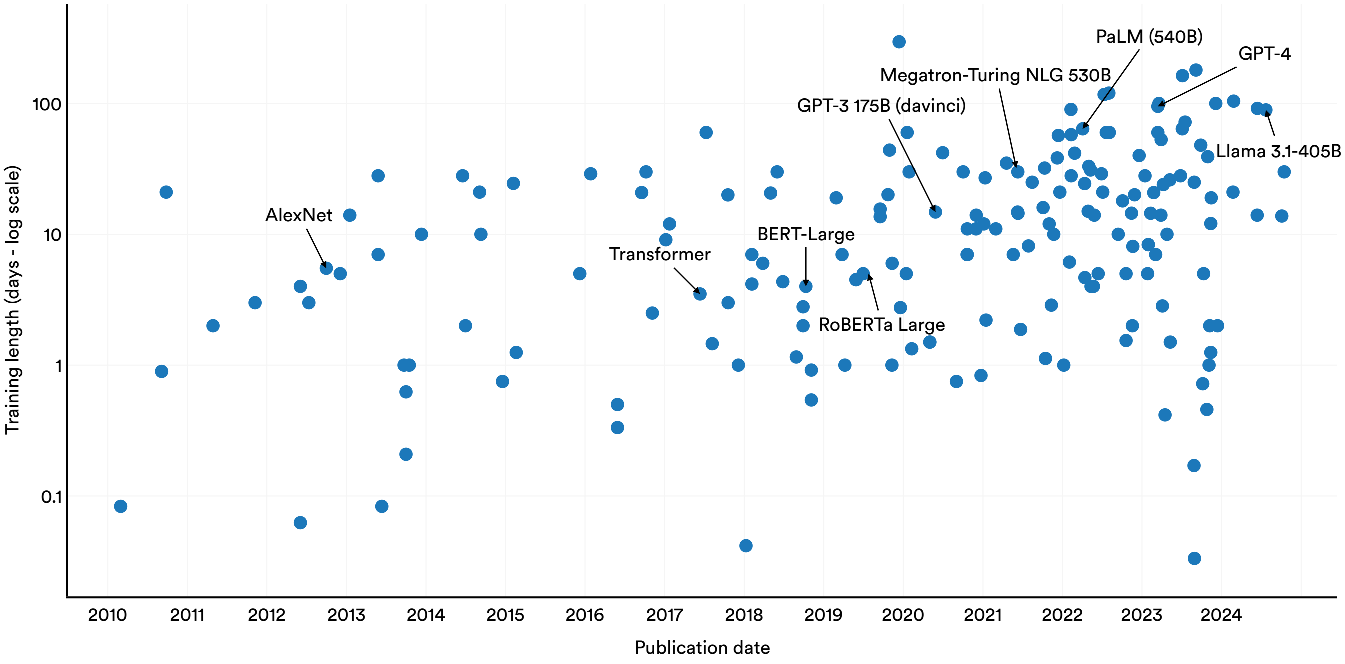


Figure 1.3.14

Compute Trends

The term "compute" in AI models denotes the computational resources required to train and operate a machine learning model. Generally, the complexity of the model and the size of the training dataset directly influence the amount of compute needed. The more complex a model is, and the larger the underlying training data, the greater the amount of compute required for training. Before the final training run, researchers conduct numerous test runs throughout the R&D phase. While training a single model is relatively inexpensive, the cumulative cost of multiple R&D runs and the necessary datasets quickly becomes significant. These figures reflect only the final training run, not the entire R&D process.

Figure 1.3.15 visualizes the training compute required for notable machine learning models over the past 22 years. Recently, the compute usage of notable AI models has increased exponentially.[22] Epoch estimates that the training compute of notable AI models doubles roughly every five months. This trend has been especially pronounced in the last five years. This rapid rise in compute demand has important implications. For instance, models requiring more computation often have larger environmental footprints, and companies typically have more access to computational resources than academic institutions. For reference, Chapter 2 of the AI Index analyzes the relationship between improvements in computational resources and model performance.

Training compute of notable AI models by sector, 2003–24

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

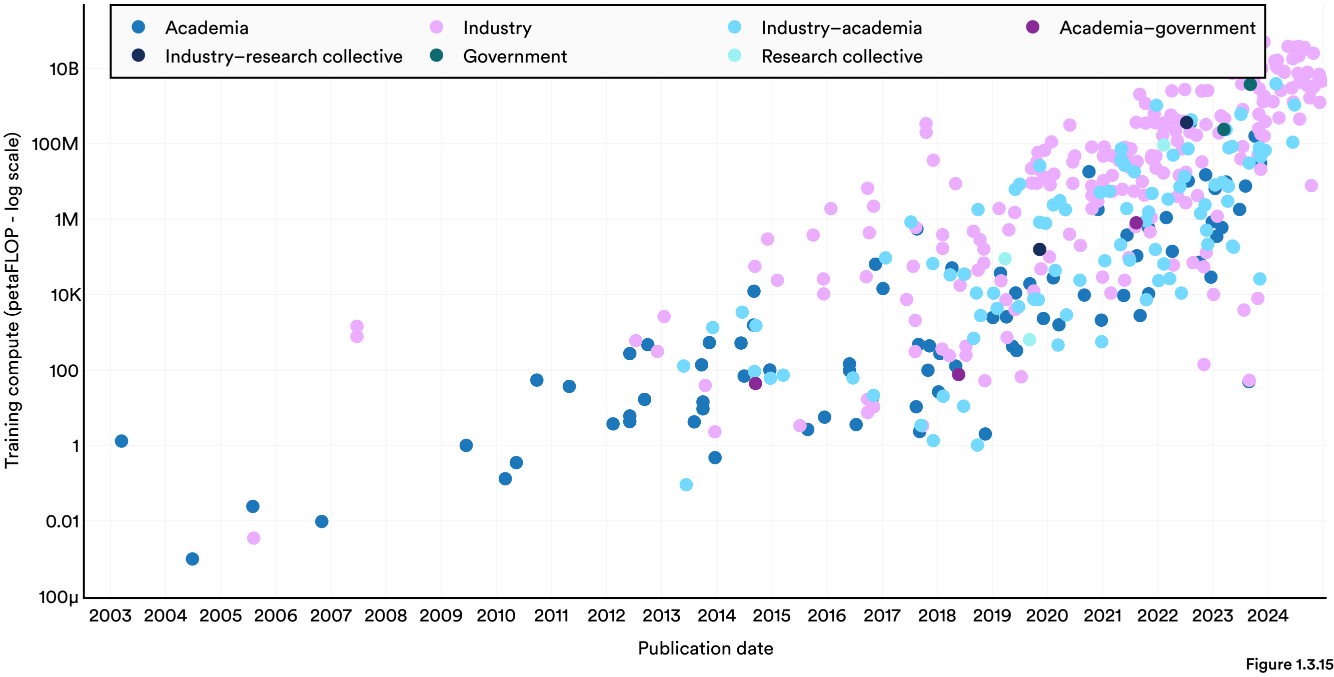


Figure 1.3.15[23]

[22] FLOP stands for "floating-point operation." A floating-point operation is a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division. The number of FLOP a processor or computer can perform per second is an indicator of its computational power. The higher the FLOP rate, the more powerful the computer. The number of floating-point operations used to train an AI model reflects its requirement for computational resources during development.

[23] 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.

Figure 1.3.16 highlights the training compute of notable machine learning models since 2012. For example, AlexNet, one of the models that popularized the now standard practice of using GPUs to improve AI models, required an estimated 470 petaFLOP for training.[24] The original Transformer, released in 2017, required around 7,400 petaFLOP. OpenAI's GPT-4o, one of the current state-of-the-art foundation models, required 38 billion petaFLOP. Creating cutting-edge AI models now demands a colossal amount of data, computing power, and financial resources that are not available to academia. Most leading AI models are coming from industry, a trend that was first highlighted in last year's AI Index. Although the gap has slightly narrowed this year, the trend persists.

Training compute of notable AI models by domain, 2012–24
Source: Epoch AI, 2025 | Chart: 2025 AI Index report

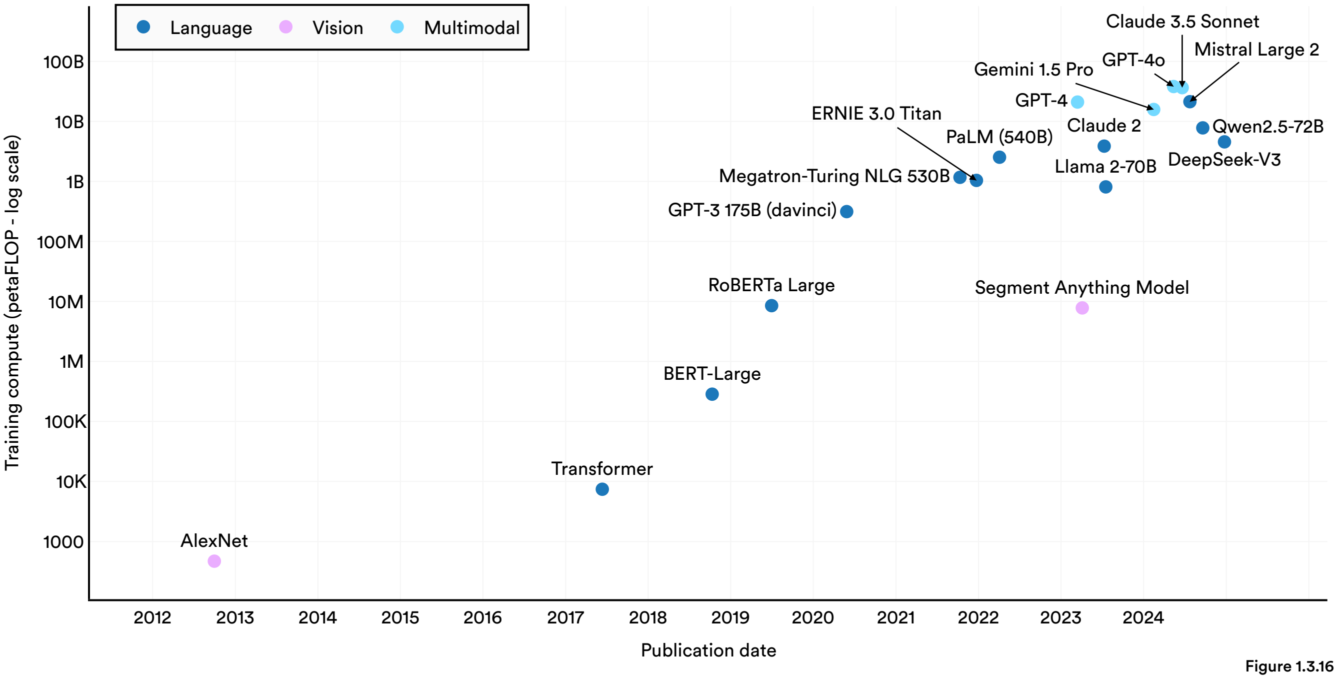


Figure 1.3.16

[24] A petaFLOP (PFLOP) is a unit of computing power equal to one quadrillion (10^{15}) floating-point operations per second.

The launch of DeepSeek's V3 model in December 2024 garnered significant attention, particularly because it achieved exceptionally high performance while requiring far fewer computational resources than many leading LLMs. Figure 1.3.17 compares the training compute of notable machine learning models from the United States and China, highlighting a key trend: Top-tier AI models from the U.S. have generally been far more computationally intensive than Chinese models. According to Epoch AI, the top 10 Chinese language models by training compute have scaled at a rate of about three times per year since late 2021—considerably slower than the five times per year trend observed in the rest of the world since 2018.

Training compute of select notable AI models in the United States and China, 2018–24

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

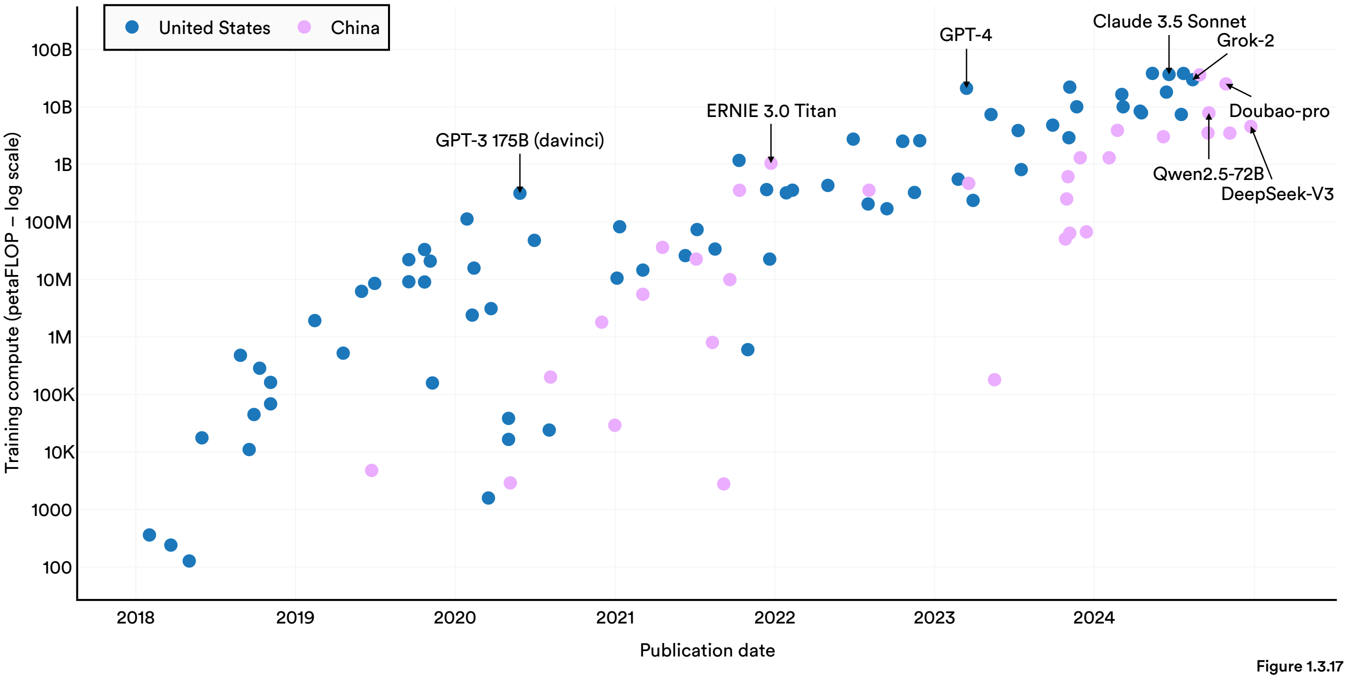


Figure 1.3.17

Highlight:

Will Models Run Out of Data?

One of the key drivers of substantive algorithmic improvements in AI systems has been the scaling of models and their training on ever-larger datasets. However, as the supply of internet training data becomes increasingly depleted, concerns have grown about the sustainability of this scaling approach and the potential for a data bottleneck, where returns to scale diminish. Last year's AI Index explored various factors in this debate, including the availability of existing internet data and the potential for training models on synthetic data. New research this year suggests that the current stock of data may last longer than previously expected.

Epoch AI has updated its previous estimates for when AI researchers might run out of data. In its latest research, the team estimated the total effective stock of data available for training models according to token count (Figure 1.3.18). Common Crawl, an open repository of web crawl data frequently used in AI training, is estimated to contain a median of 130 trillion tokens. The indexed web holds approximately 510 trillion tokens, while the entire web contains around 3,100 trillion. Additionally, the total stock of images is estimated at 300 trillion, and video at 1,350 trillion.

Estimated median data stocks

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

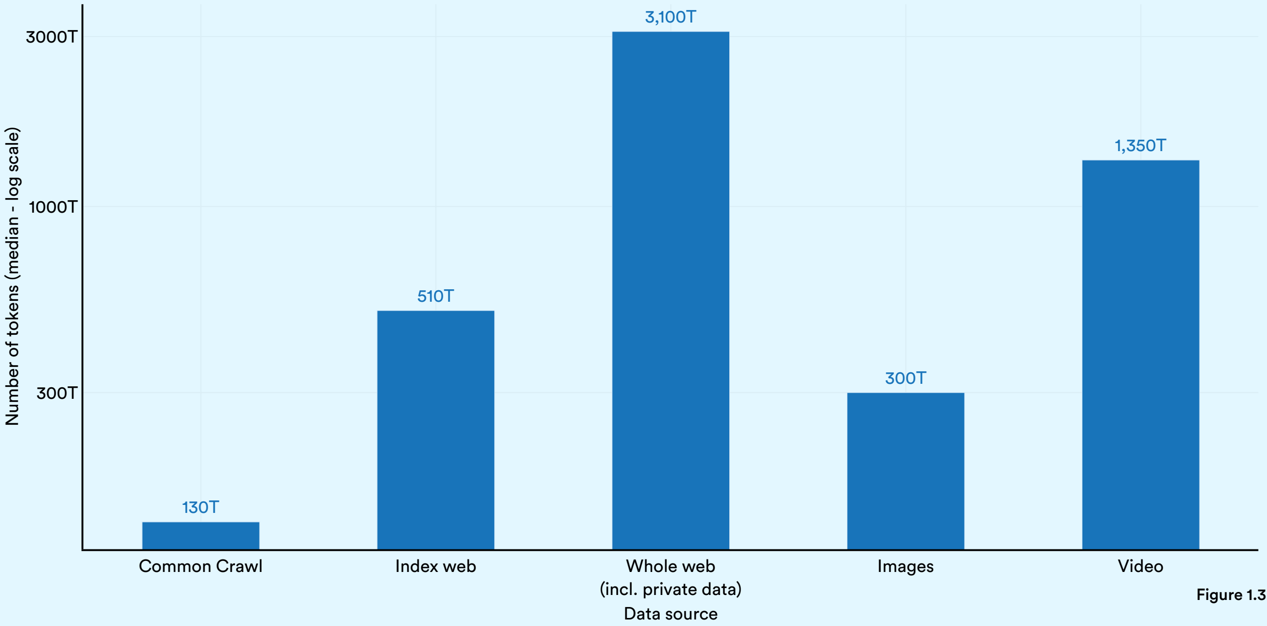


Figure 1.3.18

Highlight:

Will Models Run Out of Data? (cont'd)

The Epoch AI research team projects, with an 80% confidence interval, that the current stock of training data will be fully utilized between 2026 and 2032 (Figure 1.3.19). Several factors influence the point in time when data is likely to run out. One key factor is the historical growth of dataset sizes, which depends on how many people generate and contribute content to the internet. Another important factor is computer usage. If models are trained in a compute-optimal manner, the available data stock can last longer. However, if models are overtrained to achieve more compute-efficient inference performance, the stock is likely to be depleted sooner. When AI models are overtrained, meaning they are trained for an extended period beyond the typical point of diminishing returns, they may achieve more compute-efficient inference—that is, they can process prompts (make predictions, generate text, etc.) using less computational power. However, this comes at a cost: The stock (i.e., data available to train the model) may be depleted more quickly.

Projections of the stock of public text and data usage

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

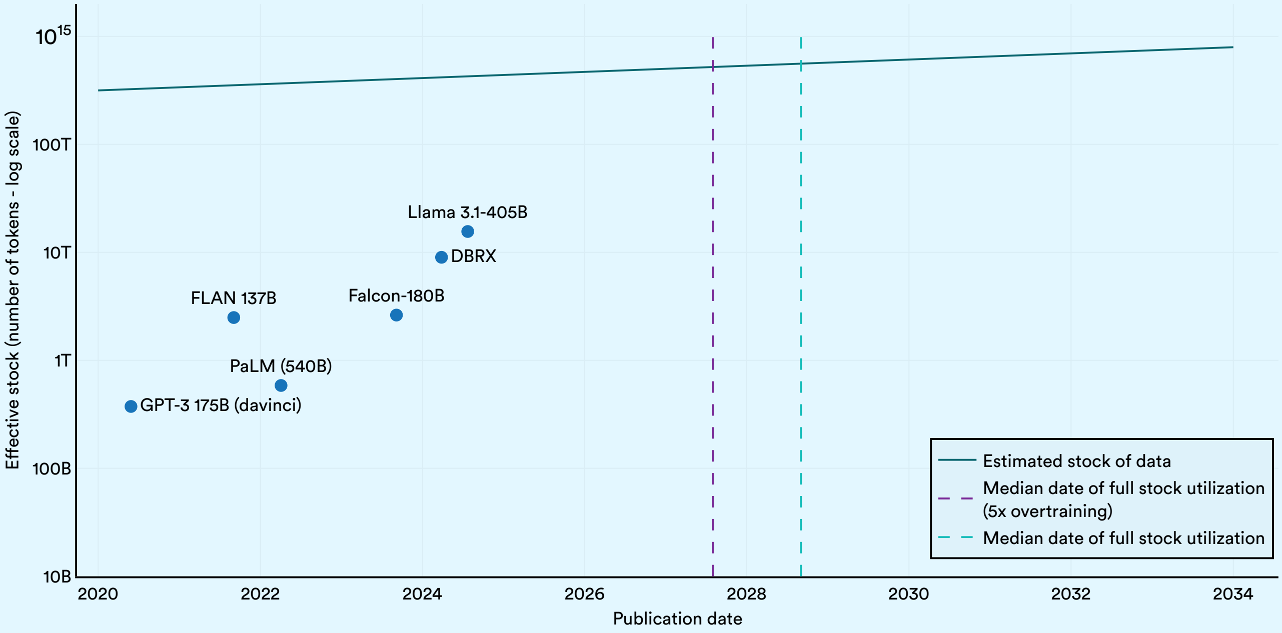


Figure 1.3.19

Highlight:

Will Models Run Out of Data? (cont'd)

These projections differ slightly from Epoch's earlier estimates, which predicted that high-quality text data would be depleted by 2024. The revised projections reflect an updated methodology that incorporates new research showing that web data performs better than curated corpora and that models can be trained on the same datasets multiple times. The realization that carefully filtered web data is effective and that repeated training on the same dataset is viable has expanded estimates of the available data stock. As a result, the Epoch researchers pushed back their forecasts of when data depletion might occur.

Using synthetic data—data generated by AI models themselves—to train models has also been suggested as a solution to potential data shortages. The 2024 AI Index suggests there are limitations associated with this approach, namely that models trained this way are likely to lose representation of the tails of distributions when performing repeated training cycles on synthetic data. This leads to degraded model output quality. This phenomenon was observed across different model architectures, including variational autoencoders (VAEs), Gaussian mixture models (GMMs), and LLMs. However, newer research suggests that when synthetic data is layered on top of real data, rather than replacing it, the model collapse phenomenon does not occur. While this accumulation does not necessarily improve performance or reduce test loss (lower test loss indicates better model performance), it also does not result in the same degree of degradation as outright data replacement (Figure 1.3.20).

Effect of data accumulation on language models pretrained on TinyStories

Source: Gerstgrasser et al., 2024 | Chart: 2025 AI Index report

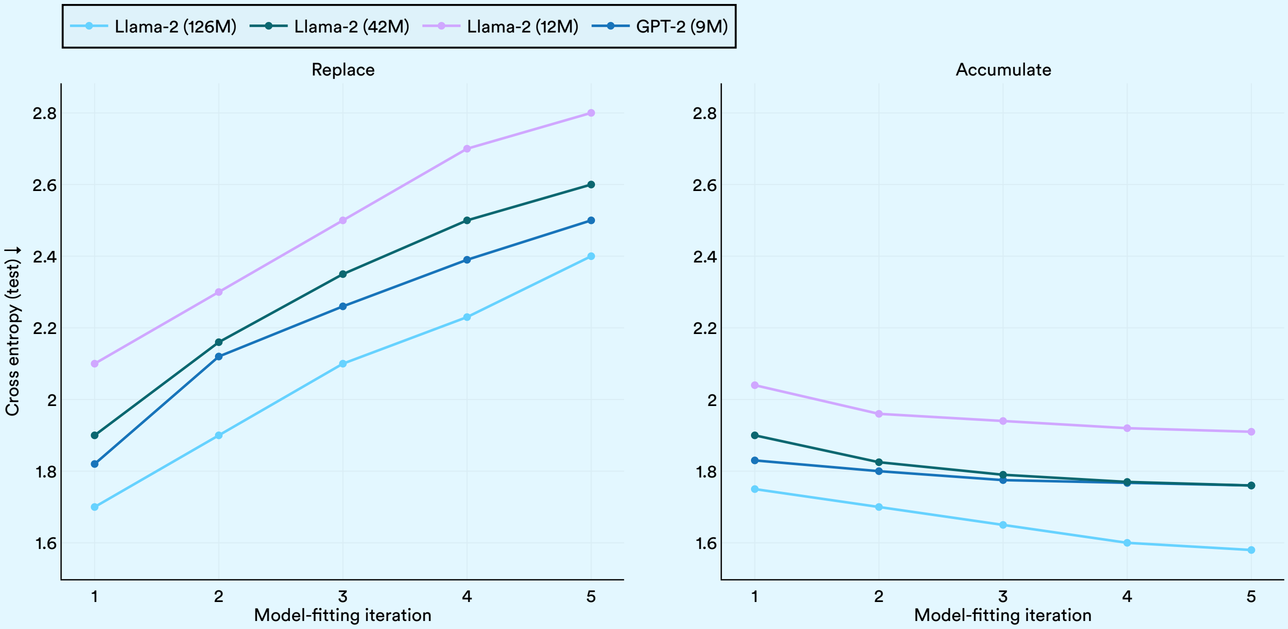


Figure 1.3.20

Highlight:

Will Models Run Out of Data? (cont'd)

This year, there have been advances in generating high-fidelity synthetic data. However, synthetic data is still generally distinguishable from real data, and there is no existing scalable method to achieve the same performance training LLMs on synthetic data compared to real data. A team of Slovenian researchers compared the performance of models trained on synthetic and real data across multiple architectures and datasets. They evaluated how well synthetic relational data preserves key characteristics of the original data ("fidelity") and remains useful for downstream tasks ("utility"). They found that most methods are systematically detectable as synthetic, especially once relational information is considered. Furthermore, performance typically deteriorates compared to real data–trained models, but some methods still yield moderately good predictive scores. In a few experiments, synthetic data outperformed real data such as using Synthetic Data Vault (SDV) vs. Walmart data to train an XGBoost classifier. The researchers showed that training on the synthetic dataset achieves a lower mean squared error (MSE). There is also evidence that synthetic data shows promise in the healthcare domain. More specifically, some model architectures lead to enhanced performance on classification and prediction tasks by training on synthetically augmented datasets, increasing F1 scores or AUROC by 5%–10% on minority classes.[25]

There are concerns around the quality and fidelity of synthetically generated data, as LLMs are known to hallucinate and provide factually incorrect outputs. When training on hallucinated content in datasets, models can experience compounded degradation in output quality. New techniques have been developed to combat this issue. For example, researchers from Stanford and the University of North Carolina at Chapel Hill have used automated fact-checking and confidence scores to rank factuality scores of model response pairs. The FactTune-FS methods introduced by these researchers have tended to outperform other RLHF and decoding-based methods for factuality improvement (Figure 1.3.21). Human-in-the-loop approaches to label preferred responses have also been used to align language models. While promising, the human-in-the-loop approaches tend to be more expensive. Finally, post hoc filtering and debiasing methods can be used to remove anomalies in synthetic data before the training stage.

[25] AUROC (area under the receiver operating characteristic) curve is a widely used metric for evaluating AI model performance, particularly in classification tasks.

Highlight:

Will Models Run Out of Data? (cont’d)

Factual accuracy: percentage of correct answers in biographies

Source: Tian et al., 2023 | Chart: 2025 AI Index report

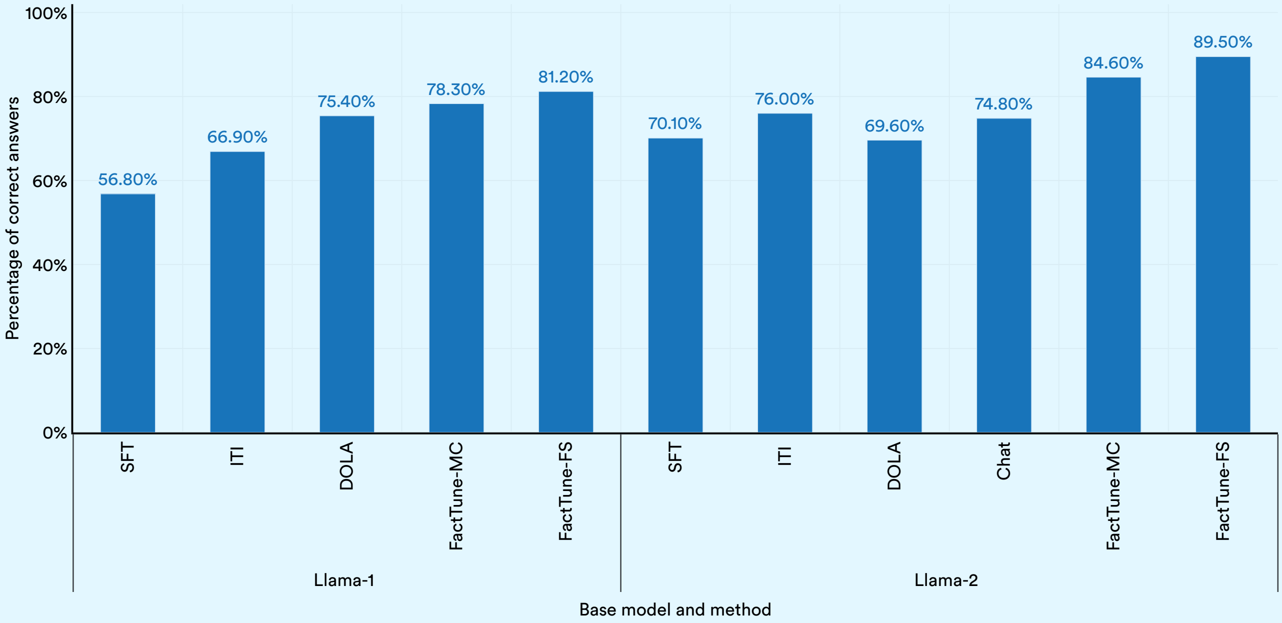


Figure 1.3.21

As the prevalence of synthetic data grows, particularly with an increasing share of web content being AI-generated, future models will inevitably be trained on non-human-generated material. While synthetic data offers the advantage of a near-infinite supply, effectively leveraging it for model training requires a deeper understanding of its impact on learning dynamics and performance. One approach to expanding datasets is data augmentation, which modifies real data—such as tilting or image mixing—to create new variations while preserving essential characteristics. Both synthetic data generation and data augmentation present opportunities to enhance AI models, but their effective use demands further research.

Inference Cost

Last year's AI Index highlighted the rapidly rising training costs of frontier LLM systems. This year, in addition to updating its analysis on training costs, the Index examines how inference costs for frontier systems have evolved over time. Inference costs refer to the expense of querying a trained model, and they are typically measured in USD per million tokens. Data on AI token pricing comes from both Artificial Analysis and Epoch AI's proprietary database on API pricing. The reported price is a 3:1 weighted average of input and output token prices.

To analyze inference costs, the AI Index worked with Epoch to measure how costs have decreased for a fixed AI performance threshold. This standardized approach facilitates a more accurate comparison. While newer models may cost more, they also tend to perform significantly better—so comparing them directly to older, less capable models can obscure the real trend: AI performance per dollar has improved substantially. For instance, the inference cost for an AI model scoring the equivalent of GPT-3.5 (64.8) on MMLU, a popular benchmark for assessing language model performance, dropped from $20.00 per million tokens in November 2022 to just $0.07 per million tokens by October 2024 (Gemini-1.5-Flash-8B)—a more than 280-fold reduction in approximately 1.5 years. A similar trend is evident in the cost of models scoring above 50% on GPQA, a substantially more challenging benchmark than MMLU. There, inference costs declined from $15 per million tokens in May 2024 to $0.12 per million tokens by December 2024 (Phi 4). Epoch AI estimates that, depending on the task, LLM inference costs have been falling anywhere from nine to 900 times per year.

Inference price across select benchmarks, 2022–24
Source: Epoch AI, 2025; Artificial Analysis, 2025 | Chart: 2025 AI Index report

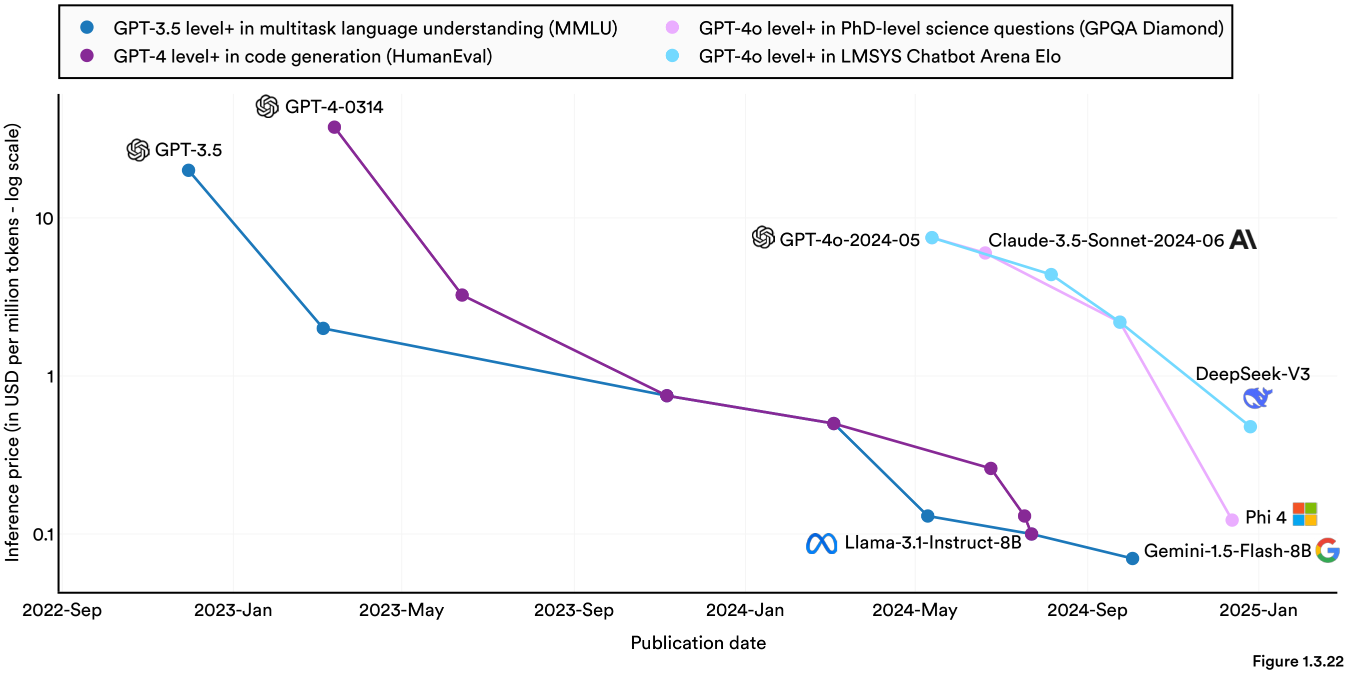


Figure 1.3.22

The inference cost to achieve a given level of performance has declined notably over time. However, state-of-the-art models remain more expensive than some of the previously mentioned alternatives. Figure 1.3.23 illustrates the cost per million tokens for leading models from developers such as OpenAI, Meta, and Anthropic.[26] These top-tier models are generally priced higher than smaller models from the same companies, reflecting the premium required for cutting-edge performance.

Output price per million tokens for select models
Source: Artificial Analysis, 2025 | Chart: 2025 AI Index report

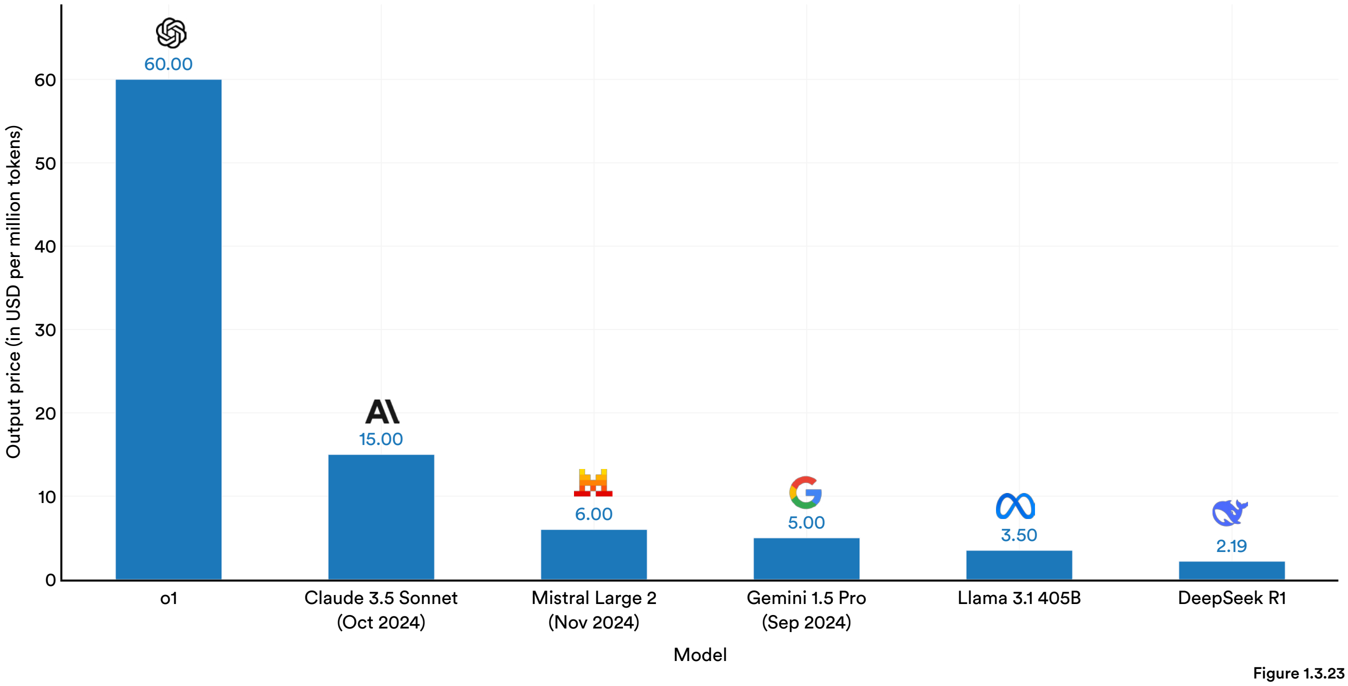


Figure 1.3.23

Training Cost

A frequent discussion around foundation models pertains to their high training costs. While AI companies rarely disclose exact figures, costs are widely estimated to reach into the millions of dollars—and continue to rise. OpenAI CEO Sam Altman, for instance, indicated that training GPT-4 exceeded $100 million. In July 2024, Anthropic CEO Dario Amodei noted that model training runs costing around $1 billion were already underway. Even more recent models, such as DeepSeek-V3, reportedly cost less—about $6 million—but overall, training remains extremely expensive.[27]

Understanding the costs associated with training AI models remains important, yet detailed cost information remains scarce. Last year, the AI Index published initial estimates on the costs of training foundation models. This year, the AI Index once again partnered with Epoch AI to update and refine these estimates. To calculate costs for cutting-edge models, the Epoch team analyzed factors such as training duration, hardware type, quantity, and utilization rates, relying on information from academic publications, press releases, and technical reports.[28]

[26] The Index visualizes a selection of state-of-the-art models with publicly available pricing as of February 2025. Since publication, newer models may have been released and pricing may have changed.

[27] Some reports have disputed the stated cost of DeepSeek-V3, arguing that when factoring in employee salaries, capital expenditures, and research expenses, the actual development costs were significantly higher.

[28] A detailed report on Epoch's research methodology is available in this paper.

Figure 1.3.24 visualizes the estimated training cost associated with select AI models, based on cloud compute rental prices. Figure 1.3.25 visualizes the training cost of all AI models for which the AI Index has estimates.

AI Index estimates validate suspicions that in recent years model training costs have significantly increased. For example, in 2017, the original Transformer model, which introduced the architecture that underpins virtually every modern LLM, cost around $670 to train. RoBERTa Large, released in 2019, which achieved state-of-the-art results on many canonical comprehension benchmarks like SQuAD and GLUE, cost around $160,000 to train. Fast-forward to 2023, and training costs for OpenAI's GPT-4 were estimated around $79 million.

One of the few 2024 models for which Epoch could estimate training costs was Llama 3.1-405B, with an estimated cost of $170 million. As the AI landscape grows more competitive, companies are disclosing less about their training processes, making it increasingly difficult to estimate computational costs.

As established in previous AI Index reports, there is a direct correlation between the training costs of AI models and their computational requirements. As illustrated in Figure 1.3.26, models with greater computational training needs cost substantially more to train.

Estimated training cost of select AI models, 2019–24
Source: Epoch AI, 2024 | Chart: 2025 AI Index report

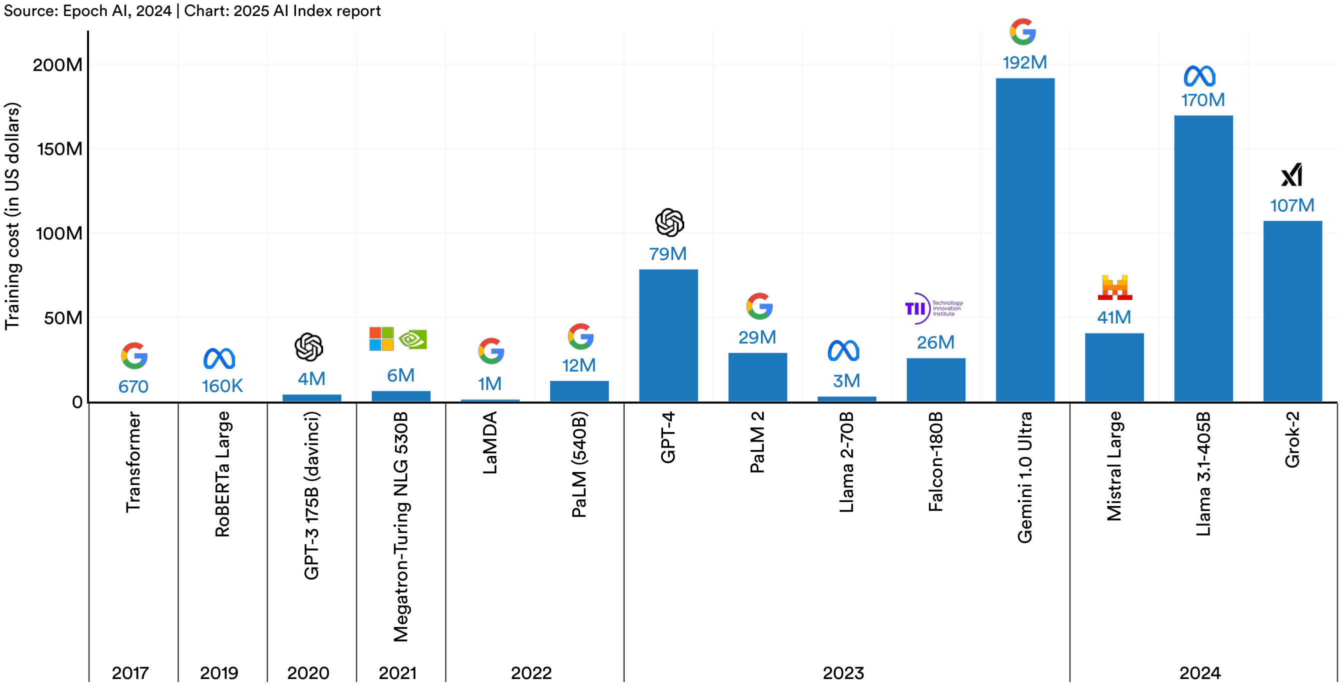


Figure 1.3.24

[29] The cost figures reported in this section are inflation-adjusted.

Estimated training cost of select AI models, 2016–24

Source: Epoch AI, 2024 | Chart: 2025 AI Index report

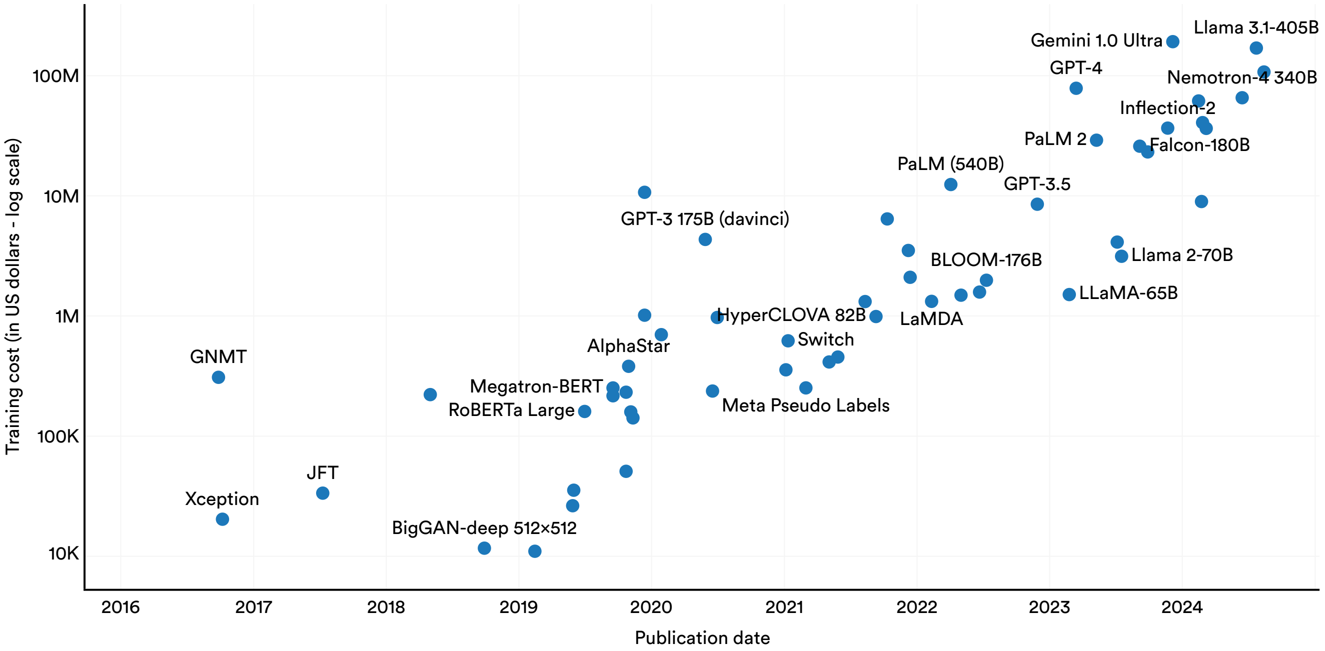


Figure 1.3.25

Estimated training cost and compute of select AI models

Source: Epoch AI, 2024 | Chart: 2025 AI Index report

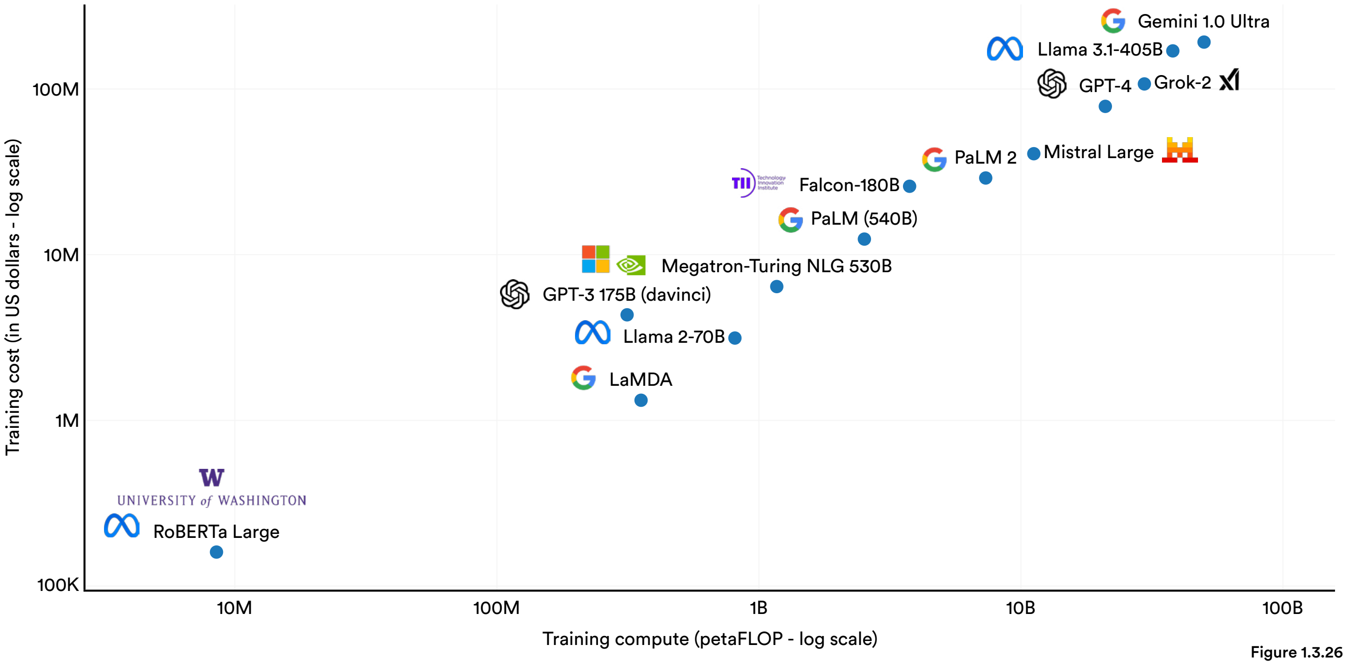


Figure 1.3.26

Hardware advancements play a critical role in driving AI progress. While scaling models and training on larger datasets have led to significant performance improvements, these advances have largely been enabled by improvements in hardware—particularly the development of more powerful and efficient GPUs (graphics processing units). GPUs accelerate complex computations, allowing models to process vast amounts of data in parallel and significantly reducing training time. This section of the Index leverages data from Epoch AI to analyze key trends in machine learning hardware and its impact on AI development.

While this section currently emphasizes compute performance (FLOP/s), network bandwidth—the speed at which GPUs communicate—is equally critical. Although data on network bandwidth of data centers is limited, future editions of the AI Index will aim to include this information.

1.4 Hardware

Overview

Figure 1.4.1 illustrates the peak computational performance of ML hardware across different precision types, where precision refers to the number of bits used to represent numerical values, particularly floating-point numbers, in computations. The choice of precision depends on the specific goal. For instance, lower-precision hardware, which requires fewer bits and has lower memory bandwidth, is ideal for optimizing computation speed and energy efficiency. This is particularly beneficial for AI models running on edge or mobile devices or in scenarios where inference speed is a priority. On the other hand, higher-precision hardware preserves greater numerical accuracy, making it essential for scientific computing and applications sensitive to precision errors. Of the precisions visualized in the figures below, FP32 has the highest precision, TF32 offers medium-high precision, and Tensor-FP16/BF16 and FP16 are lower-precision formats optimized for speed and efficiency.

Measured in 16-bit floating-point operations, Epoch estimates that machine learning hardware performance has grown over the period 2008–2024 at an annual rate of approximately 43%, doubling every 1.9 years. According to Epoch, this progress has been driven by increased transistor counts, advancements in semiconductor manufacturing, and the development of specialized hardware for AI workloads.

Peak computational performance of ML hardware for different precisions, 2008–24
Source: Epoch AI, 2025 | Chart: 2025 AI Index report

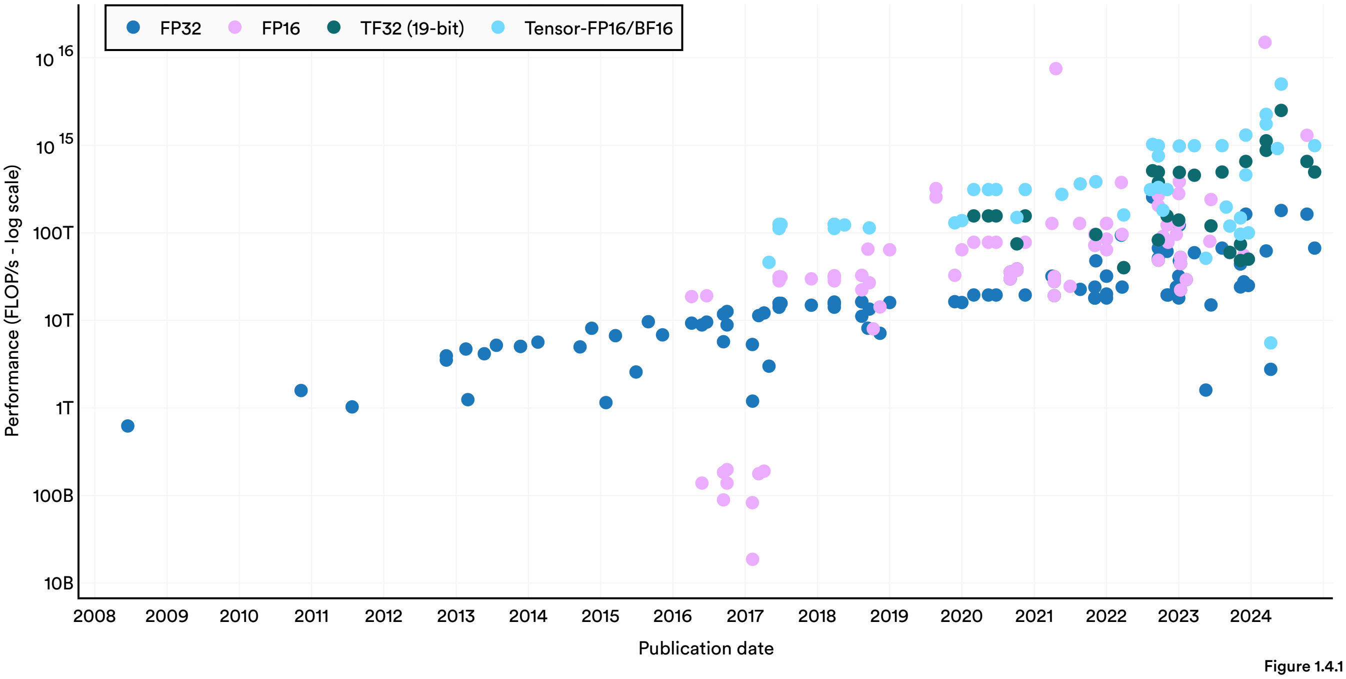


Figure 1.4.1

The price-performance of leading machine learning hardware has steadily improved. Figure 1.4.2 illustrates the performance of selected Nvidia data center GPUs—among the most commonly used for AI training—in FLOP per second. Figure 1.4.3 visualizes the price-performance of those same GPUs, measured in FLOP per second per dollar. For example, the H100 GPU, announced in March 2022, achieves 22 billion FLOP per second per dollar, which is approximately 1.7 times the price-performance of the A100 (launched in June 2020) and 16.9 times that of the P100 (released in April 2016). Epoch estimates that hardware with a fixed performance level decreases in cost by 30% annually, making AI training increasingly affordable, scalable, and conducive to model improvements.

Performance of leading Nvidia data center GPUs for machine learning

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

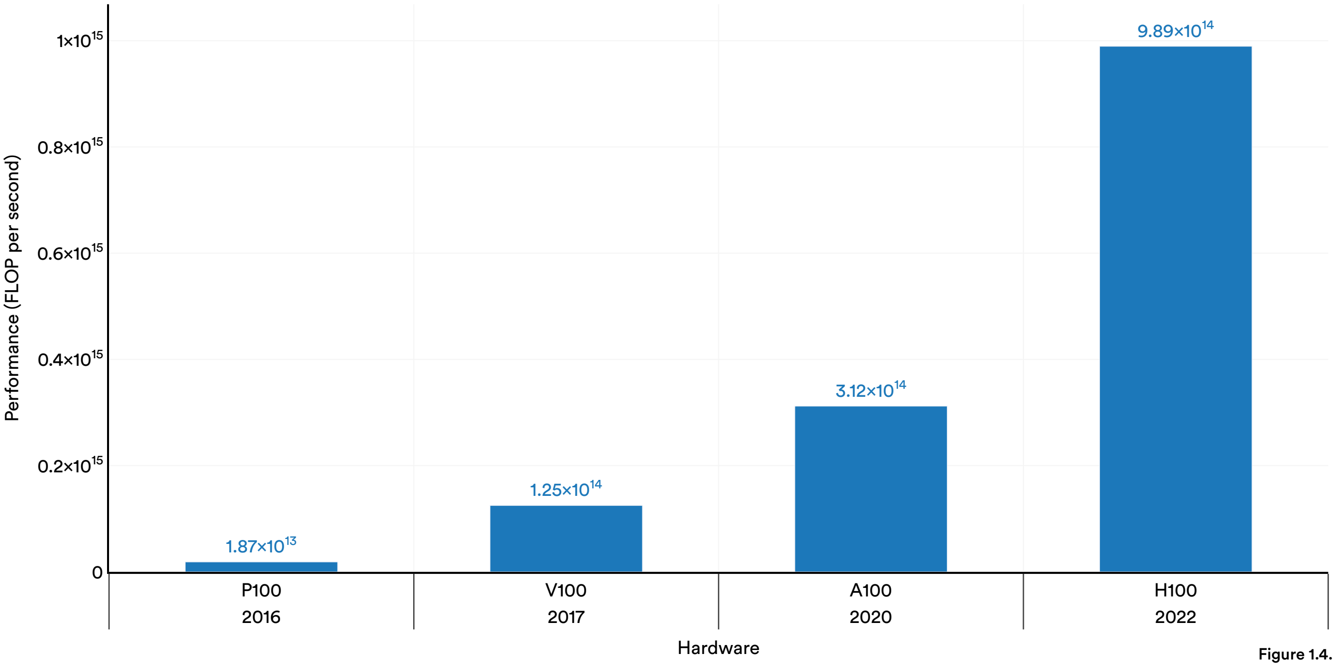


Figure 1.4.2

Figure 1.4.4, based on the Epoch AI notable machine learning models dataset, examines the hardware used to train notable machine learning models. As of 2024, the most commonly reported hardware was the A100, used by 64 models, followed by the V100. An increasing number of models are now being trained on the H100, with 15 reported by the end of 2024.

Price-performance of leading Nvidia data center GPUs for machine learning

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

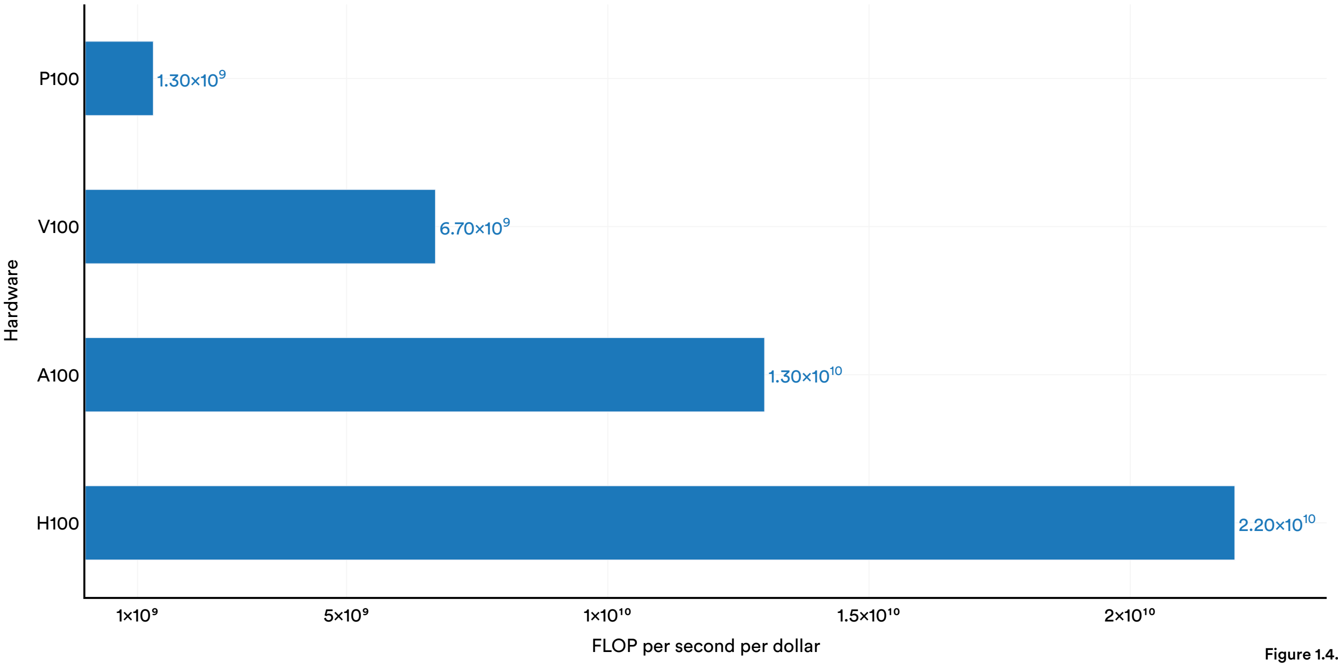


Figure 1.4.3

Cumulative number of notable AI models trained by accelerator, 2017–24

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

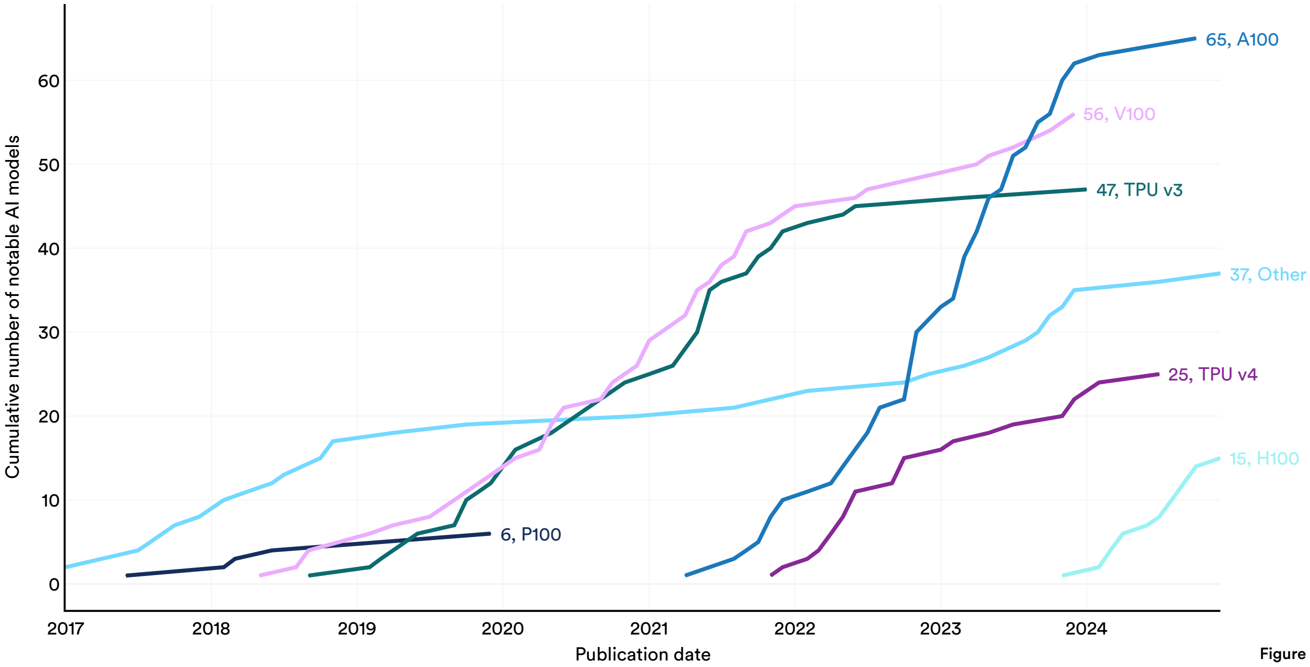


Figure 1.4.4

Highlight:

Energy Efficiency and Environmental Impact

Training AI systems requires substantial energy, making the energy efficiency of machine learning hardware a critical factor. Epoch AI reports that ML hardware has become increasingly energy efficient over time, improving by approximately 40% per year. Figure 1.4.5 illustrates the energy efficiency of Tensor-FP16 precision hardware, measured in FLOP per watt. For instance, the Nvidia B100, released in March 2024, achieved an energy efficiency of 2.5 trillion FLOP per watt, compared to the Nvidia P100, released in April 2016, which reported 74 billion FLOP per watt. This means the B100 is 33.8 times more energy efficient than the P100.

Energy efficiency of leading machine learning hardware, 2016–24

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

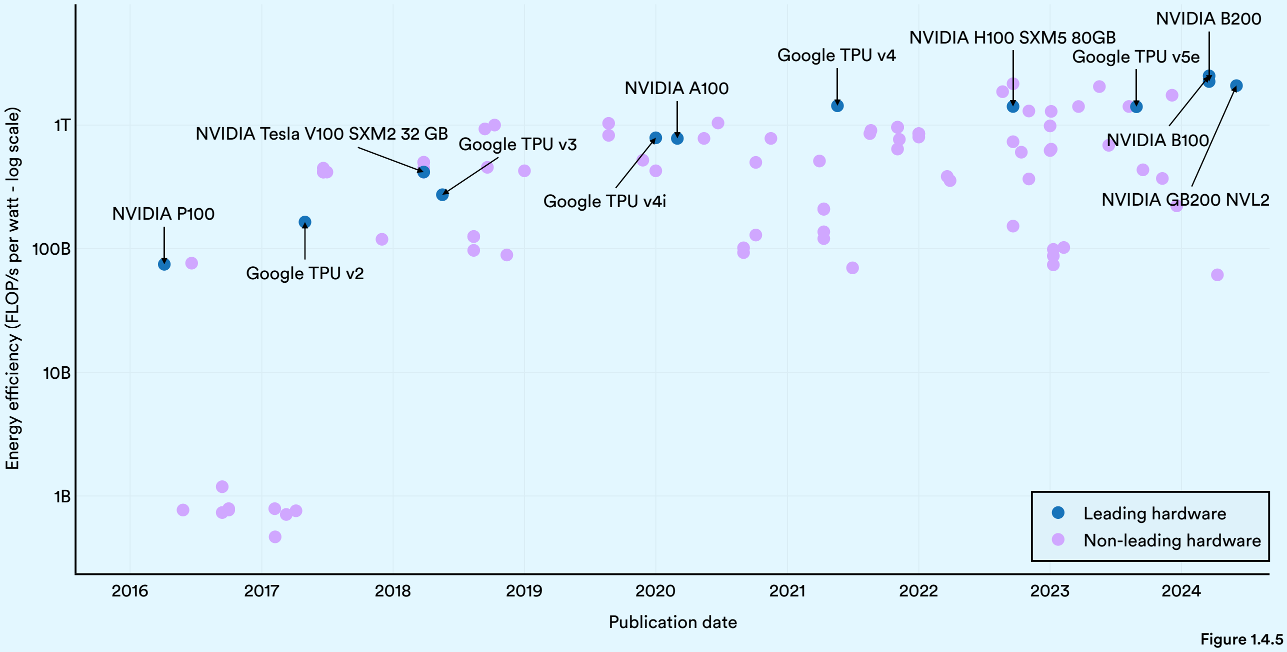


Figure 1.4.5

Highlight:

Energy Efficiency and Environmental Impact (cont'd)

Despite significant improvements in the energy efficiency of AI hardware, the overall power consumption required to train AI systems continues to rise rapidly. Figure 1.4.6 illustrates the total power draw, measured in watts, for training various state-of-the-art AI models. For example, the original Transformer, introduced in 2017, consumed an estimated 4,500 watts. In contrast, PaLM, one of Google's first flagship LLMs, had a power draw of 2.6 million watts—almost 600 times that of the Transformer. Llama 3.1-405B, released in the summer of 2024, required 25.3 million watts, consuming over 5,000 times more power than the original Transformer. According to Epoch AI, the power required to train frontier AI models is doubling annually. The rising power consumption of AI models reflects the trend of training on increasingly larger datasets.

Unsurprisingly, given that the total amount of power used to train AI systems has increased over time, so has the amount of carbon emitted by the models. Many factors determine the amount of carbon emitted by AI systems, including the number of parameters in a model, the power usage effectiveness of a data center, and the grid carbon intensity.[30]

Total power draw required to train frontier models, 2011–24

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

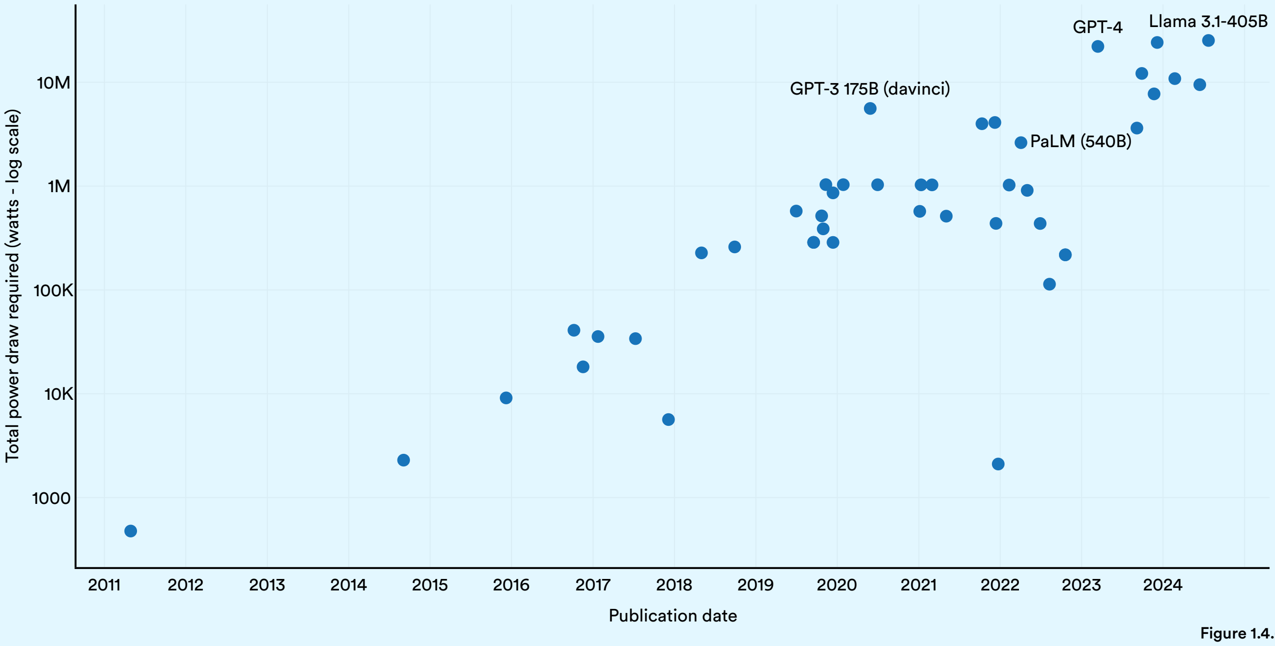


Figure 1.4.6

[30] Power usage effectiveness (PUE) is a metric used to evaluate the energy efficiency of data centers. It is the ratio of the total amount of energy used by a computer data center facility, including air conditioning, to the energy delivered to computing equipment. The higher the PUE, the less efficient the data center.

Highlight:

Energy Efficiency and Environmental Impact (cont'd)

Figure 1.4.7 illustrates the carbon emissions of selected AI models, sorted by their release year. To estimate these emissions, the AI Index used carbon data published by model developers and supplemented it with calculations from a widely used online AI training emissions calculator. This step was necessary as many developers do not disclose their models' carbon footprints. The calculator estimates emissions based on the type of hardware used for training, total training hours, cloud provider, and training region.[31]

The carbon emissions from training frontier AI models have steadily increased over time. While AlexNet's emissions were negligible, GPT-3 (released in 2020) reportedly emitted around 588 tons of carbon during training, GPT-4 (2023) emitted 5,184 tons, and Llama 3.1 405B (2024) emitted 8,930 tons. DeepSeek V3, released in 2024, and whose performance is comparable to OpenAI's o1, is estimated to have emissions comparable to the GPT-3, released five years ago. For context, on average, Americans emit 18.08 tons of carbon per capita per year.

Estimated carbon emissions from training select AI models and real-life activities, 2012–24

Source: AI Index, 2025; Strubell et al., 2019 | Chart: 2025 AI Index report

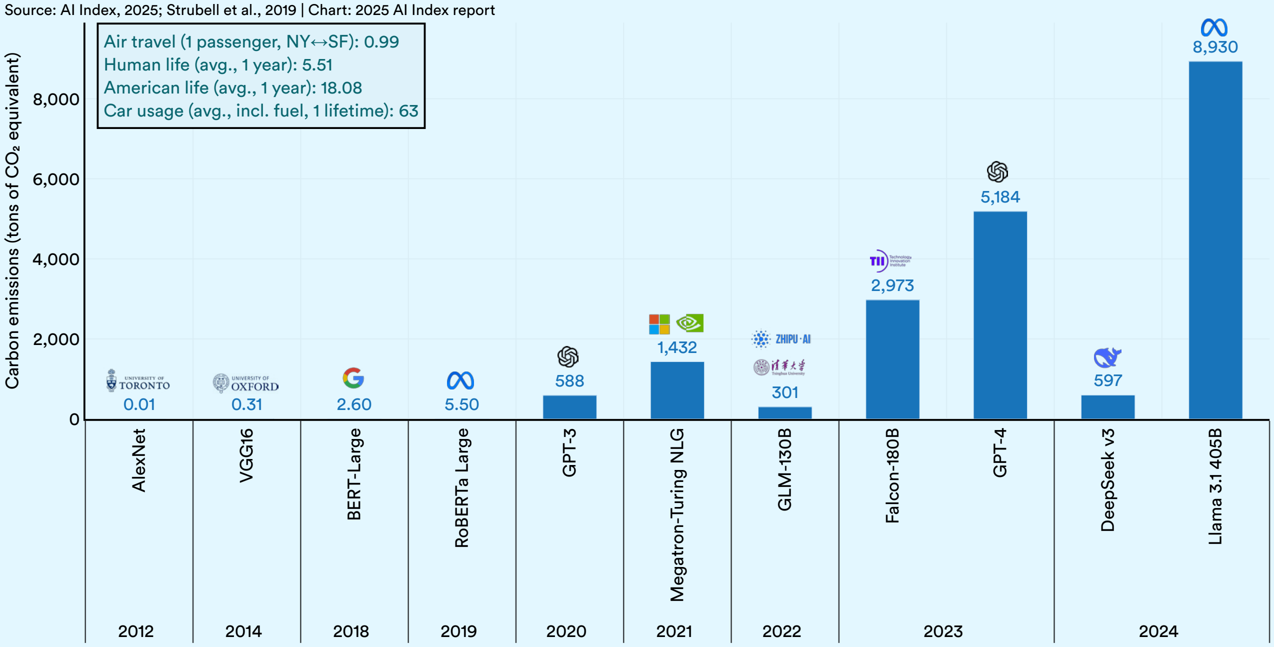


Figure 1.4.7

[31] The AI Index sourced input data—such as training hardware and duration—for the emissions calculator from various online sources. To validate the accuracy of the calculator, the Index compared the calculator's estimates with actual emissions reported by developers and found that the results were largely consistent. The full estimation methodology is detailed in the Appendix.

Highlight:

Energy Efficiency and Environmental Impact (cont'd)

Estimated carbon emissions and number of parameters by select AI models

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

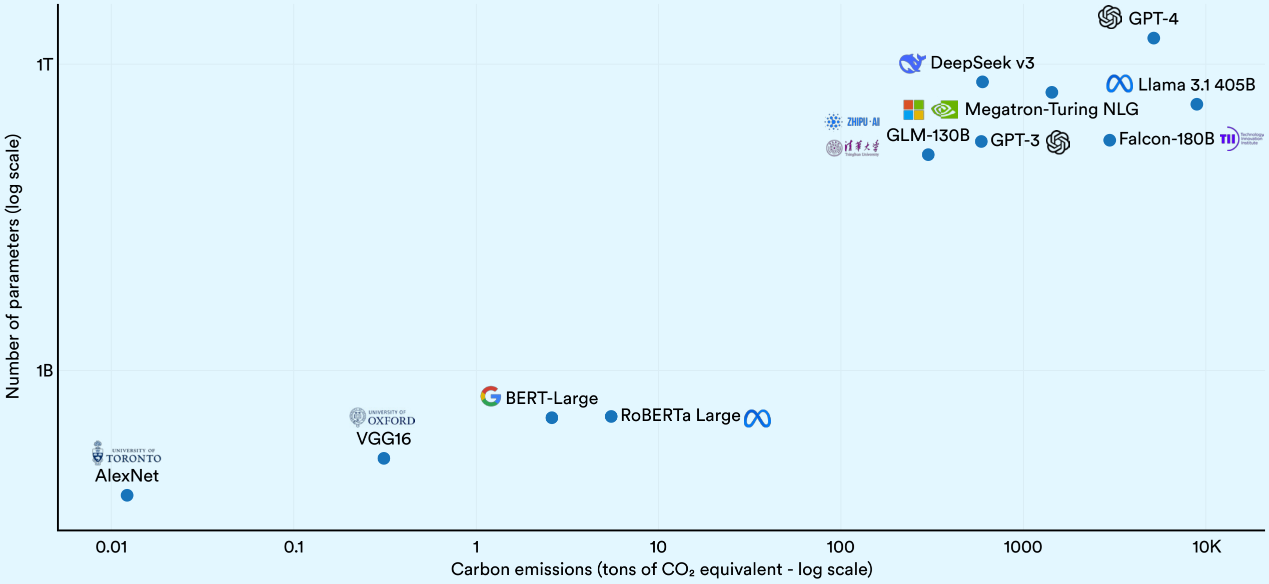


Figure 1.4.8

AI conferences serve as essential platforms for researchers to present their findings and network with peers and collaborators. Over the past two decades, these conferences have expanded in scale, quantity, and prestige. This section explores trends in attendance at major AI conferences.

1.5 AI Conferences

Conference Attendance

Figure 1.5.1 graphs attendance at a selection of AI conferences since 2010. In 2020 the pandemic forced conferences to be held fully online, increasing attendance significantly. This was followed by a decline in attendance, likely due to the shift back to in-person formats, returning attendance in 2022 to prepandemic levels. Since then, there has been a steady growth in conference attendance, increasing almost 21.7% from 2023 to 2024.[32] Since 2014, the annual number of attendees has risen by more than 60,000, reflecting not just a growing interest in AI research but also the emergence of new AI conferences.

Neural Information Processing Systems (NeurIPS) remains the most attended AI conference, attracting almost 20,000 participants in 2024 (Figure 1.5.2 and Figure 1.5.3). Among the major AI conferences, NeurIPS, CVPR, ICML, ICRA, ICLR, IROS and AAAI experienced increases in attendance over the last year.

Attendance at select AI conferences, 2010–24

Source: AI Index, 2024 | Chart: 2025 AI Index report

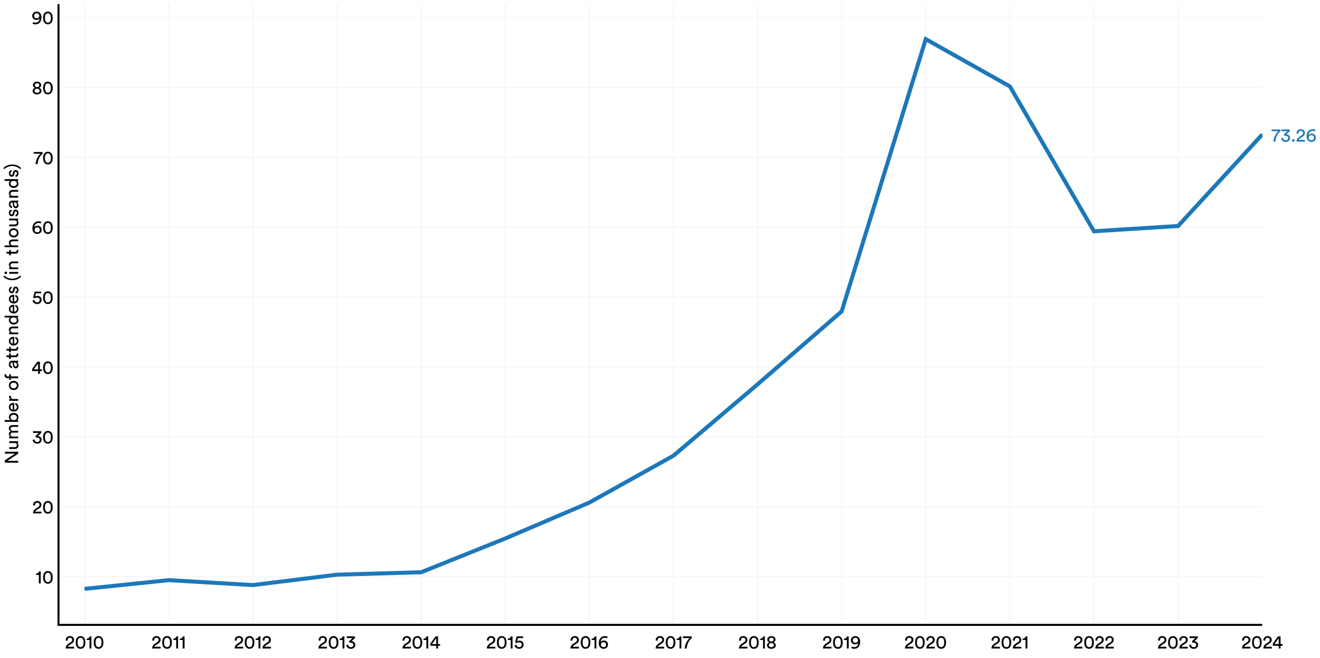


Figure 1.5.1

[32] This data should be interpreted with caution given that many conferences in the last few years have had 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 of researchers from around the world. The AI Index reports total attendance figures, encompassing virtual, hybrid, and in-person participation. The conferences for which the AI Index tracked data include AAAI, AAMAS, CVPR, EMNLP, FAccT, ICAPS, ICCV, ICLR, ICML, ICRA, IJCAI, IROS, KR, NeurIPS, and UAI.

Attendance at large conferences, 2010–24

Source: AI Index, 2024 | Chart: 2025 AI Index report

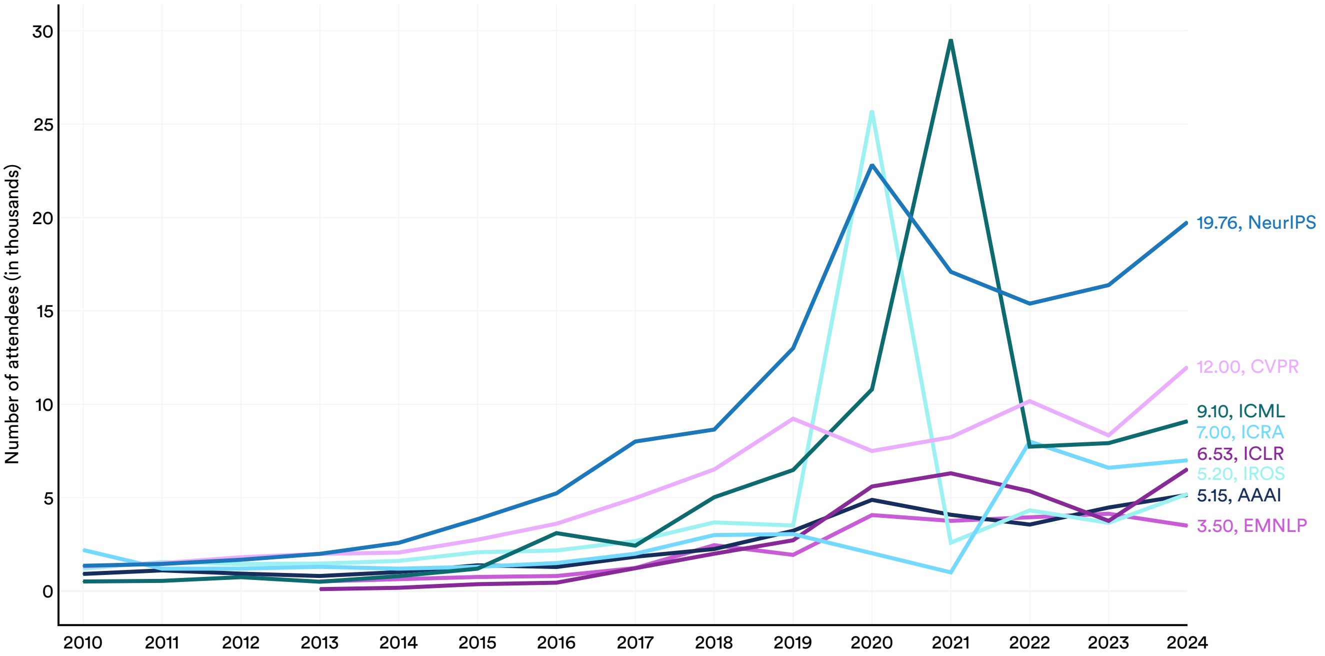


Figure 1.5.2[33]

Attendance at small conferences, 2010–24

Source: AI Index, 2024 | Chart: 2025 AI Index report

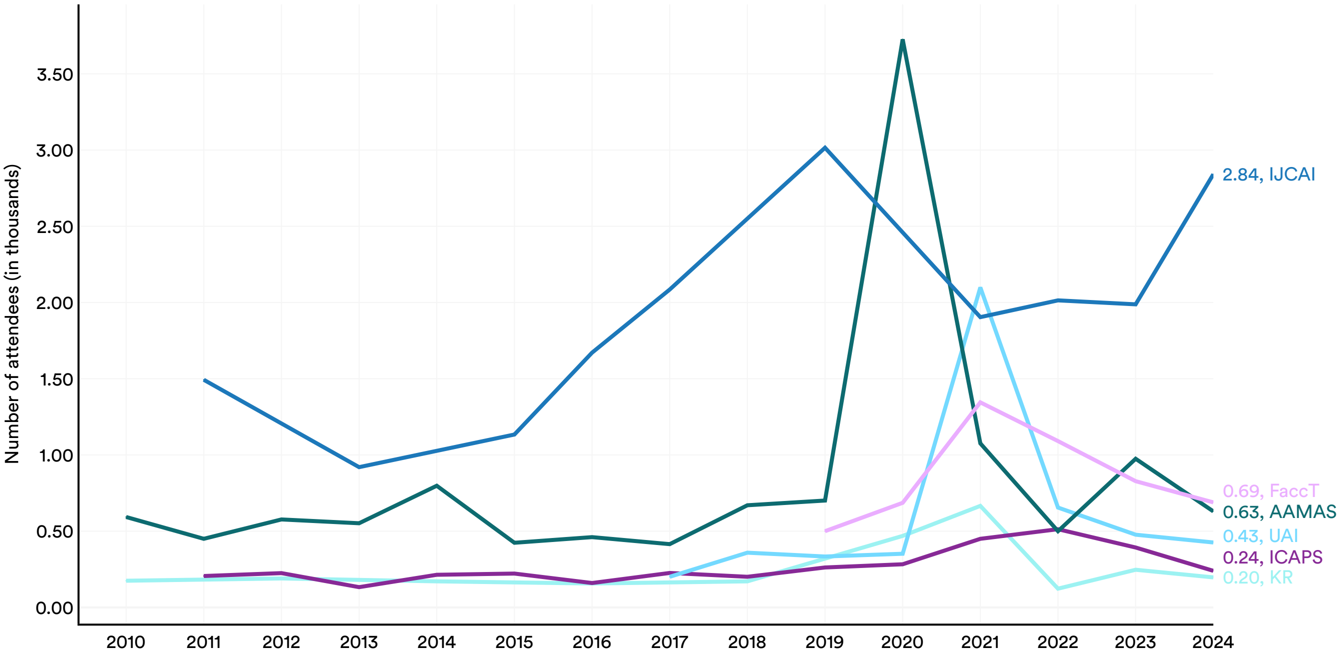


Figure 1.5.3

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

GitHub is a web-based platform that enables individuals and teams to host, review, and collaborate on code repositories. Widely used by software developers, GitHub facilitates code management, project collaboration, and open-source software support. This section draws on data from GitHub that provides insights into broader trends in open-source AI software development not reflected in academic publication data.[34]

1.6 Open-Source AI Software

Projects

A GitHub project comprises a collection of files, including source code, documentation, configuration files, and images, that together make up a software project. Figure 1.6.1 looks at the total number of GitHub AI projects over time.[35] Since 2011, the number of AI-related GitHub projects has consistently increased, growing from 1,549 in 2011 to approximately 4.3 million in 2024. Notably, there was a sharp 40.3% rise in the total number of GitHub AI projects in the last year alone.

Number of GitHub AI projects, 2011–24

Source: GitHub, 2024 | Chart: 2025 AI Index report

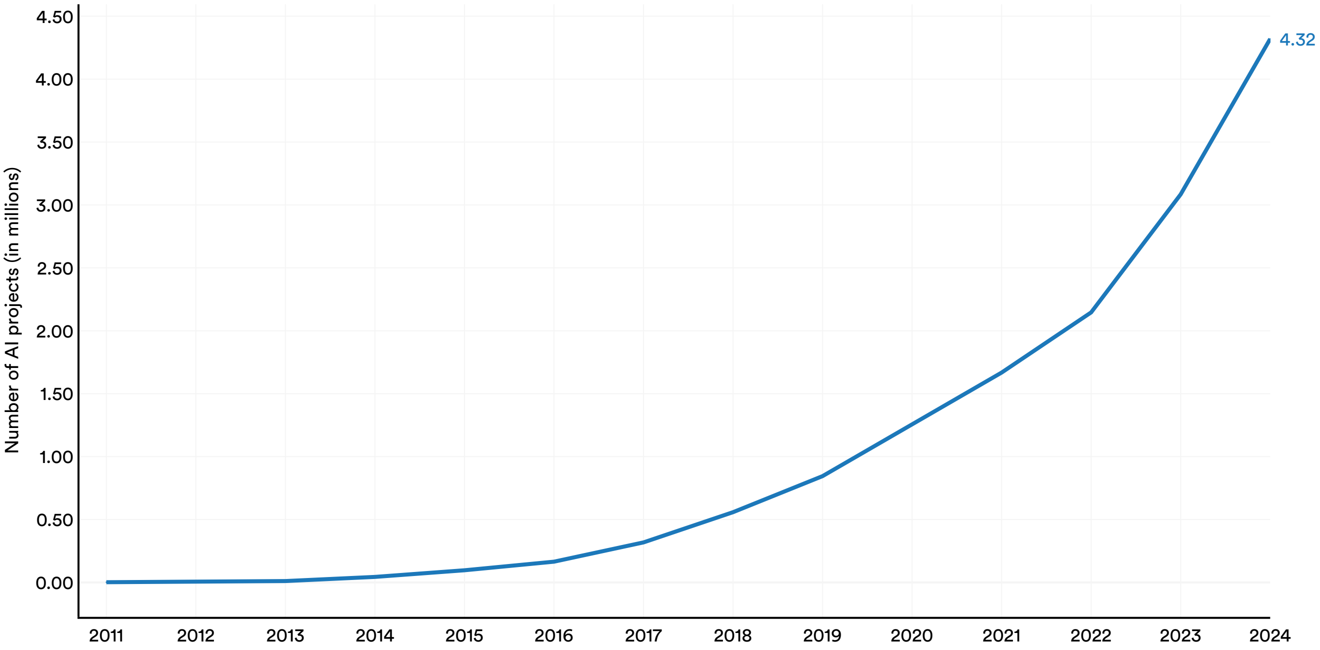


Figure 1.6.1

[34] This year, GitHub updated its methodology to capture a broader range of AI-related topics, including more recent developments. As a result, the figures in this year's AI Index may not align with those from previous editions. Chinese researchers often use alternative sites to GitHub for code sharing, such as Gitee and GitCode, but the data from those sites is not included in this report. A full methodological description is available in the Appendix.

[35] GitHub used AI-topic classification methods to identify AI-related repositories. Details on the methodology are available in the Appendix.

Figure 1.6.2 reports GitHub AI projects by geographic area since 2011. As of 2024, a significant share of GitHub AI projects were located in the United States, accounting for 23.4% of contributions. India was the second largest contributor with 19.9%, followed closely by Europe, which accounted for 19.5%. Notably, the share of open-source AI projects on GitHub from U.S.-based developers has declined since 2016 and appears to have stabilized in recent years.

GitHub AI projects (% of total) by geographic area, 2011–24

Source: GitHub, 2024 | Chart: 2025 AI Index report

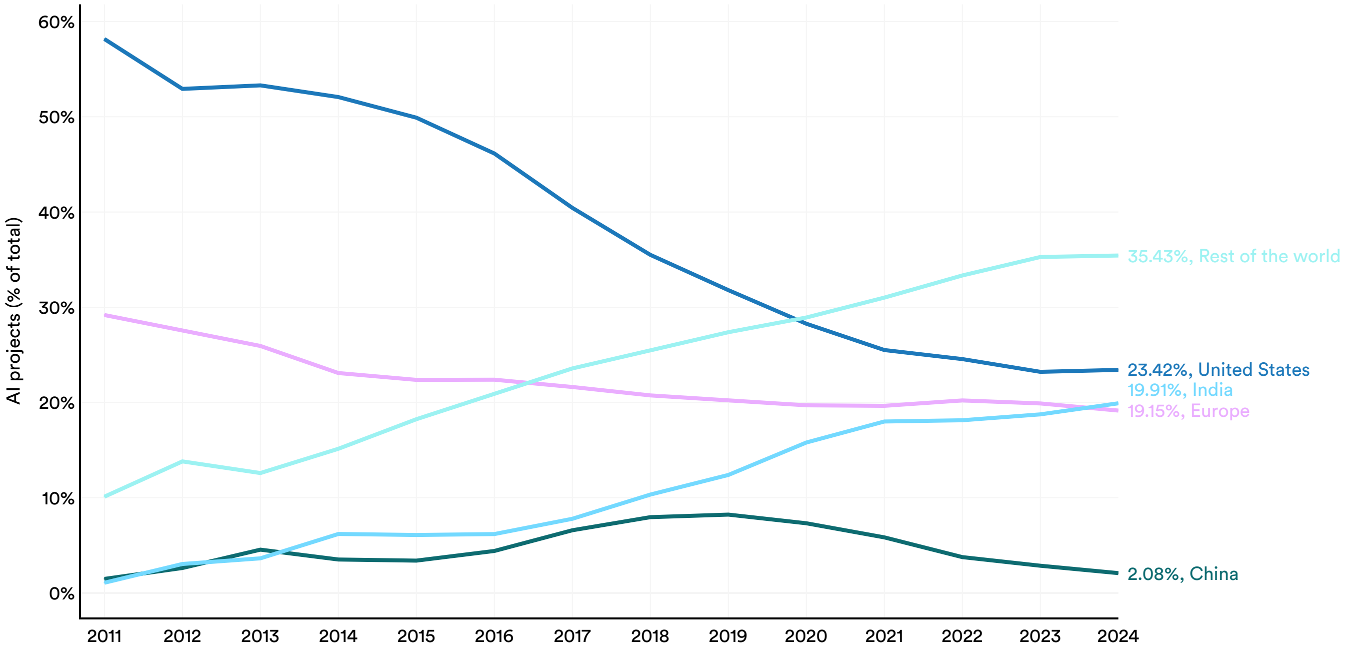


Figure 1.6.2

Stars

GitHub users can show their interest in a repository by "starring" it, a feature similar to liking a post on social media, which signifies support for an open-source project. Among the most starred repositories are libraries such as TensorFlow, OpenCV, Keras, and PyTorch, which enjoy widespread popularity among software developers in the broader developer community beyond AI. TensorFlow, Keras, and PyTorch are popular libraries for building and deploying machine learning models, while OpenCV offers a variety of tools for computer vision, such as object detection and feature extraction.

The total number of stars for AI-related projects on GitHub continued to rise last year, increasing from 14.0 million in 2023 to 17.7 million in 2024 (Figure 1.6.3).[36] This follows a particularly sharp rise from 2022 to 2023, when the total more than doubled.

Number of GitHub stars in AI projects, 2011–24

Source: GitHub, 2024 | Chart: 2025 AI Index report

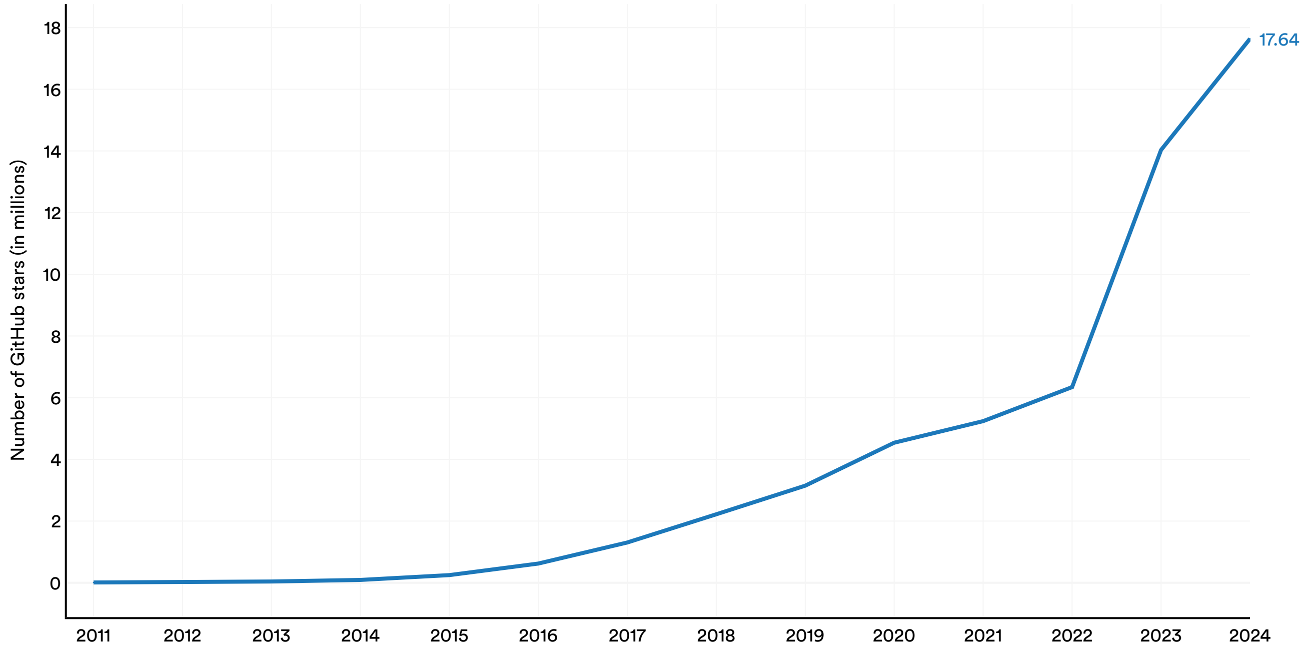


Figure 1.6.3

[36] Figure 1.6.3 shows new stars given to GitHub projects within a year, not the total accumulated over time.

In 2024, the United States led in receiving the highest number of GitHub stars, totaling 21.1 million (Figure 1.6.4). All major geographic regions sampled, including Europe, China, and India, saw a year-over-year increase in the total number of GitHub stars awarded to projects located in their countries.

Number of GitHub stars by geographic area, 2011–24

Source: GitHub, 2024 | Chart: 2025 AI Index report

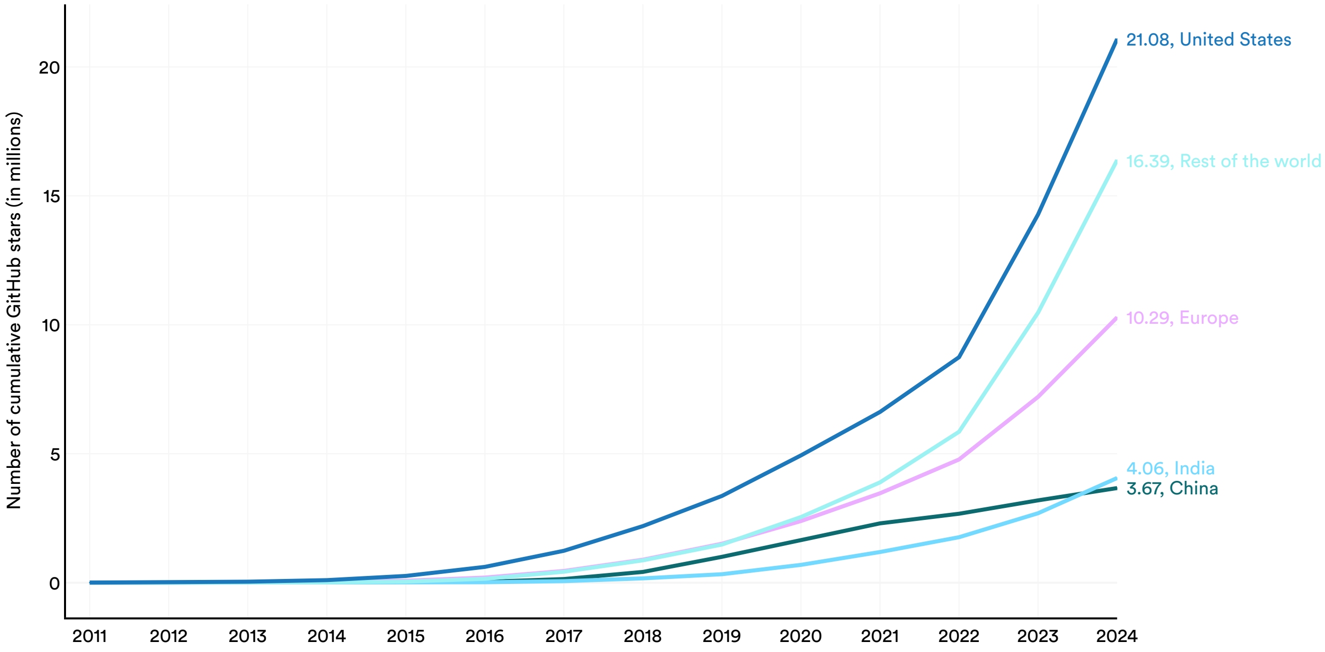


Figure 1.6.4

Appendix

Acknowledgments

The AI Index would like to acknowledge Angelo Salatino for his contributions to AI publication classification, Ben Cottier for leading the analysis of machine learning inference costs, Lapo Santarlasci for leading the analysis of AI patents, and Andrew Shi for leading the analysis of the environmental impact of AI models.

AI Publication Analysis

For this analysis, the AI Index used OpenAlex, an open scholarly database with over 260 million research publications, as its primary data source. OpenAlex classifies papers using its own knowledge organization system, known as OpenAlex Topics—a taxonomy of around 4,500 topics combining Scopus codes and CWTS classification. The system uses a deep learning model that considers titles, abstracts, journal names, and citation networks for classification. To identify AI-related topics more precisely, the AI Index analyzed computer science publications identified by OpenAlex and refined the classifications using the Computer Science Ontology and the CSO Classifier.

The Computer Science Ontology (CSO) is a large-scale, automatically generated ontology of research areas derived from 16 million publications using the Klink-2 algorithm. It features a hierarchical structure with thousands of subtopics, allowing for precise mapping of specific terms to broader research fields. Compared to general-purpose scholarly databases like OpenAlex, Scopus, and Web of Science, CSO offers a more detailed and fine-grained representation of the research landscape. As a result, it has been widely used for scholarly data exploration, analysis, modeling, and expert identification and recommendation. Version 3.4.1—used in this analysis—includes approximately 15,000 topics and 166,000 relationships within computer science. Released on Jan. 17, 2025, this version introduces over 150 new research topics in artificial intelligence, bringing the total to 2,369 AI-related topics and 12,620 hierarchical relationships within the AI domain alone.

To analyze research trends, the AI Index used the CSO Classifier—an unsupervised method that automatically categorizes research papers based on CSO topics. The classifier follows a three-stage pipeline that processes paper titles and abstracts: A syntactic module detects direct mentions of CSO topics; a semantic module uses word embeddings to identify related concepts; and a postprocessing module merges results, filters out irrelevant topics, and adds broader categories for a more refined classification. For this analysis, the AI Index extended the CSO Classifier to focus specifically on artificial intelligence and its subtopics. Since its initial release, the classifier has gained significant and growing interest due to its versatility. For example, Springer Nature uses it to routinely classify proceedings books, improving metadata quality. Beyond academic publishing, it has been successfully applied to categorize research software, YouTube videos, press releases, job ads, and IT museum collections.

Accurately categorizing research papers as either conference proceedings or journal articles is essential for this analysis. OpenAlex's metadata fields—type, crossref_type, and source_type—can sometimes conflict. To resolve these inconsistencies, the AI Index mapped OpenAlex records to DBLP, a leading bibliographic database for computer science publications. Known for its high metadata quality, DBLP continuously adds new publications through a rigorous, semiautomated curation process and currently indexes 3.6 million conference papers and 3 million journal articles. The initial matching between OpenAlex and DBLP was performed using DOIs. For remaining unmatched papers, the AI Index used a combination of title and publication year. To streamline this process, the AI Index built a title index to optimize search and ensure efficient mapping across the datasets.

AI publications are aggregated based on several parameters to provide a comprehensive analysis. Publications are grouped by year, considering the publication date of the most recent versions. Additionally, the AI Index groups publications by geographic areas or World Bank regions using the affiliations of authors. This means a single paper can contribute to multiple counts if coauthored by researchers from different countries, with each country receiving a count. When authors' affiliations are missing, these publications are mapped as "Unknown." Furthermore, sectors are associated with publications through authors' affiliations when available, which may lead to a publication being counted for multiple sectors. Citation counts are included when available; those without citation data are classified as "Unknown."

Top 100 Publications Analysis

The AI Index conducted a comprehensive analysis of influential AI publications by collecting and analyzing citation data from multiple sources including OpenAlex, Google Scholar, and Semantic Scholar. Initially gathering the top 150 most-cited papers per publication year from OpenAlex, the list was refined to 100 publications through careful review.

The methodology attributes publications to all countries and regions represented by authors' affiliations, meaning a single paper can contribute to multiple counts. For instance, a paper coauthored by researchers from the United States and China counts once for each country. This approach may result in overlapping totals in aggregate statistics. Publication years are based on the most recent versions, whether in journals, conferences, or repositories like arXiv. To maintain accuracy, organizational affiliations were verified and standardized, with countries assigned according to headquarters' locations.

The full list of the top 100 AI publications is available here.

AI Patent Analysis

The AI Index identifies AI-related patents using a hybrid classification approach, combining keyword-based text analysis with classification-code-based identification. Patent-level bibliographic data is sourced from PATSTAT Global, a comprehensive database issued by the European Patent Office (EPO). The analysis focuses on granted patents from 2010 onward, aggregated at the DOCDB family level to avoid duplicate counting of the same invention.[1] Patents are attributed to countries based on the publication authority of the earliest recorded grant publication.

Patent abstracts and titles originally published in languages other than English were translated using the deep-translator tool, Google Translate engine, and the Meta NLLB-200 machine translation model. Post-translation, patent texts were processed using natural language processing (NLP) techniques. These included the removal of stop words and special characters, part-of-speech (POS) tagging to retain key grammatical categories, lowercase conversion, lemmatization, and replacement of numerical measures with a <NUM> tag.

AI-related patents are identified by searching for relevant terms in patent titles and abstracts using regular expressions (regex). An AI-specific keyword dictionary was developed through a structured multistep process, incorporating keywords generated by AI models, expanded using established AI lexicons such as those from Yamashita et al. (2021), and refined through Word2Vec-based synonym identification. Further validation was conducted using BERTopic topic modeling and DeBERTA-based zero-shot classification, with manual checks applied to reduce false positives.

In addition to keyword-based classification, AI-related patents were identified using International Patent Classification (IPC) and Cooperative Patent Classification (CPC) codes. A curated list of AI-relevant codes was compiled through a combination of AI model analysis, regex-based searches, and prior research, including classifications from Pairolero et al. (2023) and WIPO (2024). The final dataset was constructed by merging results from both approaches, balancing coverage and accuracy.

[1] Despite this aggregation procedure, duplicates occasionally appear in marginal cases where applications within the same DOCDB family share the same earliest filing date. The AI Index removes duplicate values with respect to the aggregation variables (e.g., counting by year) when presenting analytics.

Epoch Notable Models Analysis

The AI forecasting research group Epoch AI maintains a dataset of landmark AI and ML models, along with accompanying information about their creators and publications, such as the list of their authors, number of citations, type of AI task accomplished, and amount of compute used in training.

The nationalities of the authors of these papers have important implications for geopolitical AI forecasting. As various research institutions and technology companies start producing advanced ML models, the global distribution of AI development may shift or concentrate in certain places, which in turn affects the geopolitical landscape because AI is expected to become a crucial component of economic and military power in the near future.

To track the distribution of AI research contributions on landmark publications by country, the Epoch dataset is coded according to the following methodology:

1. A snapshot of the dataset was taken in March 2025. This includes papers about landmark models, selected using the inclusion criteria of importance, relevance, and uniqueness, as described in the Compute Trends dataset documentation.
2. The authors are attributed to countries based on their affiliation credited on the paper. For international organizations, authors are attributed to the country where the organization is headquartered, unless a more specific location is indicated.
3. All of the landmark publications are aggregated within time periods (e.g., monthly or yearly) and the national contributions compiled to determine the extent of each country's contribution to landmark AI research during each time period.
4. The contributions of different countries are compared over time to identify any trends.

Training Cost Analysis

To create the dataset of cost estimates, the Epoch database was filtered for models released during the large-scale ML era[2] that were in the top 10 of training compute at the time of release. This filtered for the largest-scale ML models. The Transformer model was added to this set of models for further context.

For the selected ML models, the training time and the type, quantity, and hardware utilization rate were determined from the publication, press release, or technical reports, as applicable. Cloud rental prices for the computing hardware used by these models were collected from online historical archives of cloud vendors' websites.[3]

Training costs were estimated from the hardware type, quantity, and time by multiplying the hourly cloud rental rates (at the time of training)[4] by the quantity of hardware hours. However, some developers purchased hardware rather than renting cloud compute, and cloud prices vary by vendor and by rental commitment, so the true costs incurred by the developers may vary.

Various challenges were encountered while estimating the training cost of these models. Often, the developers did not disclose the duration of training or the hardware that was used. In other cases, cloud compute pricing was not available for the hardware. The investigation of training cost trends is more thoroughly detailed in a separate report by Epoch AI.

AI Conference Attendance

The AI Index reached out to the organizers of various AI conferences in 2024 and asked them to provide information on total attendance. For conferences that posted their attendance totals online, the AI Index used those reported totals and did not reach out to the conference organizers.

[2] The selected cutoff date was Sept. 1, 2015, in accordance with Compute Trends Across Three Eras of Machine Learning (Epoch, 2022).

[3] Historic prices were collected from archived snapshots of Amazon Web Services, Microsoft Azure, and Google Cloud Platform price catalogs viewed through the Internet Archive Wayback Machine.

[4] The chosen rental rate was the most recent published price for the hardware and cloud vendor used by the developer of the model, at a three-year commitment rental rate, after subtracting the training duration and two months from the publication date. If this price was not available, the most analogous price was used—either the same hardware and vendor at a different date, or the same hardware from a different cloud vendor. If a three-year commitment rental rate was unavailable, this was imputed from other rental rates based on the empirical average discount for the given cloud vendor. If the exact hardware type was not available (e.g., Nvidia A100 SXM4 40GB), a generalization was used (e.g., Nvidia A100).

GitHub

Identifying AI Projects

In partnership with researchers from Harvard Business School, Microsoft Research, and Microsoft's AI for Good Lab, GitHub identifies public AI repositories following the methodologies of Gonzalez, Zimmerman, and Nagappan (2020) and Dohmke, Iansiti, and Richards (2023), using topic labels related to AI/ML and generative AI, respectively, along with other relevant keywords identified through snowball sampling, such as "machine learning," "deep learning," and "artificial intelligence." GitHub further augments the dataset with repositories that have a dependency on the PyTorch, TensorFlow, OpenAI, Transformers, XGBoost, scikit-learn, and SciPy libraries for Python.

Mapping AI Projects to Geographic Areas

Public AI projects are mapped to geographic areas using IP address geolocation to determine the mode location of a project's owners each year. Each project owner is assigned a location based on their IP address when interacting with GitHub. If a project owner changes locations within a year, the location for the project would be determined by the mode location of its owners sampled daily in the year. Additionally, the last known location of the project owner is carried forward on a daily basis even if no activities were performed by the project owner that day. For example, if a project owner performed activities within the United States and then became inactive for six days, that project owner would be considered to be in the United States for that seven-day span.

Environmental Impact Analysis

The AI Index estimated the carbon emissions of training language and vision models using a calculator proposed by Lacoste et al. (2019). The analysis focused on the training stage emissions, excluding embodied hardware production, idle infrastructure, and deployment emissions. The study examined four model categories: industry language models, academic language models, industry vision models, and academic vision models.

The calculator's accuracy was verified against published emission values. Calculator inputs included hardware type, GPU hours, provider, and compute region. For newer hardware like the H100 GPU (released in 2022), the A100 SXM4 80GB was used as a substitute in calculations. Provider selection was based on known partnerships (e.g., Google models using GCP, OpenAI using Azure), while compute regions were determined by team locations.

Special consideration was given to models trained on custom hardware, such as BLOOM's use of the Jean Zay supercomputer in France. In these cases, private infrastructure calculations incorporated carbon efficiency (kg/kWh) and offset percentages.

The study evaluated 50 models in total: 34 industry language models (2018–24), eight industry vision models (2019–23), four academic language models (2020–23), and four academic vision models (2011–22), selecting particularly influential models in their respective domains.