Artificial Intelligence
Index Report 2023

CHAPTER 1:
Research and Development

CHAPTER 1 PREVIEW:

Research and Development

ACCESS THE PUBLIC DATA

Overview

This chapter captures trends in AI R&D. It begins by examining AI publications, including journal articles, conference papers, and repositories. Next it considers data on significant machine learning systems, including large language and multimodal models. Finally, the chapter concludes by looking at AI conference attendance and open-source AI research. Although the United States and China continue to dominate AI R&D, research efforts are becoming increasingly geographically dispersed.

Chapter Highlights

The United States and China had the greatest number of cross-country collaborations in AI publications from 2010 to 2021, although the pace of collaboration has since slowed.

The number of AI research collaborations between the United States and China increased roughly 4 times since 2010, and was 2.5 times greater than the collaboration totals of the next nearest country pair, the United Kingdom and China. However, the total number of U.S.-China collaborations only increased by 2.1% from 2020 to 2021, the smallest year-over-year growth rate since 2010.

AI research is on the rise, across the board.

The total number of AI publications has more than doubled since 2010. The specific AI topics that continue to dominate research include pattern recognition, machine learning, and computer vision.

China continues to lead in total AI journal, conference, and repository publications.

The United States is still ahead in terms of AI conference and repository citations, but those leads are slowly eroding. Still, the majority of the world's large language and multimodal models (54% in 2022) are produced by American institutions.

Industry races ahead of academia.

Until 2014, most significant machine learning models were released by academia. Since then, industry has taken over. In 2022, there were 32 significant industry-produced machine learning models compared to just three produced by academia. Building state-of-the-art AI systems increasingly requires large amounts of data, computer power, and money—resources that industry actors inherently possess in greater amounts compared to nonprofits and academia.

Large language models are getting bigger and more expensive.

GPT-2, released in 2019, considered by many to be the first large language model, had 1.5 billion parameters and cost an estimated $50,000 USD to train. PaLM, one of the flagship large language models launched in 2022, had 540 billion parameters and cost an estimated $8 million USD—PaLM was around 360 times larger than GPT-2 and cost 160 times more. It's not just PaLM: Across the board, large language and multimodal models are becoming larger and pricier.

This section draws on data from the Center for Security and Emerging Technology (CSET) at Georgetown University. CSET maintains a merged corpus of scholarly literature that includes Digital Science's Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers With Code. In that corpus, CSET applied a classifier to identify English-language publications related to the development or application of AI and ML since 2010. For this year's report, CSET also used select Chinese AI keywords to identify Chinese-language AI papers; CSET did not deploy this method for previous iterations of the AI Index report.[1]

In last year's edition of the report, publication trends were reported up to the year 2021. However, 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 2021, which we, along with CSET, are confident yields a more fully representative report.

1.1 Publications

Overview

The figures below capture the total number of English-language and Chinese-language AI publications globally from 2010 to 2021—by type, affiliation, cross-country collaboration, and cross-industry collaboration. The section also breaks down publication and citation data by region for AI journal articles, conference papers, repositories, and patents.

Total Number of AI Publications

Figure 1.1.1 shows the number of AI publications in the world. From 2010 to 2021, the total number of AI publications more than doubled, growing from 200,000 in 2010 to almost 500,000 in 2021.

Number of AI Publications in the World, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

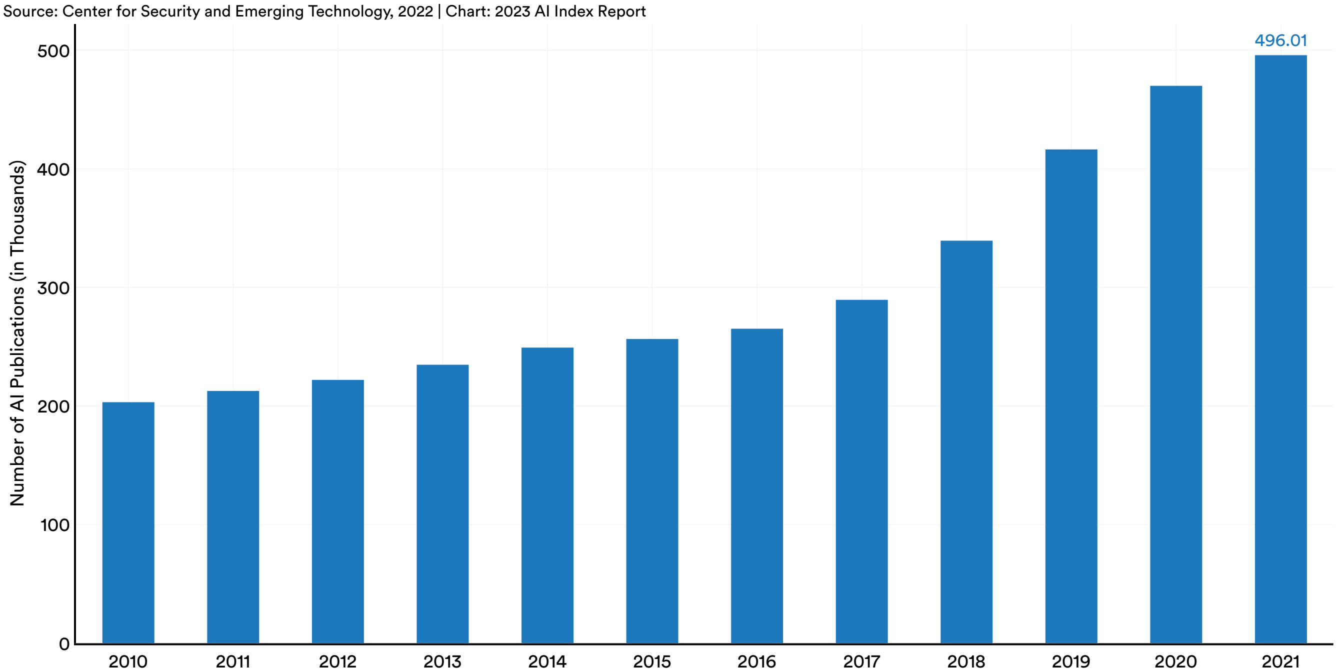

Figure 1.1.1

1 See the Appendix for more information on CSET's methodology. For more on the challenge of defining AI and correctly capturing relevant bibliometric data, see the AI Index team's discussion in the paper "Measurement in AI Policy: Opportunities and Challenges."

By Type of Publication

Figure 1.1.2 shows the types of AI publications released globally over time. In 2021, 60% of all published AI documents were journal articles, 17% were conference papers, and 13% were repository submissions. Books, book chapters, theses, and unknown document types made up the remaining 10% of publications. While journal and repository publications have grown 3 and 26.6 times, respectively, in the past 12 years, the number of conference papers has declined since 2019.

Number of AI Publications by Type, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

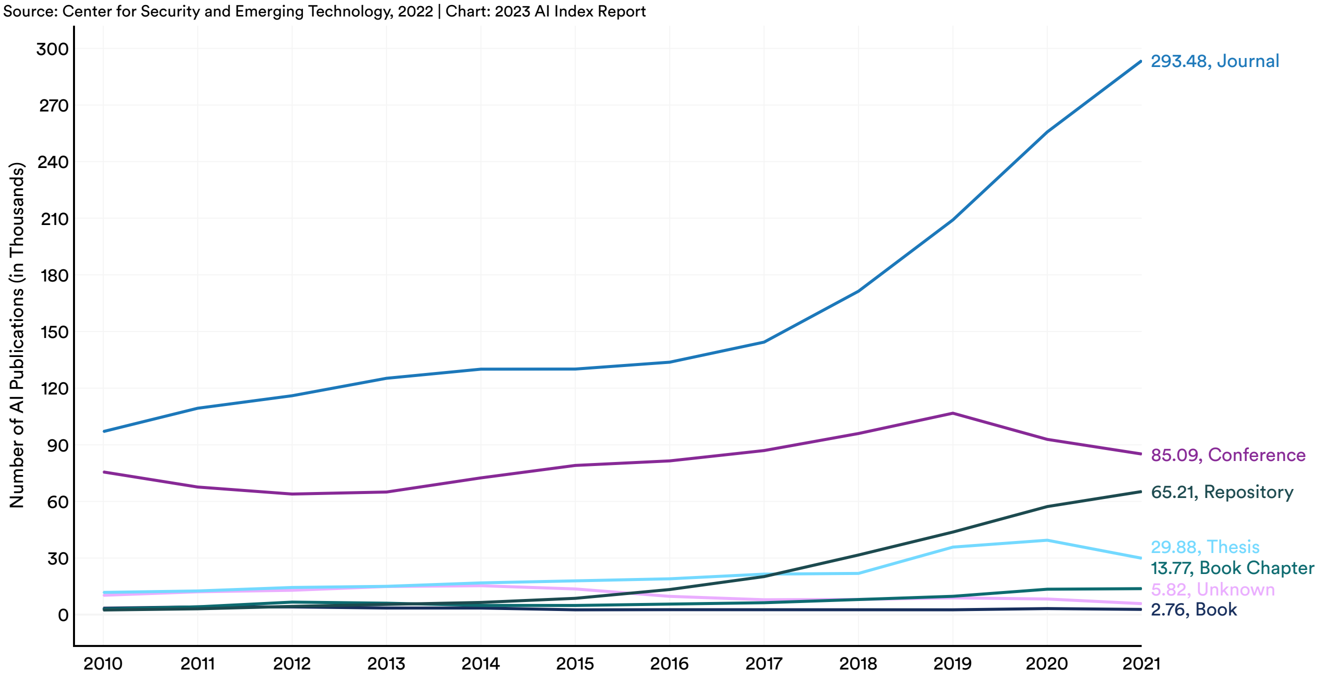

Figure 1.1.2

By Field of Study

Figure 1.1.3 shows that publications in pattern recognition and machine learning have experienced the sharpest growth in the last half decade. Since 2015, the number of pattern recognition papers has roughly doubled while the number of machine learning papers has roughly quadrupled. Following those two topic areas, in 2021, the next most published AI fields of study were computer vision (30,075), algorithm (21,527), and data mining (19,181).

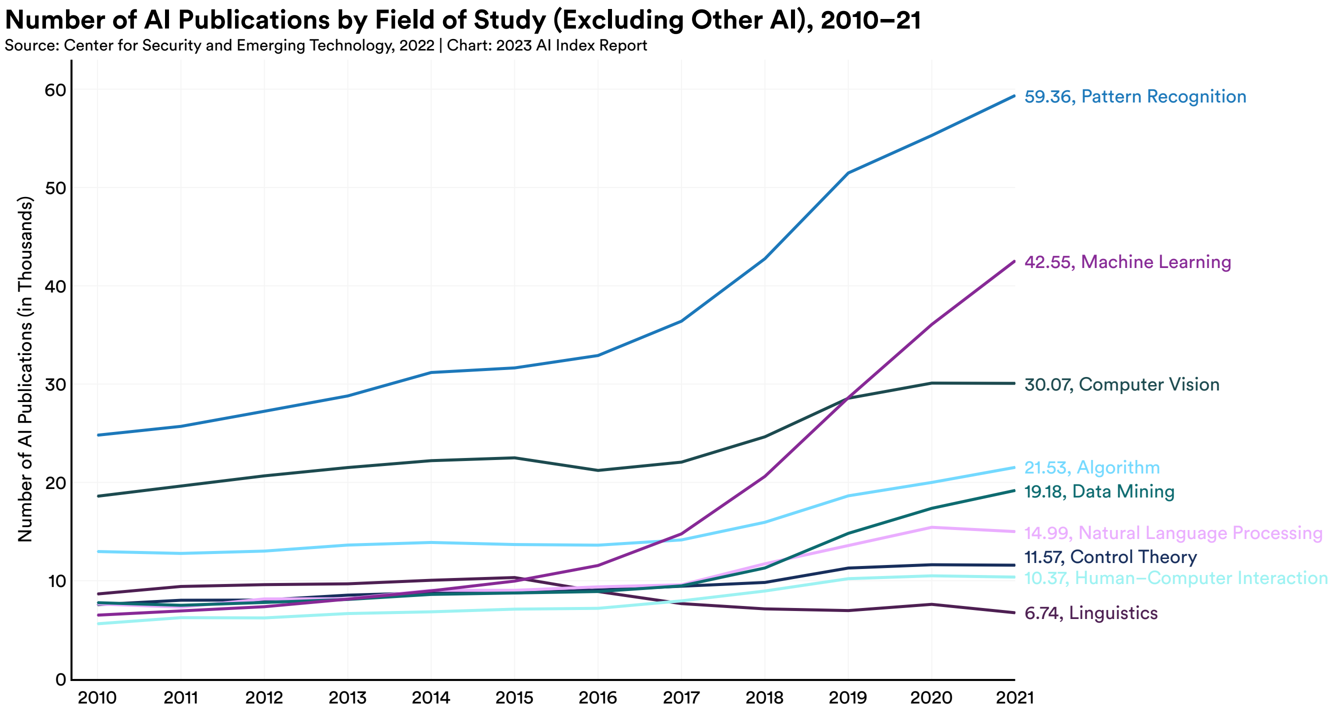

Figure 1.1.3

By Sector

This section shows the number of AI publications affiliated with education, government, industry, nonprofit, and other sectors—first globally (Figure 1.1.4), then looking at the United States, China, and the European Union plus the United Kingdom (Figure 1.1.5).[2] The education sector dominates in each region. The level of industry participation is highest in the United States, then in the European Union. Since 2010, the share of education AI publications has been dropping in each region.

AI Publications (% of Total) by Sector, 2010–21

Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

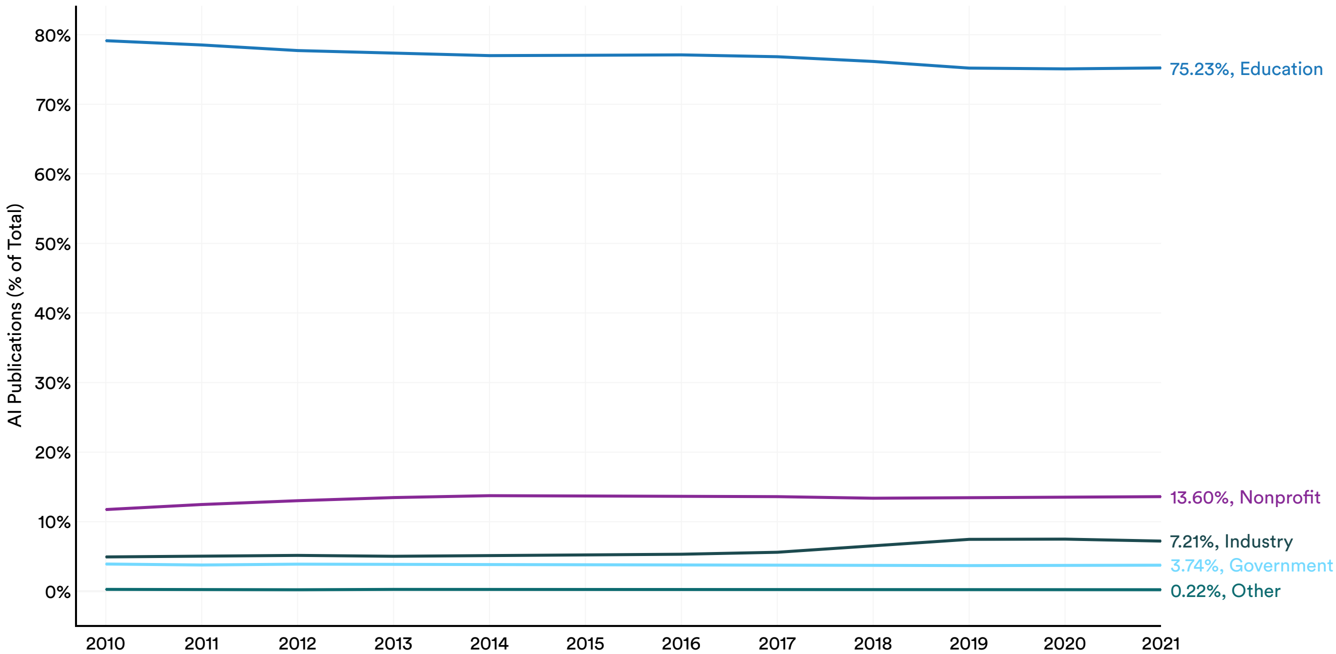

Figure 1.1.4

2 The categorization is adapted based on the Global Research Identifier Database (GRID). Healthcare, including hospitals and facilities, is included under nonprofit. Publications affiliated with state-sponsored universities are included in the education sector.

AI Publications (% of Total) by Sector and Geographic Area, 2021

Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

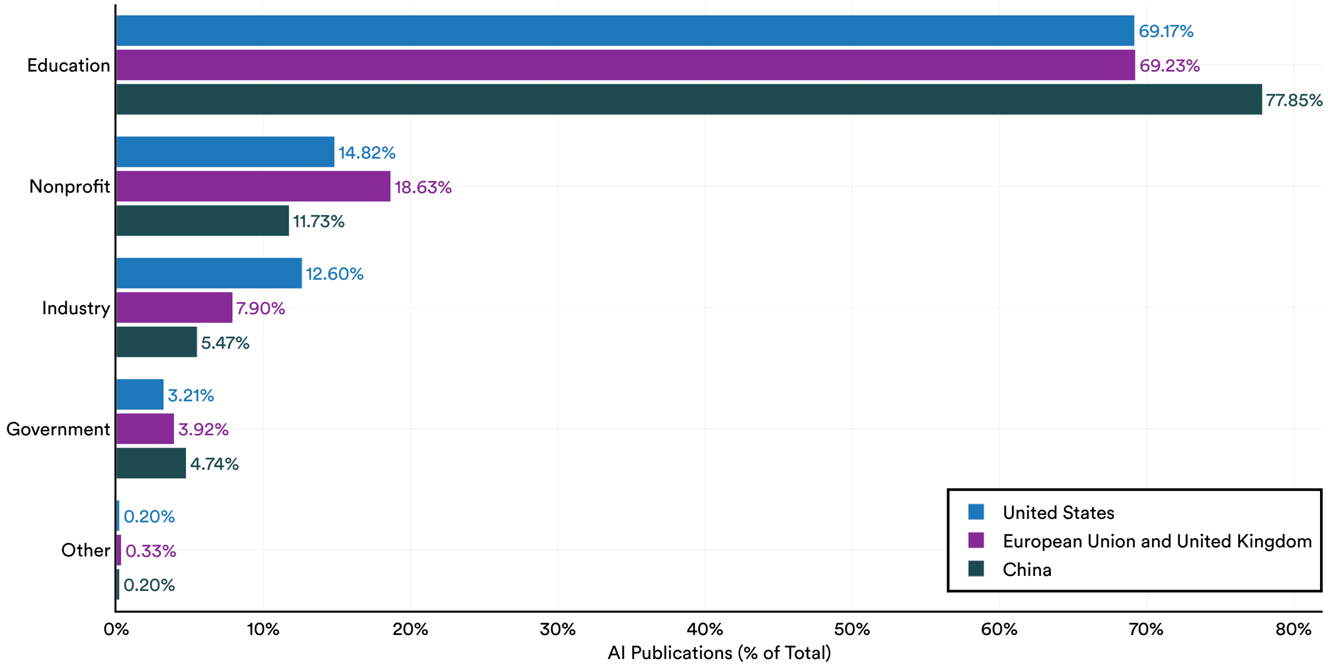

Figure 1.1.5

Cross-Country Collaboration

Cross-border collaborations between academics, researchers, industry experts, and others are a key component of modern STEM (science, technology, engineering, and mathematics) development that accelerate the dissemination of new ideas and the growth of research teams. Figures 1.1.6 and 1.1.7 depict the top cross-country AI collaborations from 2010 to 2021. CSET counted cross-country collaborations as distinct pairs of countries across authors for each publication (e.g., four U.S. and four Chinese-affiliated authors on a single publication are counted as one U.S.-China collaboration; two publications between the same authors count as two collaborations).

By far, the greatest number of collaborations in the past 12 years took place between the United States and China, increasing roughly four times since 2010. However the total number of U.S.-China collaborations only increased by 2.1% from 2020 to 2021, the smallest year-over-year growth rate since 2010.

The next largest set of collaborations was between the United Kingdom and both China and the United States. In 2021, the number of collaborations between the United States and China was 2.5 times greater than between the United Kingdom and China.

United States and China Collaborations in AI Publications, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

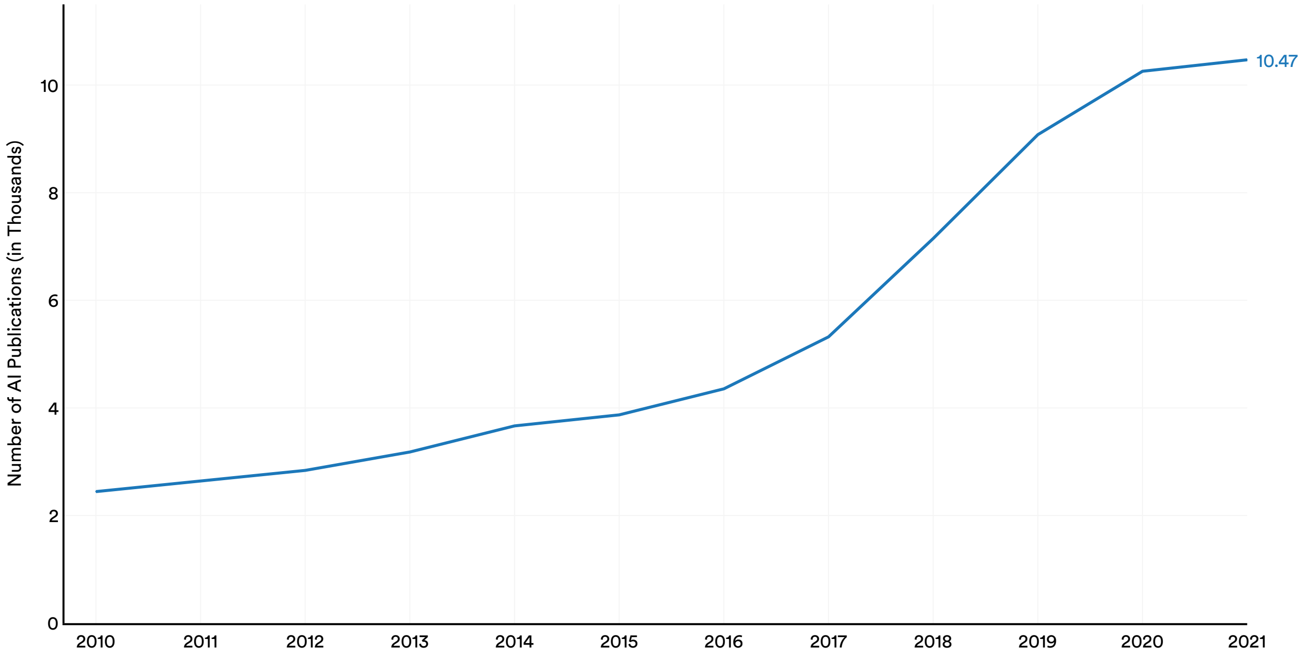

Figure 1.1.6

Cross-Country Collaborations in AI Publications (Excluding U.S. and China), 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

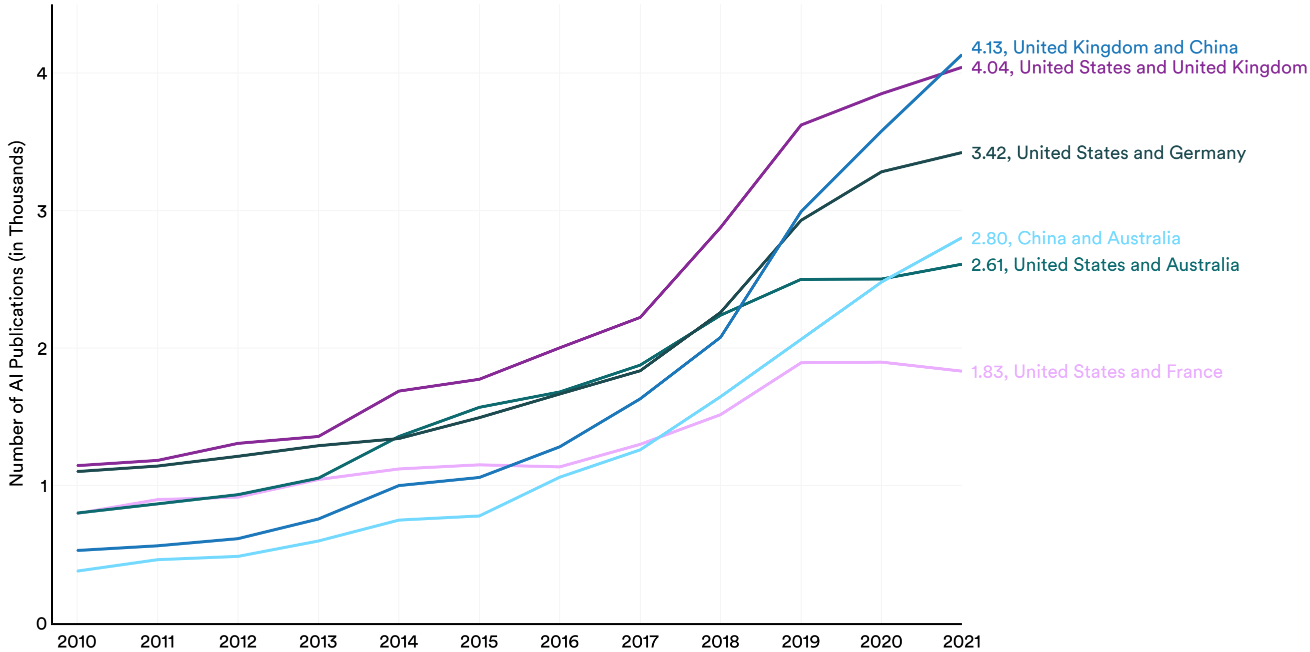

Figure 1.1.7

Cross-Sector Collaboration

The increase in AI research outside of academia has broadened and grown collaboration across sectors in general. Figure 1.1.8 shows that in 2021 educational institutions and nonprofits (32,551) had the greatest number of collaborations; followed by industry and educational institutions (12,856); and educational and government institutions (8,913). Collaborations between educational institutions and industry have been among the fastest growing, increasing 4.2 times since 2010.

Cross-Sector Collaborations in AI Publications, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

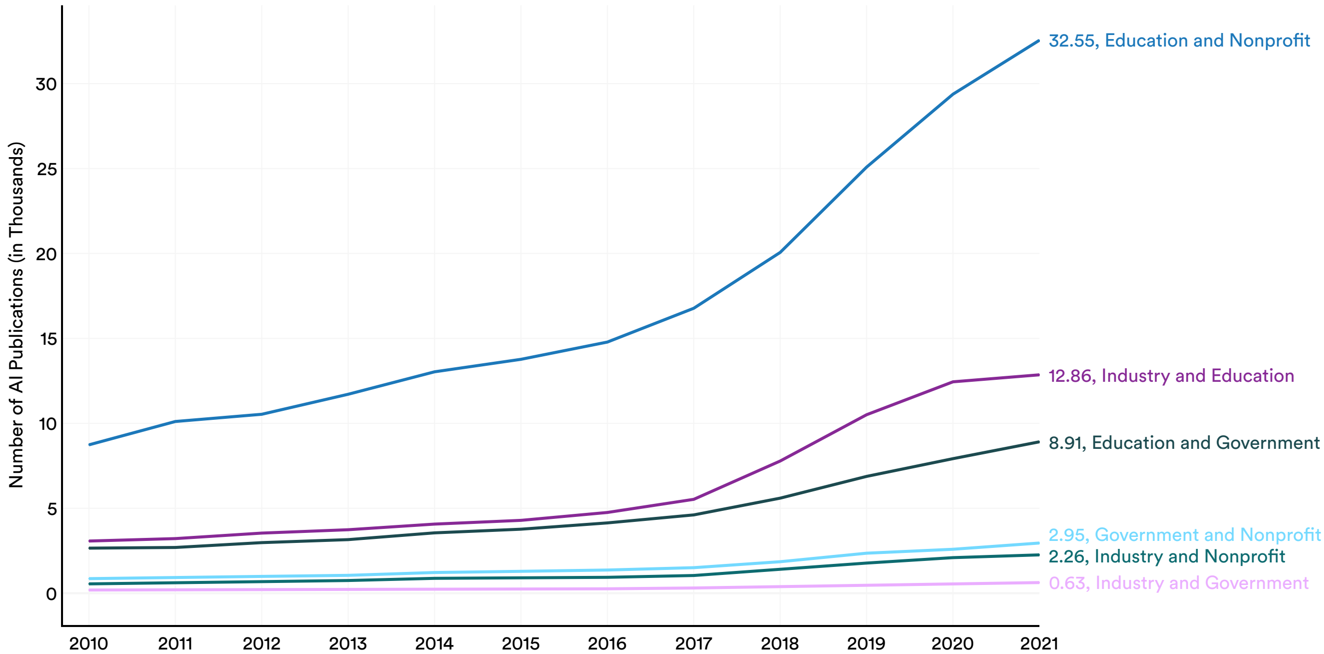

Figure 1.1.8

AI Journal Publications

Overview

After growing only slightly from 2010 to 2015, the number of AI journal publications grew around 2.3 times since 2015. From 2020 to 2021, they increased 14.8% (Figure 1.1.9).

Number of AI Journal Publications, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

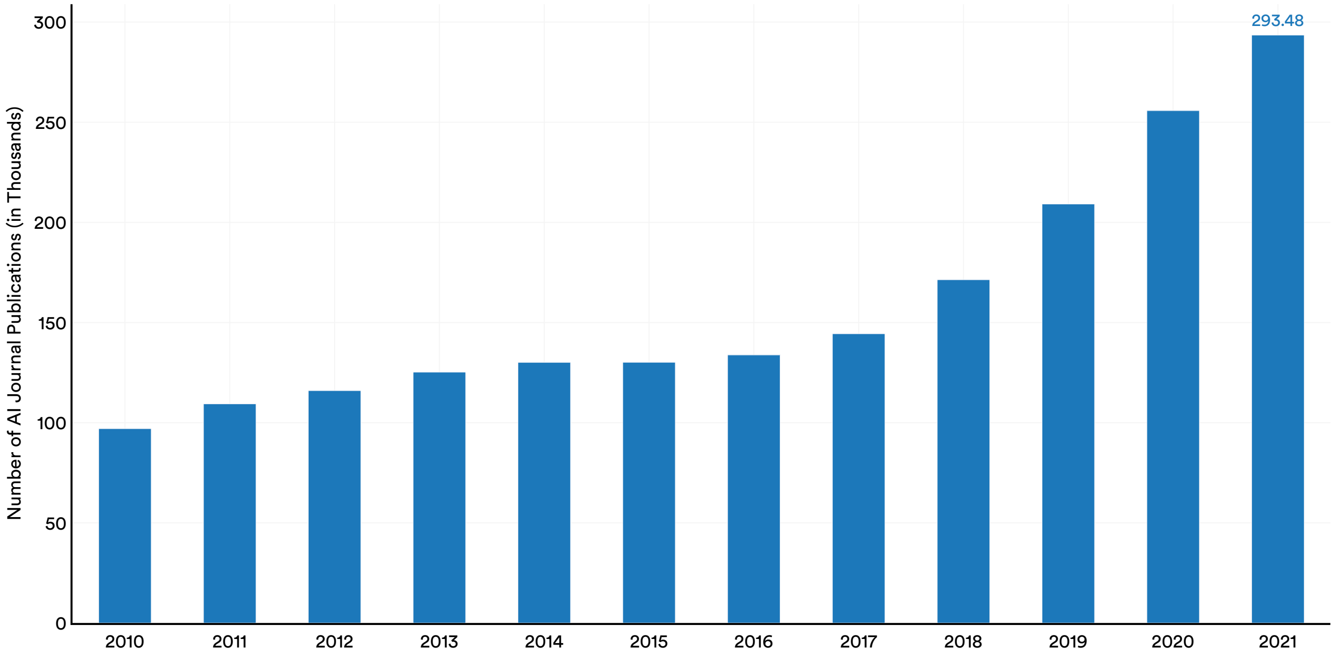

Figure 1.1.9

By Region[3]

Figure 1.1.10 shows the share of AI journal publications by region between 2010 and 2021. In 2021, East Asia and the Pacific led with 47.1%, followed by Europe and Central Asia (17.2%), and then North America (11.6%). Since 2019, the share of publications from East Asia and the Pacific; Europe and Central Asia; as well as North America have been declining. During that period, there has been an increase in publications from other regions such as South Asia; and the Middle East and North Africa.

AI Journal Publications (% of World Total) by Region, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

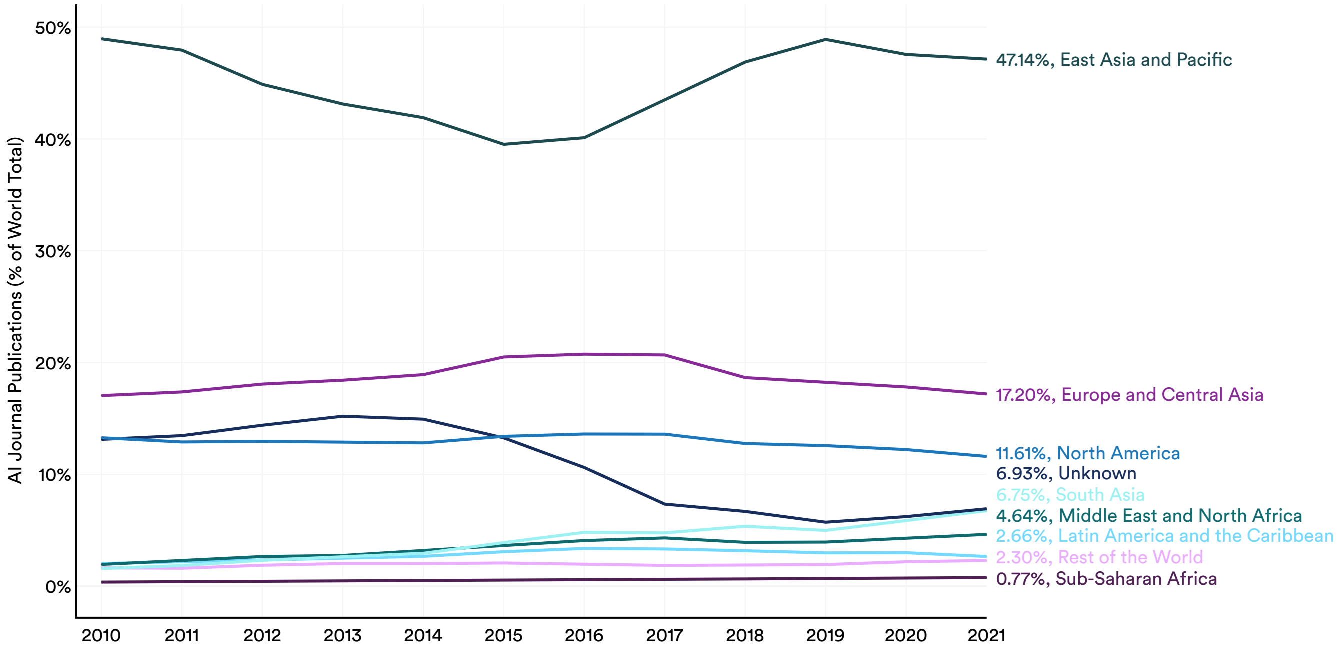

Figure 1.1.10

3 Regions in this chapter are classified according to the World Bank analytical grouping.

By Geographic Area[4]

Figure 1.1.11 breaks down the share of AI journal publications over the past 12 years by geographic area. This year's AI Index included India in recognition of the increasingly important role it plays in the AI ecosystem. China has remained the leader throughout, with 39.8% in 2021, followed by the European Union and the United Kingdom (15.1%), then the United States (10.0%). The share of Indian publications has been steadily increasing—from 1.3% in 2010 to 5.6% in 2021.

AI Journal Publications (% of World Total) by Geographic Area, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

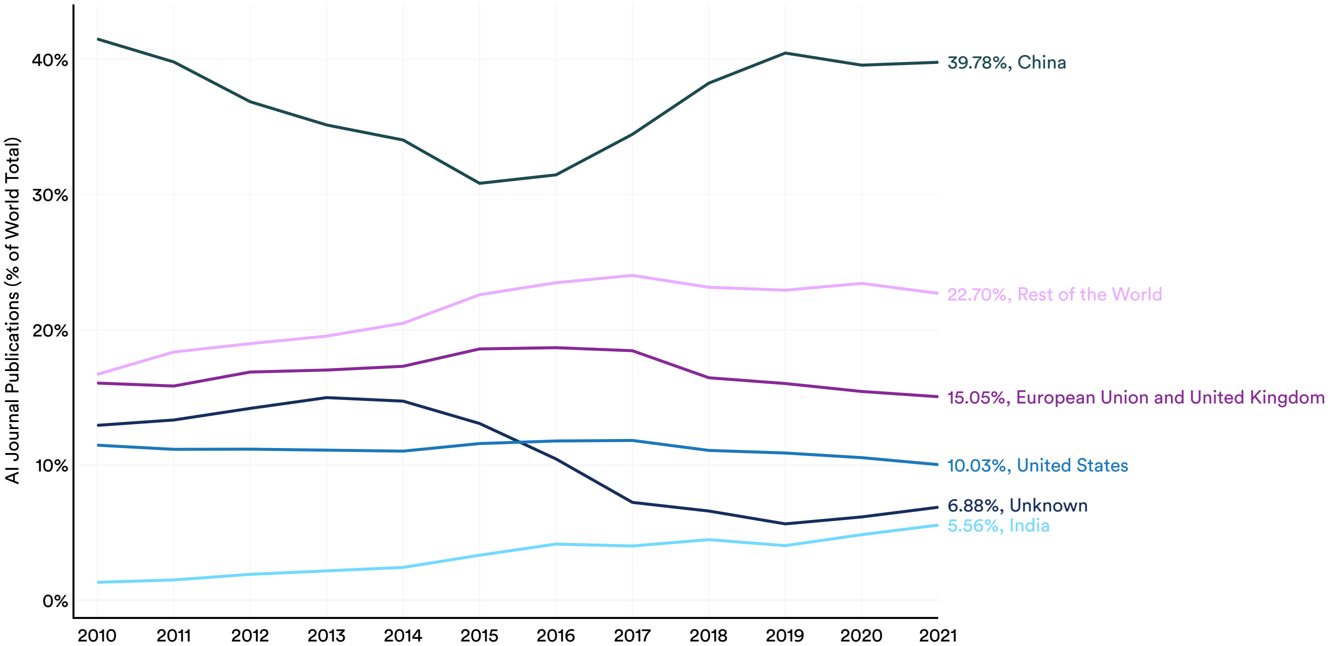

Figure 1.1.11

4 In this chapter we use "geographic area" based on CSET's classifications, which are disaggregated not only by country, but also by territory. Further, we count the European Union and the United Kingdom as a single geographic area to reflect the regions' strong history of research collaboration.

Citations

China's share of citations in AI journal publications has gradually increased since 2010, while those of the European Union and the United Kingdom, as well as those of the United States, have decreased (Figure 1.1.12). China, the European Union and the United Kingdom, and the United States accounted for 65.7% of the total citations in the world.

AI Journal Citations (% of World Total) by Geographic Area, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

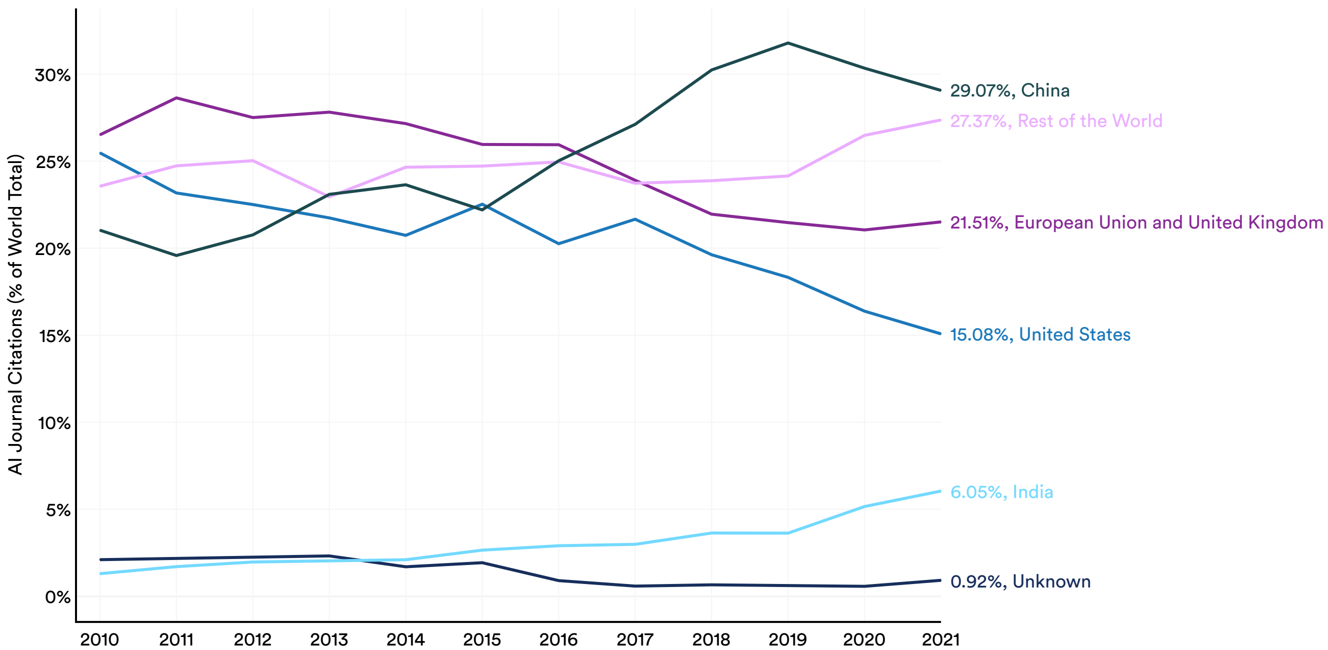

Figure 1.1.12

AI Conference Publications

Overview

The number of AI conference publications peaked in 2019, and fell 20.4% below the peak in 2021 (Figure 1.1.13). The total number of 2021 AI conference publications, 85,094, was marginally greater than the 2010 total of 75,592.

Number of AI Conference Publications, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

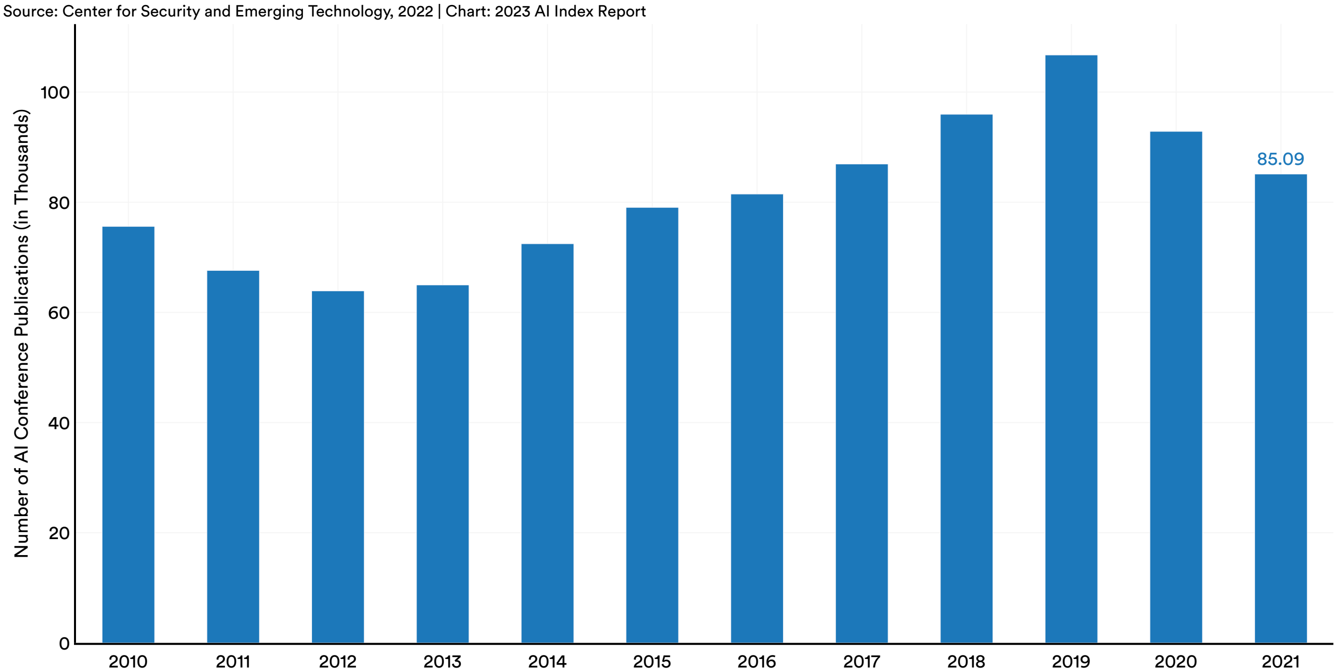

Figure 1.1.13

By Region

Figure 1.1.14 shows the number of AI conference publications by region. As with the trend in journal publications, East Asia and the Pacific; Europe and Central Asia; and North America account for the world's highest numbers of AI conference publications. Specifically, the share represented by East Asia and the Pacific continues to rise, accounting for 36.7% in 2021, followed by Europe and Central Asia (22.7%), and then North America (19.6%). The percentage of AI conference publications in South Asia saw a noticeable rise in the past 12 years, growing from 3.6% in 2010 to 8.5% in 2021.

AI Conference Publications (% of World Total) by Region, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

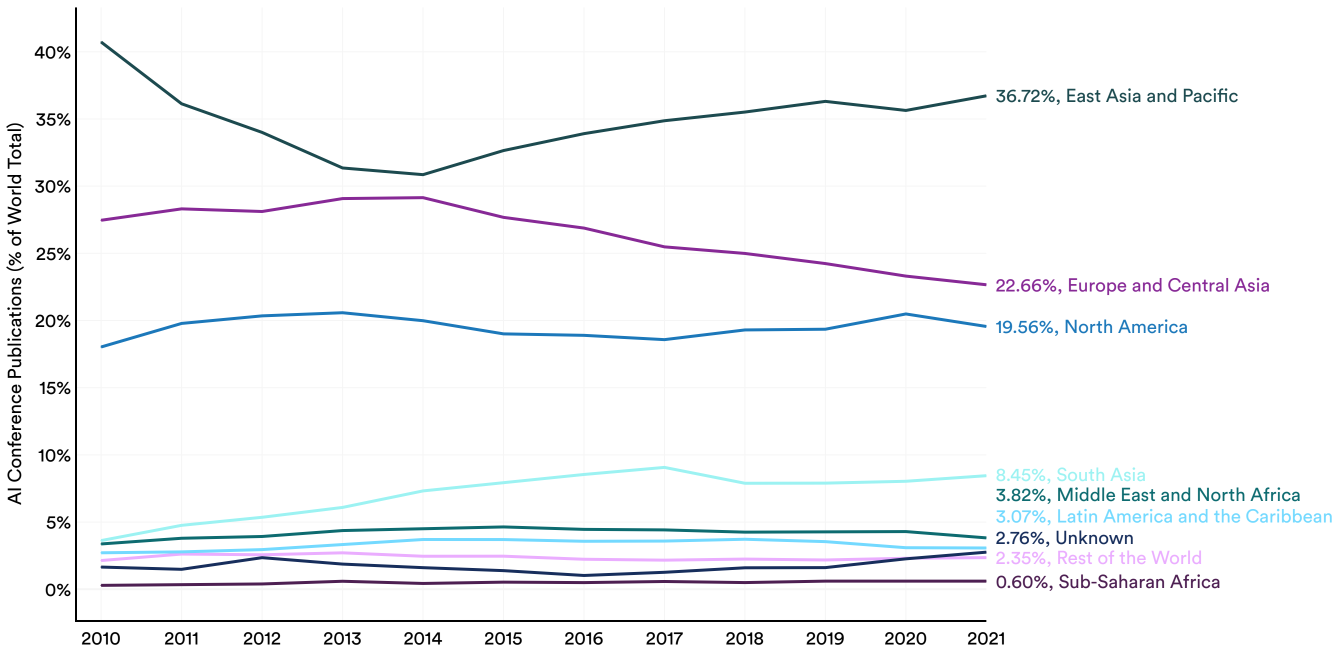

Figure 1.1.14

By Geographic Area

In 2021, China produced the greatest share of the world's AI conference publications at 26.2%, having overtaken the European Union and the United Kingdom in 2017. The European Union plus the United Kingdom followed at 20.3%, and the United States came in third at 17.2% (Figure 1.1.15). Mirroring trends seen in other parts of the research and development section, India's share of AI conference publications is also increasing.

AI Conference Publications (% of World Total) by Geographic Area, 2010–21

Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

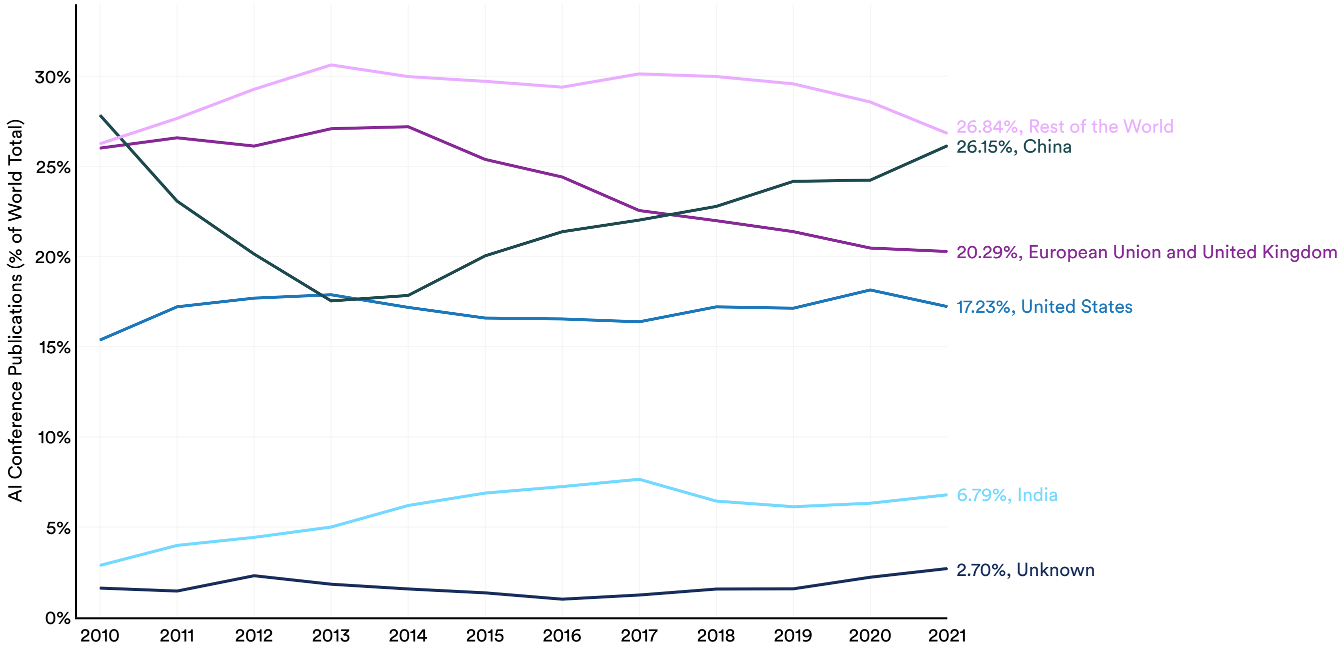

Figure 1.1.15

Citations

Despite China producing the most AI conference publications in 2021, Figure 1.1.16 shows that the United States had the greatest share of AI conference citations, with 23.9%, followed by China's 22.0%. However, the gap between American and Chinese AI conference citations is narrowing.

AI Conference Citations (% of World Total) by Geographic Area, 2010–21

Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

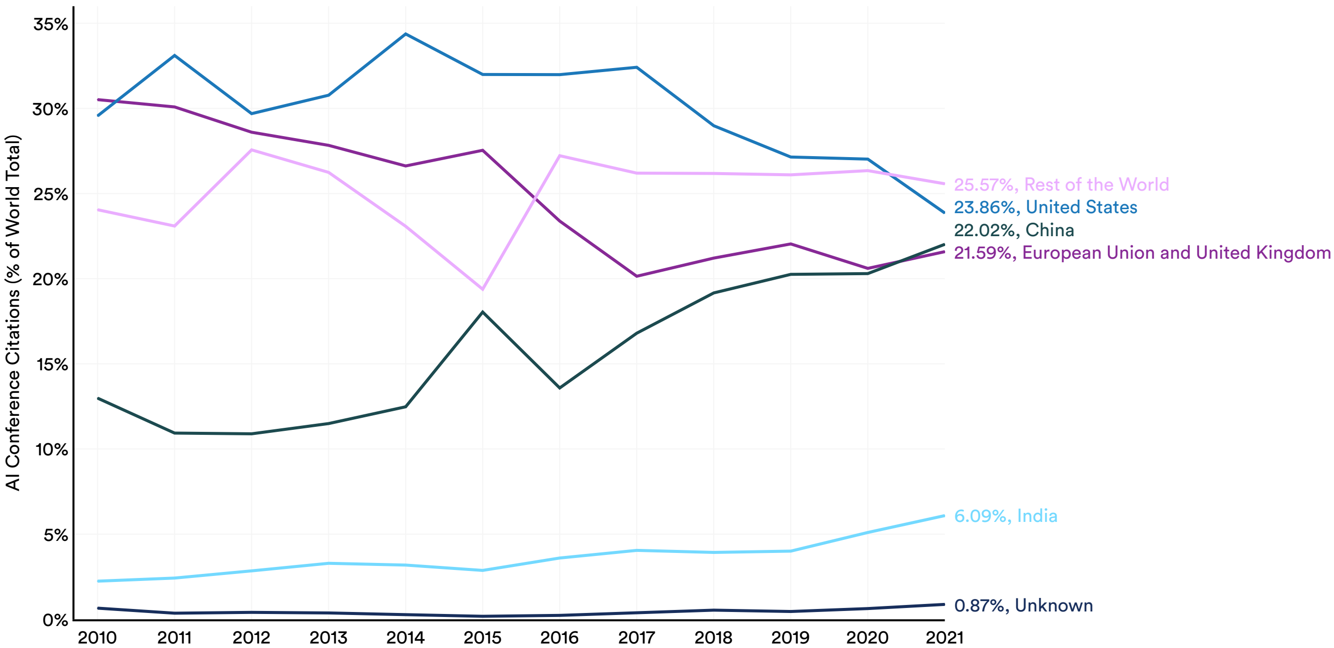

Figure 1.1.16

AI Repositories

Overview

Publishing pre-peer-reviewed papers on repositories of electronic preprints (such as arXiv and SSRN) has become a popular way for AI researchers to disseminate their work outside traditional avenues for publication. These repositories allow researchers to share their findings before submitting them to journals and conferences, thereby accelerating the cycle of information discovery. The number of AI repository publications grew almost 27 times in the past 12 years (Figure 1.1.17).

Number of AI Repository Publications, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

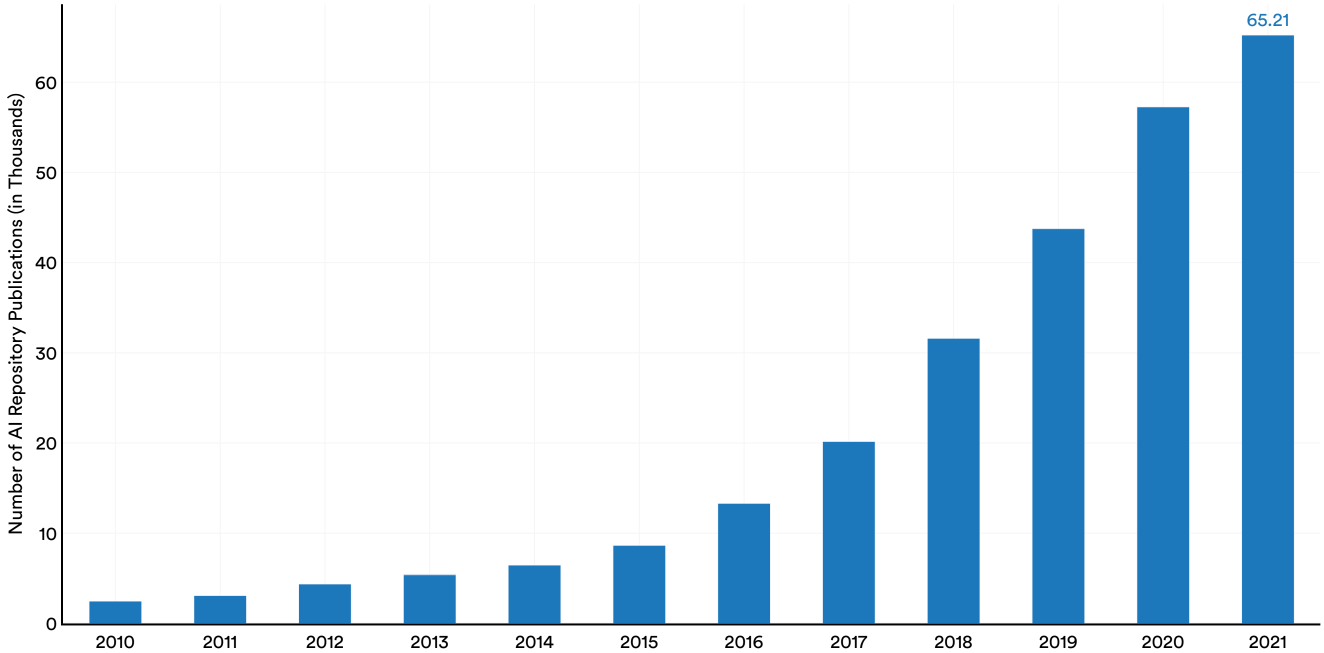

Figure 1.1.17

By Region

Figure 1.1.18 shows that North America has maintained a steady lead in the world share of AI repository publications since 2016. Since 2011, the share of repository publications from Europe and Central Asia has declined. The share represented by East Asia and the Pacific has grown significantly since 2010 and continued growing from 2020 to 2021, a period in which the year-over-year share of North American as well European and Central Asian repository publications declined.

AI Repository Publications (% of World Total) by Region, 2010–21

Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

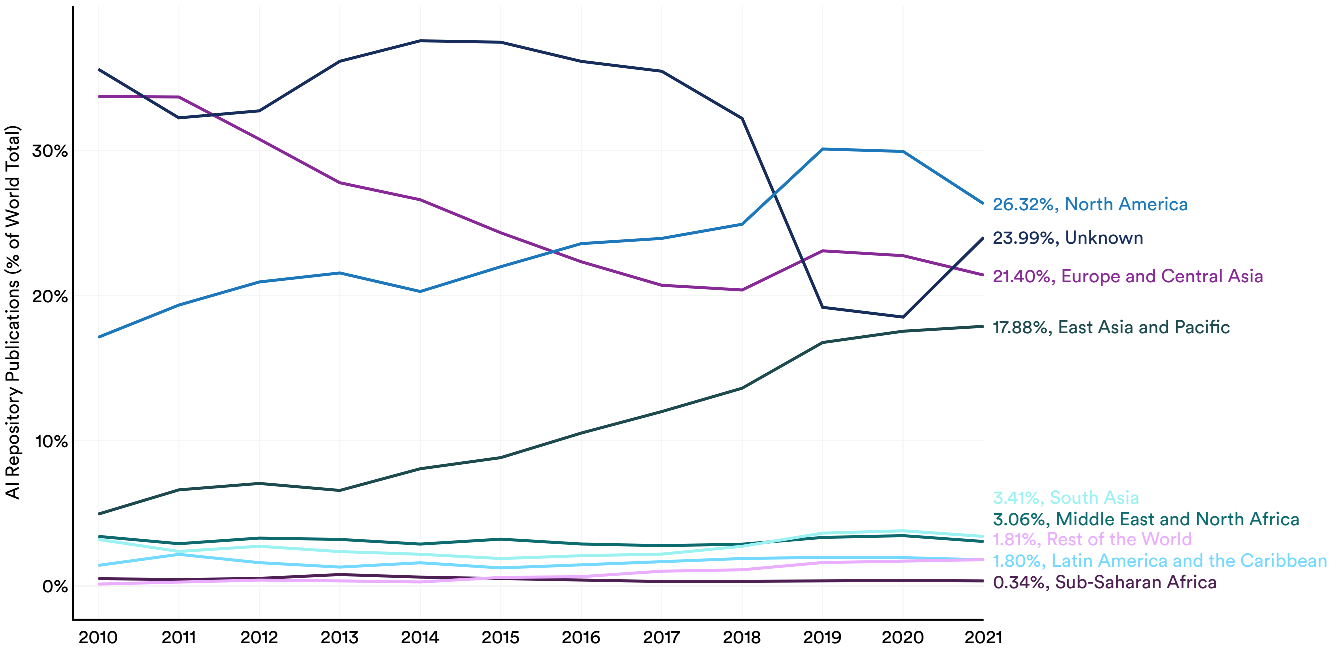

Figure 1.1.18

By Geographic Area

While the United States has held the lead in the percentage of global AI repository publications since 2016, China is catching up, while the European Union plus the United Kingdom's share continues to drop (Figure 1.1.19). In 2021, the United States accounted for 23.5% of the world's AI repository publications, followed by the European Union plus the United Kingdom (20.5%), and then China (11.9%).

AI Repository Publications (% of World Total) by Geographic Area, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

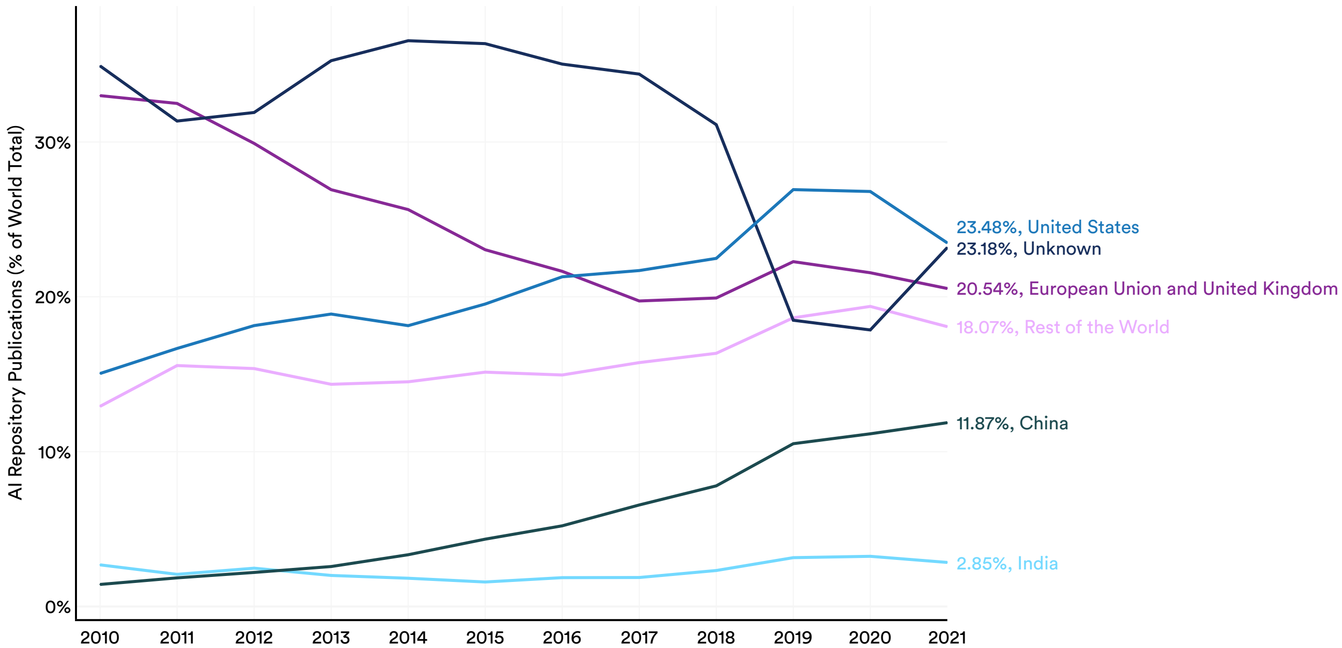

Figure 1.1.19

Citations

In the citations of AI repository publications, Figure 1.1.20 shows that in 2021 the United States topped the list with 29.2% of overall citations, maintaining a dominant lead over the European Union plus the United Kingdom (21.5%), as well as China (21.0%).

AI Repository Citations (% of World Total) by Geographic Area, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

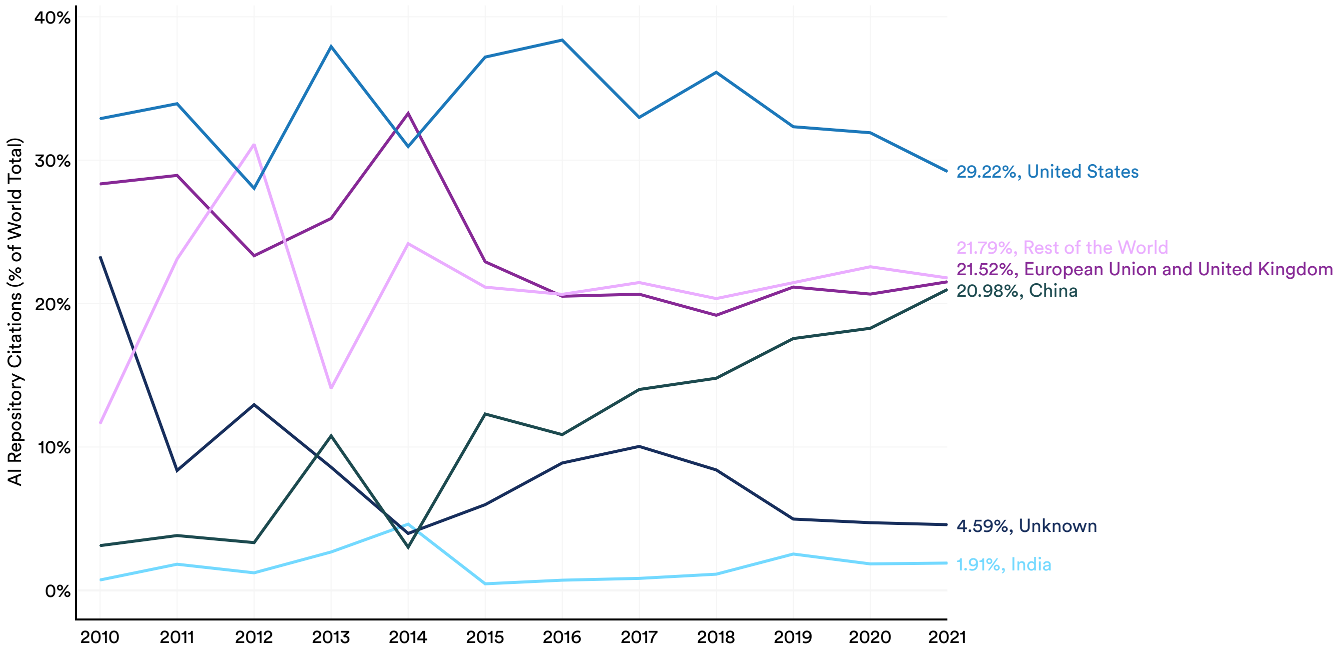

Figure 1.1.20

Narrative Highlight:
Top Publishing Institutions

All Fields

Since 2010, the institution producing the greatest number of total AI papers has been the Chinese Academy of Sciences (Figure 1.1.21). The next top four are all Chinese universities: Tsinghua University, the University of the Chinese Academy of Sciences, Shanghai Jiao Tong University, and Zhejiang University.[5] The total number of publications released by each of these institutions in 2021 is displayed in Figure 1.1.22.

Top Ten Institutions in the World in 2021 Ranked by Number of AI Publications in All Fields, 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

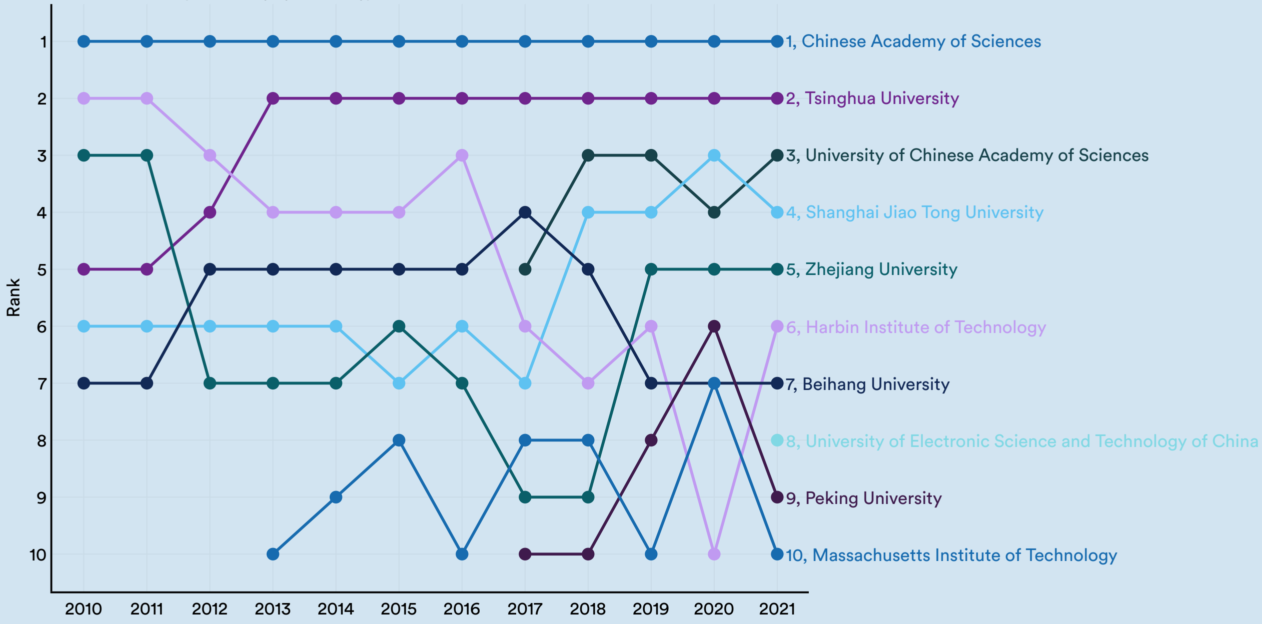

Figure 1.1.21

5 It is important to note that many Chinese research institutions are large, centralized organizations with thousands of researchers. It is therefore not entirely surprising that, purely by the metric of publication count, they outpublish most non-Chinese institutions.

Narrative Highlight:
Top Publishing Institutions (cont'd)

Top Ten Institutions in the World by Number of AI Publications in All Fields, 2021
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

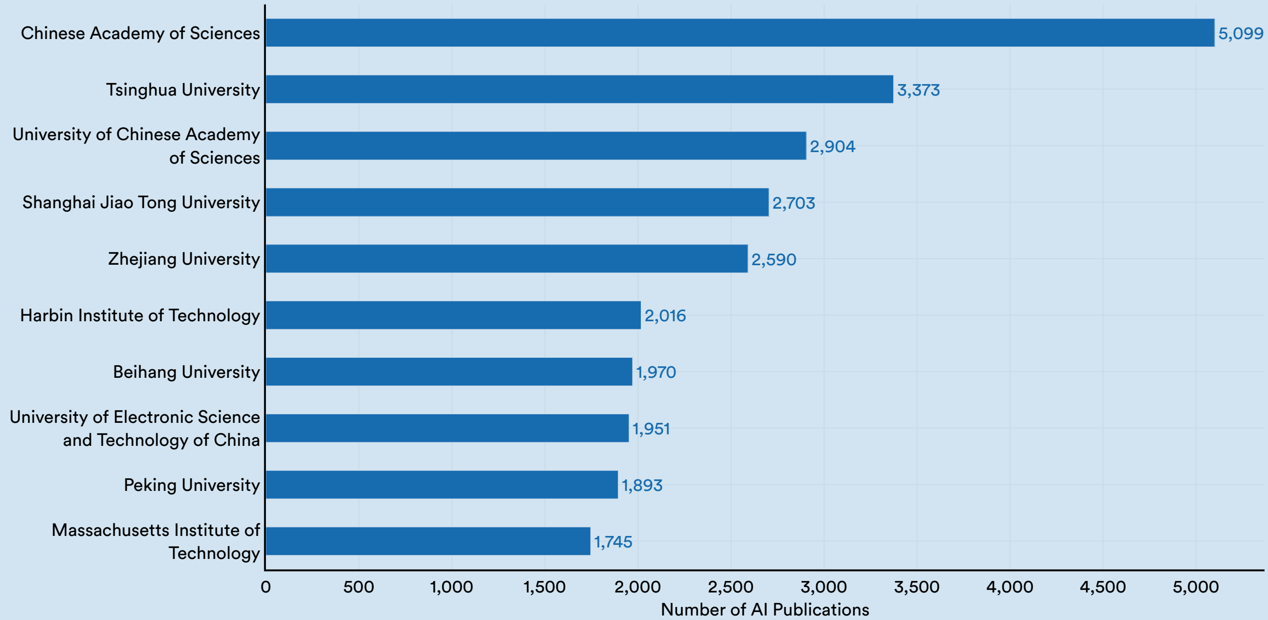

Figure 1.1.22

Narrative Highlight:
Top Publishing Institutions (cont'd)

Computer Vision

In 2021, the top 10 institutions publishing the greatest number of AI computer vision publications were all Chinese (Figure 1.1.23). The Chinese Academy of Sciences published the largest number of such publications, with a total of 562.

Top Ten Institutions in the World by Number of AI Publications in Computer Vision, 2021
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

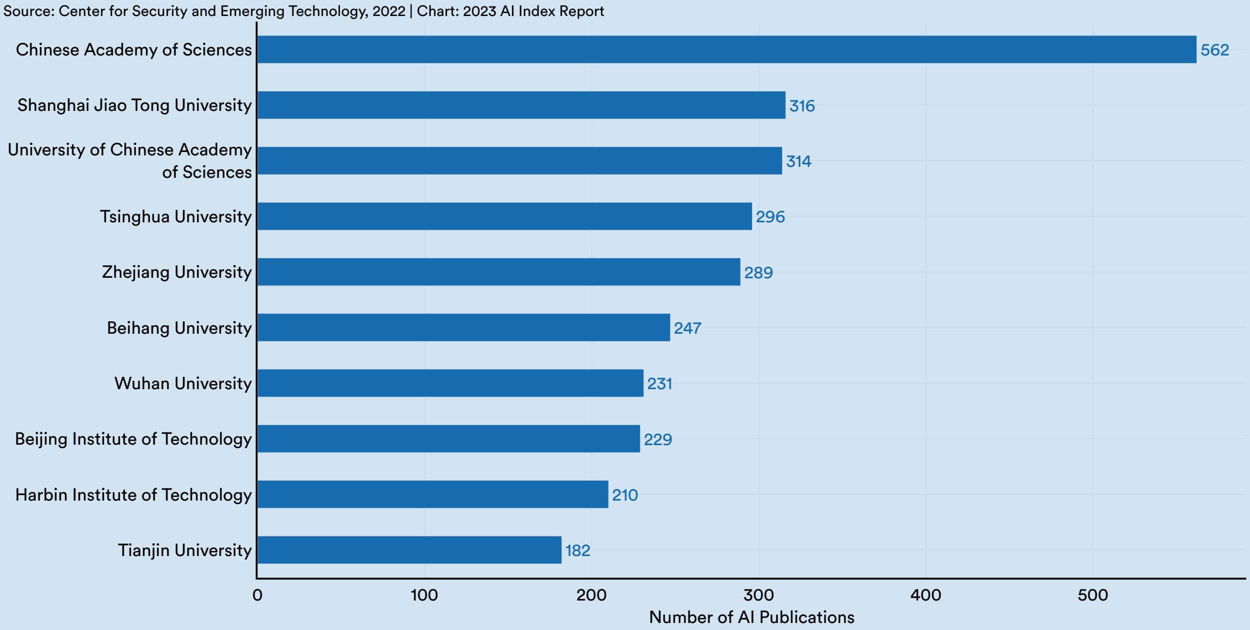

Figure 1.1.23

Narrative Highlight:
Top Publishing Institutions (cont'd)

Natural Language Processing

American institutions are represented to a greater degree in the share of top NLP publishers (Figure 1.1.24). Although the Chinese Academy of Sciences was again the world's leading institution in 2021 (182 publications), Carnegie Mellon took second place (140 publications), followed by Microsoft (134). In addition, 2021 was the first year Amazon and Alibaba were represented among the top-ten largest publishing NLP institutions.

Top Ten Institutions in the World by Number of AI Publications in Natural Language Processing, 2021
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

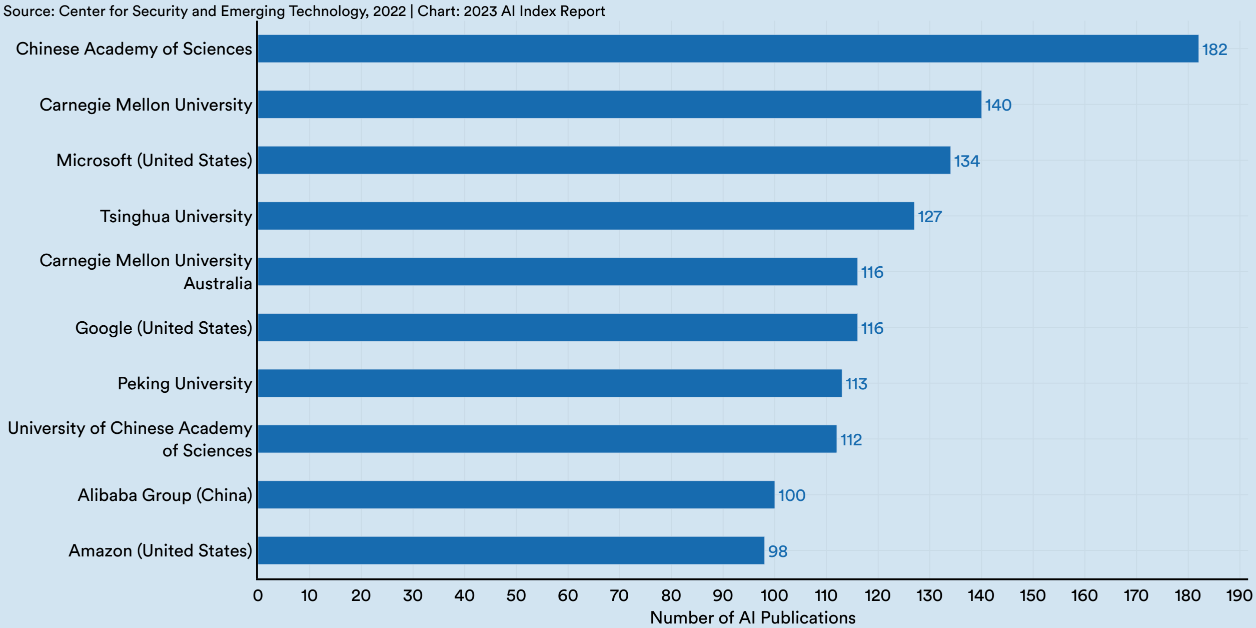

Figure 1.1.24

Narrative Highlight:
Top Publishing Institutions (cont'd)

Speech Recognition

In 2021, the greatest number of speech recognition papers came from the Chinese Academy of Sciences (107), followed by Microsoft (98) and Google (75) (Figure 1.1.25). The Chinese Academy of Sciences reclaimed the top spot in 2021 from Microsoft, which held first position in 2020.

Top Ten Institutions in the World by Number of AI Publications in Speech Recognition, 2021
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

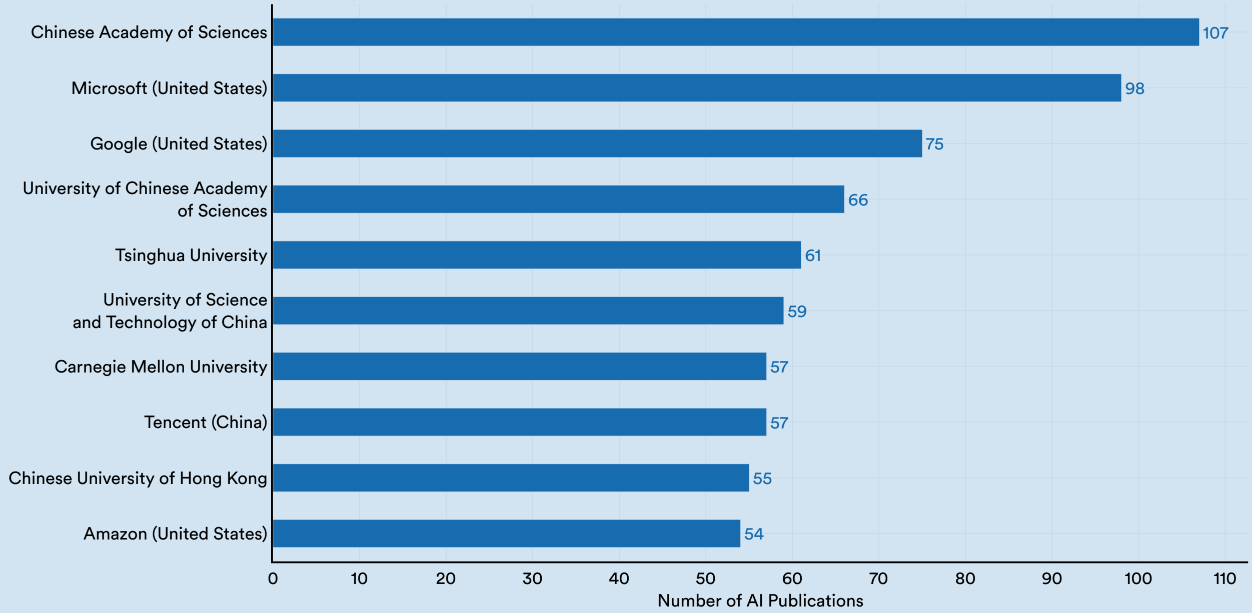

Figure 1.1.25

Epoch AI is a collective of researchers investigating and forecasting the development of advanced AI. Epoch curates a database of significant AI and machine learning systems that have been released since the 1950s. There are different criteria under which the Epoch team decides to include particular AI systems in their database; for example, the system may have registered a state-of-the-art improvement, been deemed to have been historically significant, or been highly cited.

This subsection uses the Epoch database to track trends in significant AI and machine learning systems. The latter half of the chapter includes research done by the AI Index team that reports trends in large language and multimodal models, which are models trained on large amounts of data and adaptable to a variety of downstream applications.

1.2 Trends in Significant Machine Learning Systems

General Machine Learning Systems

The figures below report trends among all machine learning systems included in the Epoch dataset. For reference, these systems are referred to as *significant machine learning systems* throughout the subsection.

System Types

Among the significant AI machine learning systems released in 2022, the most common class of system was language (Figure 1.2.1). There were 23 significant AI language systems released in 2022, roughly six times the number of the next most common system type, multimodal systems.

Number of Significant Machine Learning Systems by Domain, 2022
Source: Epoch, 2022 | Chart: 2023 AI Index Report

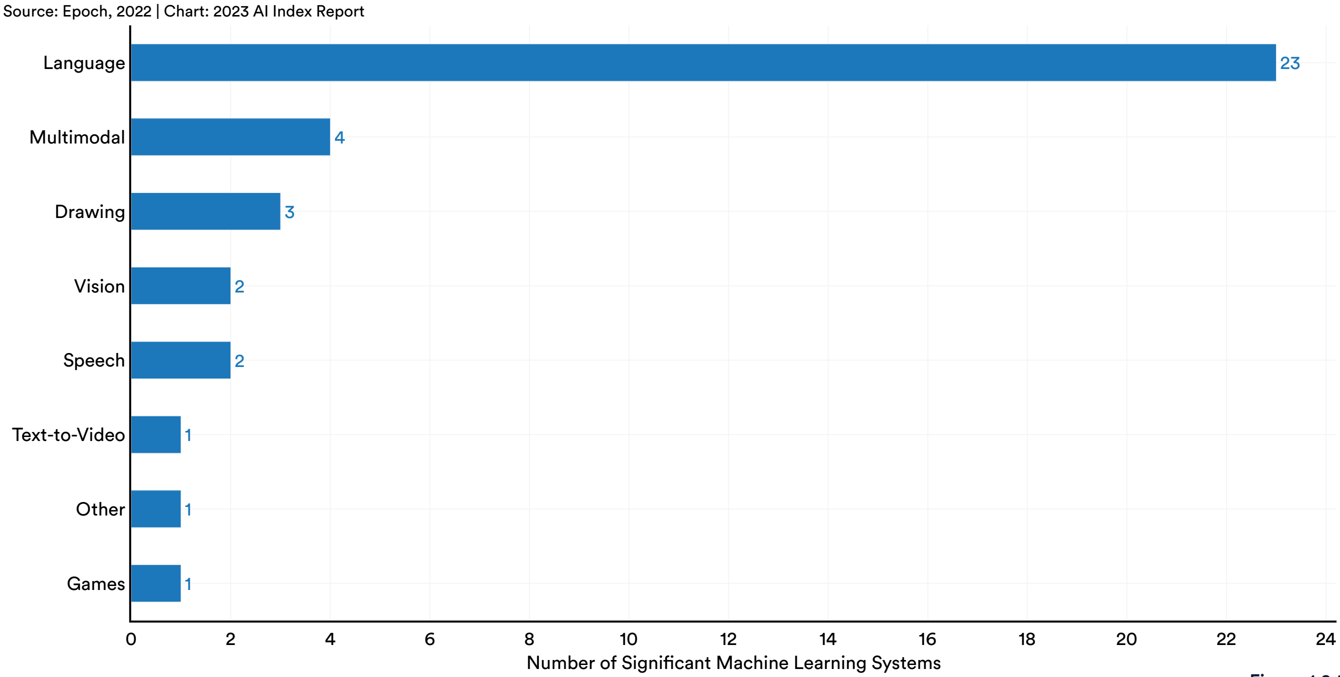

Figure 1.2.1[6]

6 There were 38 total significant AI machine learning systems released in 2022, according to Epoch; however, one of the systems, BaGuaLu, did not have a domain classification and is therefore omitted from Figure 1.2.1.

Sector Analysis

Which sector among industry, academia, or nonprofit has released the greatest number of significant machine learning systems? Until 2014, most machine learning systems were released by academia. Since then, industry has taken over (Figure 1.2.2). In 2022, there were 32 significant industry-produced machine learning systems compared to just three produced by academia. Producing state-of-the-art AI systems increasingly requires large amounts of data, computing power, and money; resources that industry actors possess in greater amounts compared to nonprofits and academia.

Number of Significant Machine Learning Systems by Sector, 2002–22
Source: Epoch, 2022 | Chart: 2023 AI Index Report

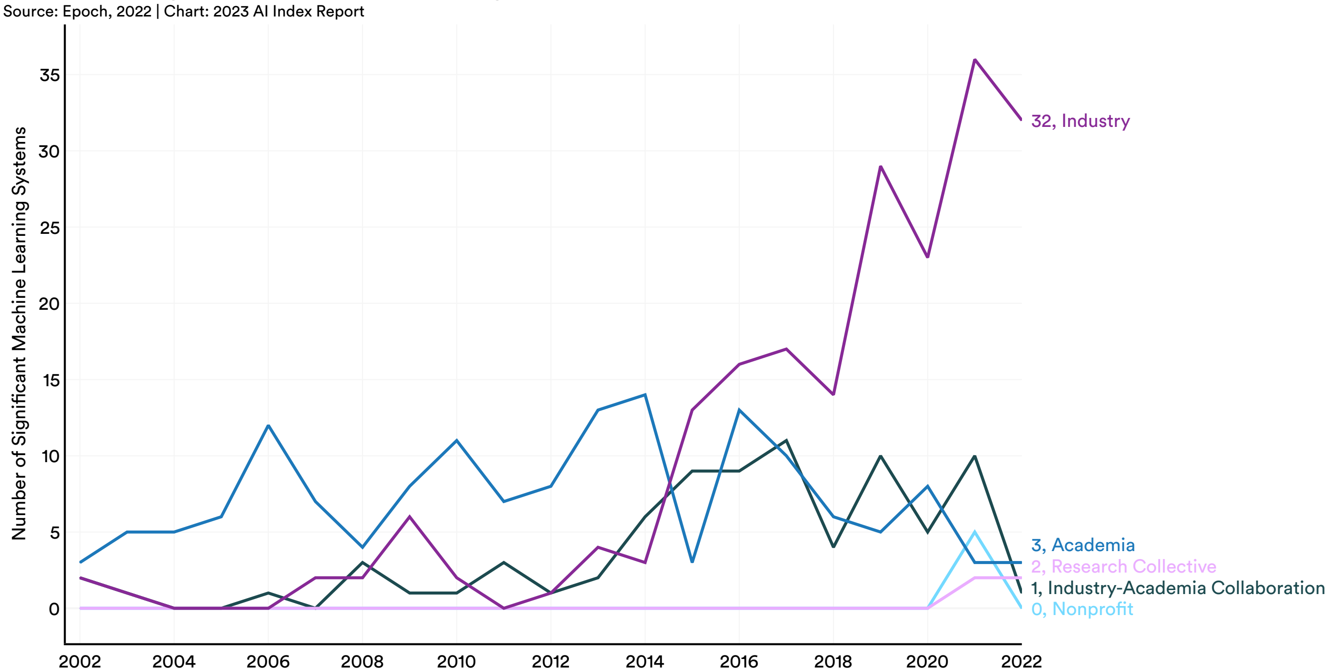

Figure 1.2.2

National Affiliation

In order to paint a picture of AI's evolving geopolitical landscape, the AI Index research team identified the nationality of the authors who contributed to the development of each significant machine learning system in the Epoch dataset.[7]

Systems

Figure 1.2.3 showcases the total number of significant machine learning systems attributed to researchers from particular countries.[8] A researcher is considered to have belonged to the country in which their institution, for example a university or AI-research firm, was headquartered. In 2022, the United States produced the greatest number of significant machine learning systems with 16, followed by the United Kingdom (8) and China (3). Moreover, since 2002 the United States has outpaced the United Kingdom and the European Union, as well as China, in terms of the total number of significant machine learning systems produced (Figure 1.2.4). Figure 1.2.5 displays the total number of significant machine learning systems produced by country since 2002 for the entire world.

Number of Significant Machine Learning Systems by Country, 2022
Source: Epoch and AI Index, 2022 | Chart: 2023 AI Index Report

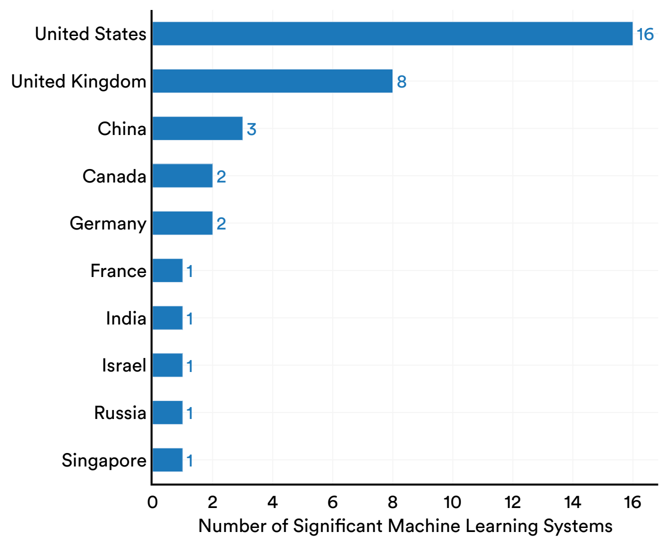

Figure 1.2.3

Number of Significant Machine Learning Systems by Select Geographic Area, 2002–22
Source: Epoch and AI Index, 2022 | Chart: 2023 AI Index Report

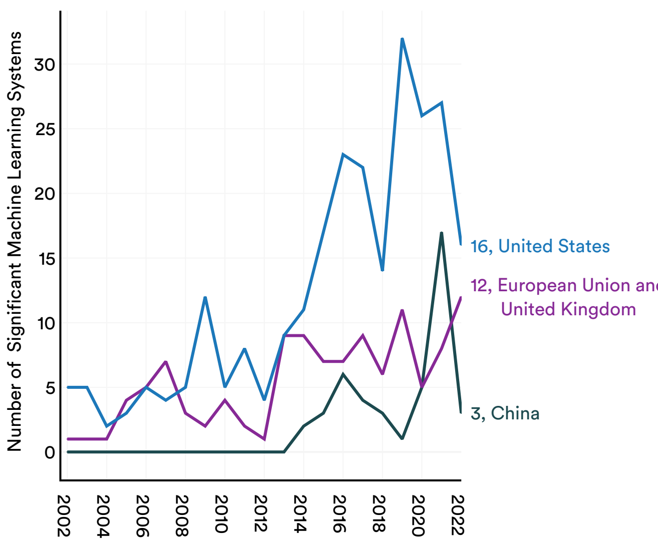

Figure 1.2.4

7 The methodology by which the AI Index identified authors' nationality is outlined in greater detail in the Appendix.
8 A machine learning system is considered to be affiliated with a particular country if at least one author involved in creating the model was affiliated with that country. Consequently, in cases where a system has authors from multiple countries, double counting may occur.

Number of Significant Machine Learning Systems by Country, 2002–22 (Sum)
Source: AI Index, 2022 | Chart: 2023 AI Index Report

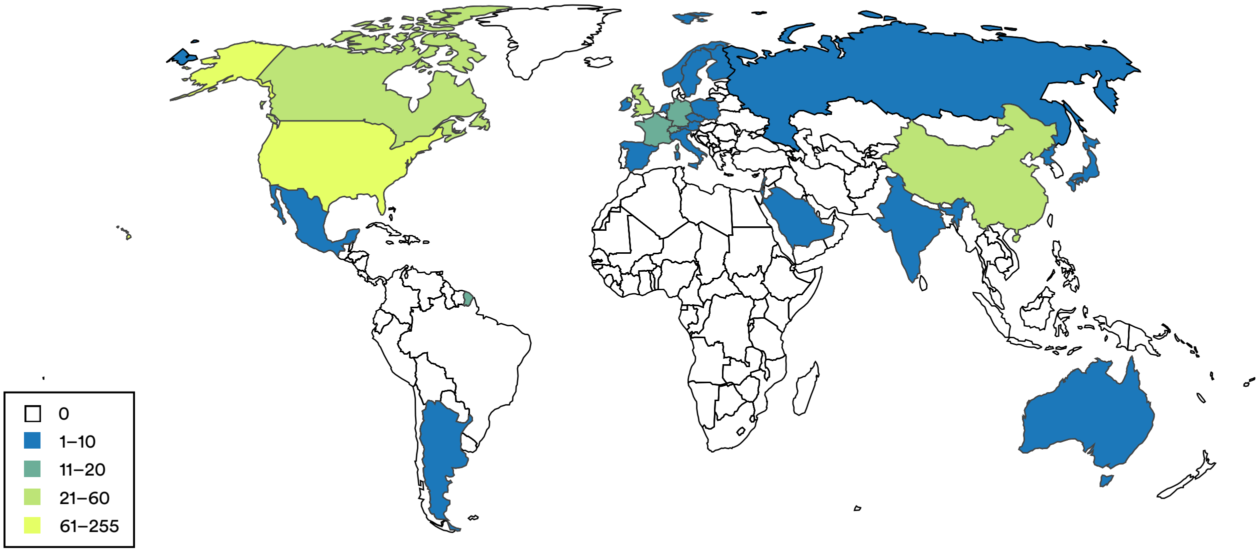

Figure 1.2.5

Authorship

Figures 1.2.6 to 1.2.8 look at the total number of authors, disaggregated by national affiliation, that contributed to the launch of significant machine learning systems. As was the case with total systems, in 2022 the United States had the greatest number of authors producing significant machine learning systems, with 285, more than double that of the United Kingdom and nearly six times that of China (Figure 1.2.6).

Number of Authors of Significant Machine Learning Systems by Country, 2022
Source: Epoch and AI Index, 2022 | Chart: 2023 AI Index Report

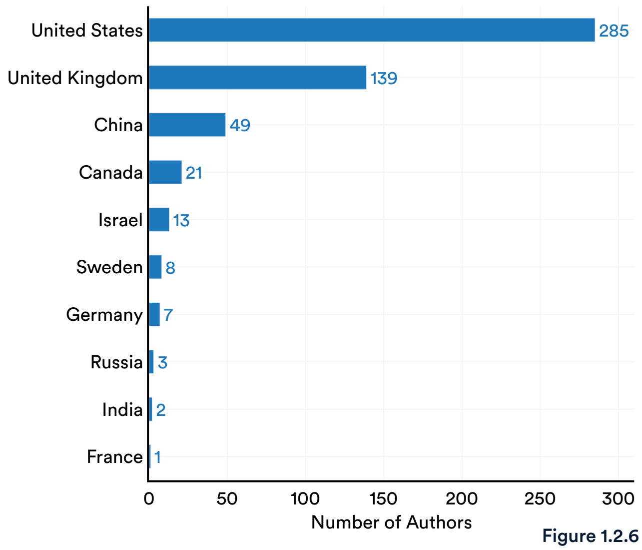

Figure 1.2.6

Number of Authors of Significant Machine Learning Systems by Select Geographic Area, 2002–22
Source: Epoch and AI Index, 2022 | Chart: 2023 AI Index Report

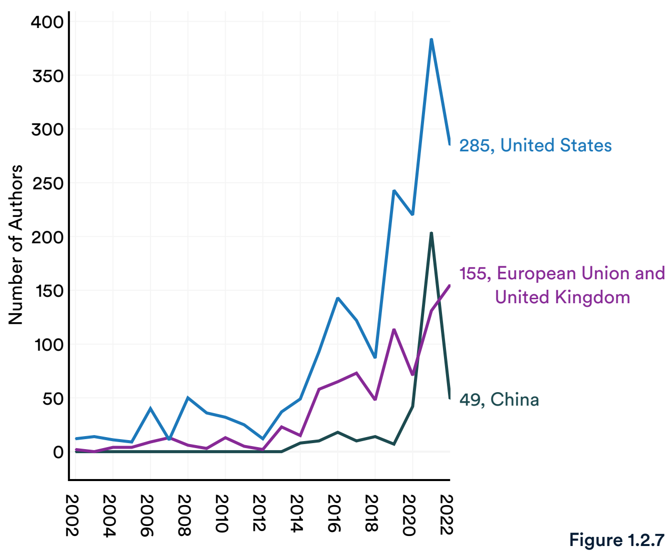

Figure 1.2.7

Number of Authors of Significant Machine Learning Systems by Country, 2002–22 (Sum)
Source: AI Index, 2022 | Chart: 2023 AI Index Report

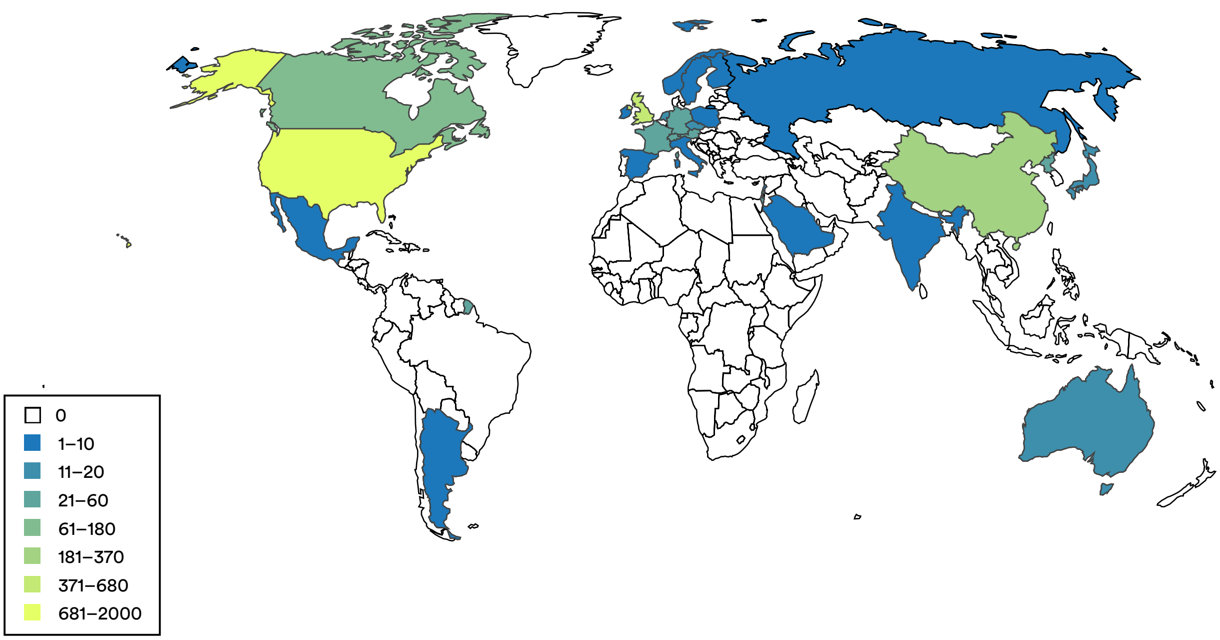

Figure 1.2.8

Parameter Trends

Parameters are numerical values that are learned by machine learning models during training. The value of parameters in machine learning models determines how a model might interpret input data and make predictions. Adjusting parameters is an essential step in ensuring that the performance of a machine learning system is optimized.

Figure 1.2.9 highlights the number of parameters of the machine learning systems included in the Epoch dataset by sector. Over time, there has been a steady increase in the number of parameters, an increase that has become particularly sharp since the early 2010s. The fact that AI systems are rapidly increasing their parameters is reflective of the increased complexity of the tasks they are being asked to perform, the greater availability of data, advancements in underlying hardware, and most importantly, the demonstrated performance of larger models.

Number of Parameters of Significant Machine Learning Systems by Sector, 1950–2022
Source: Epoch, 2022 | Chart: 2023 AI Index Report

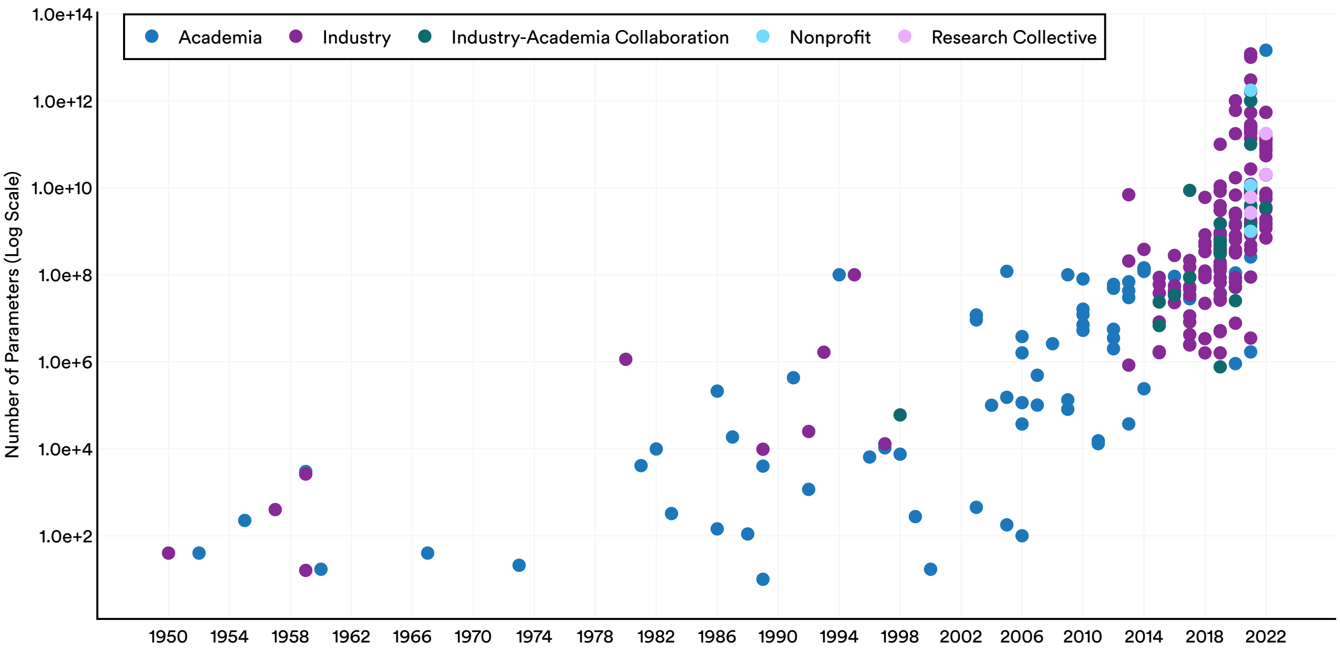

Figure 1.2.9

Figure 1.2.10 demonstrates the parameters of machine learning systems by domain. In recent years, there has been a rise in parameter-rich systems.

Number of Parameters of Significant Machine Learning Systems by Domain, 1950–2022
Source: Epoch, 2022 | Chart: 2023 AI Index Report

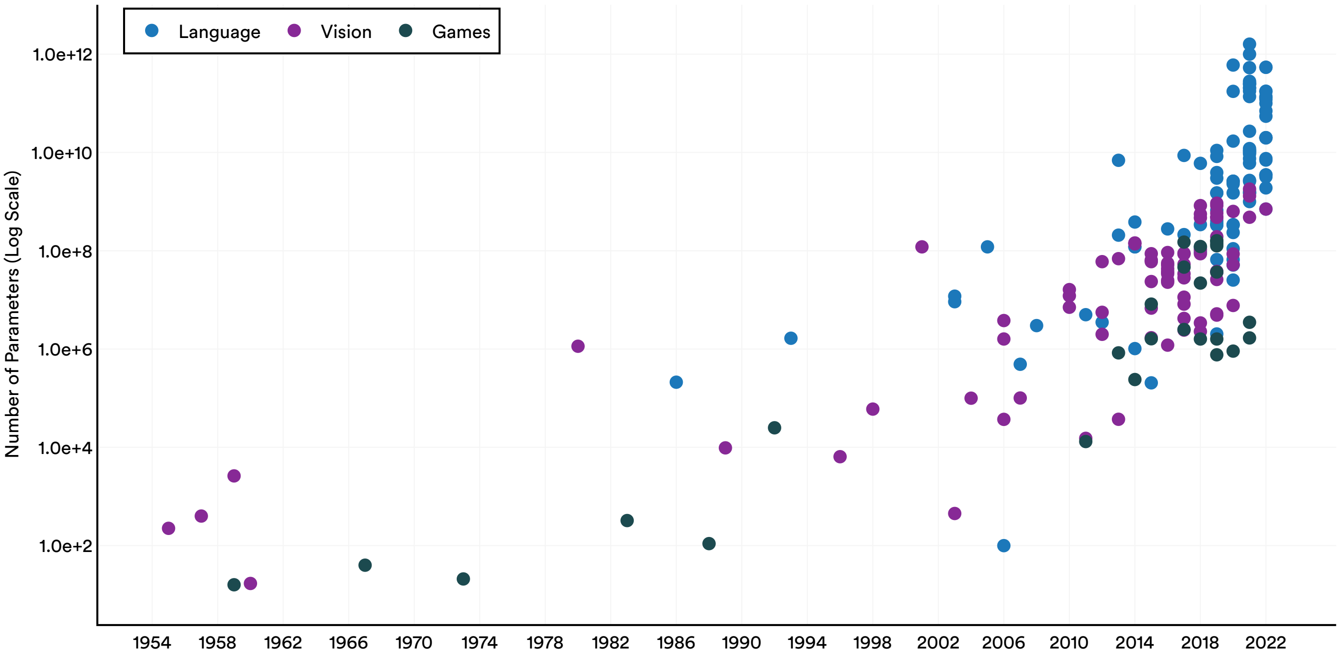

Figure 1.2.10

Compute Trends

The computational power, or "compute," of AI systems refers to the amount of computational resources needed to train and run a machine learning system. Typically, the more complex a system is, and the larger the dataset on which it is trained, the greater the amount of compute required.

The amount of compute used by significant AI machine learning systems has increased exponentially in the last half-decade (Figure 1.2.11).[9] The growing demand for compute in AI carries several important implications. For example, more compute-intensive models tend to have greater environmental impacts, and industrial players tend to have easier access to computational resources than others, such as universities.

Training Compute (FLOP) of Significant Machine Learning Systems by Sector, 1950–2022
Source: Epoch, 2022 | Chart: 2023 AI Index Report

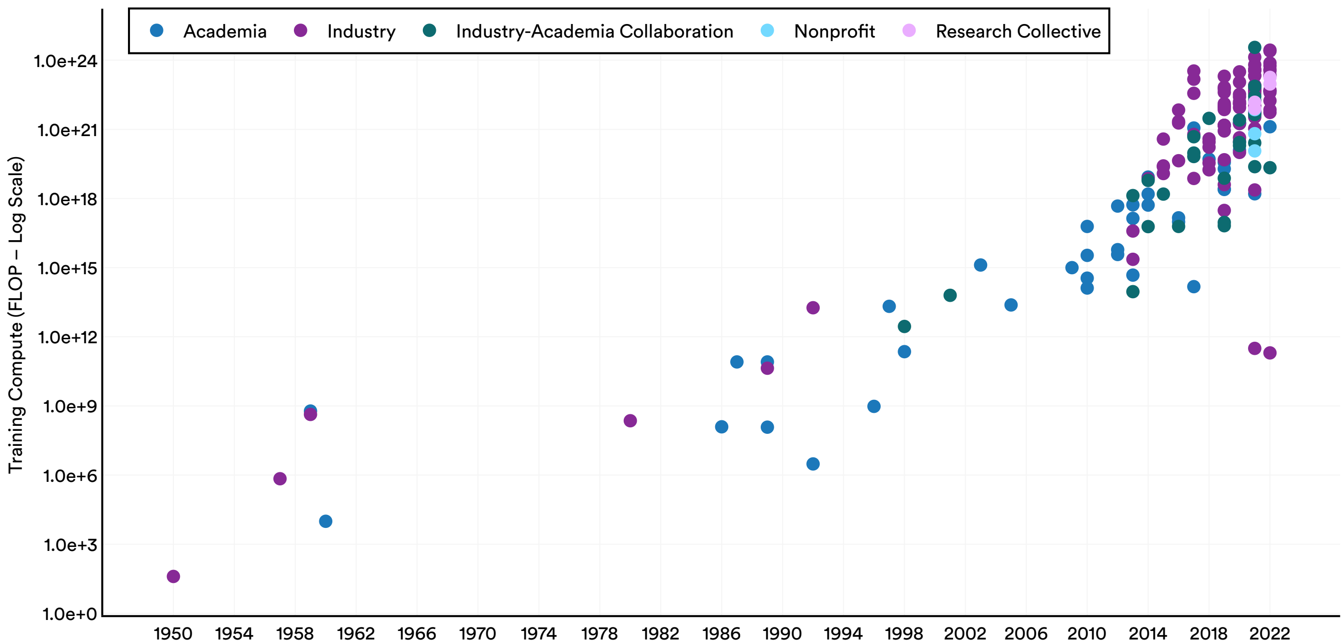

Figure 1.2.11

9 FLOP stands for "Floating Point Operations" and is a measure of the performance of a computational device.

Since 2010, it has increasingly been the case that of all machine learning systems, language models are demanding the most computational resources.

Training Compute (FLOP) of Significant Machine Learning Systems by Domain, 1950–2022
Source: Epoch, 2022 | Chart: 2023 AI Index Report

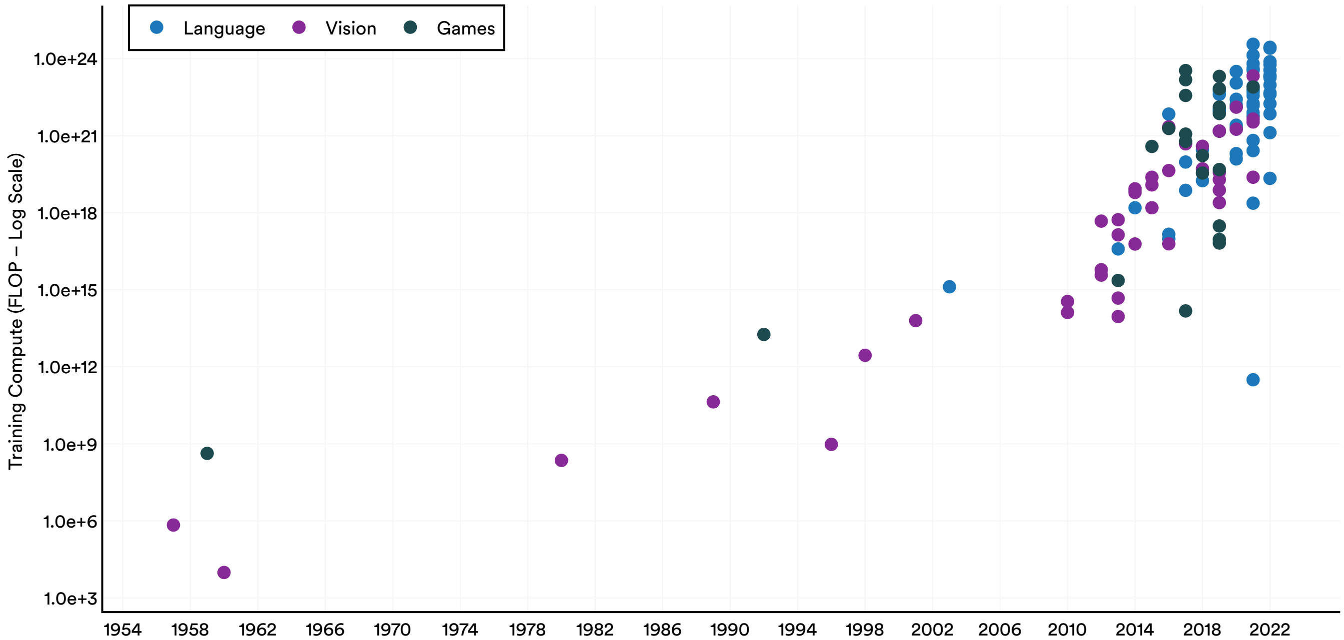

Figure 1.2.12

Large Language and Multimodal Models

Large language and multimodal models, sometimes called foundation models, are an emerging and increasingly popular type of AI model that is trained on huge amounts of data and adaptable to a variety of downstream applications. Large language and multimodal models like ChatGPT, DALL-E 2, and Make-A-Video have demonstrated impressive capabilities and are starting to be widely deployed in the real world.

National Affiliation

This year the AI Index conducted an analysis of the national affiliation of the authors responsible for releasing new large language and multimodal models.[10] The majority of these researchers were from American institutions (54.2%) (Figure 1.2.13). In 2022, for the first time, researchers from Canada, Germany, and India contributed to the development of large language and multimodal models.

Authors of Select Large Language and Multimodal Models (% of Total) by Country, 2019–22
Source: Epoch and AI Index, 2022 | Chart: 2023 AI Index Report

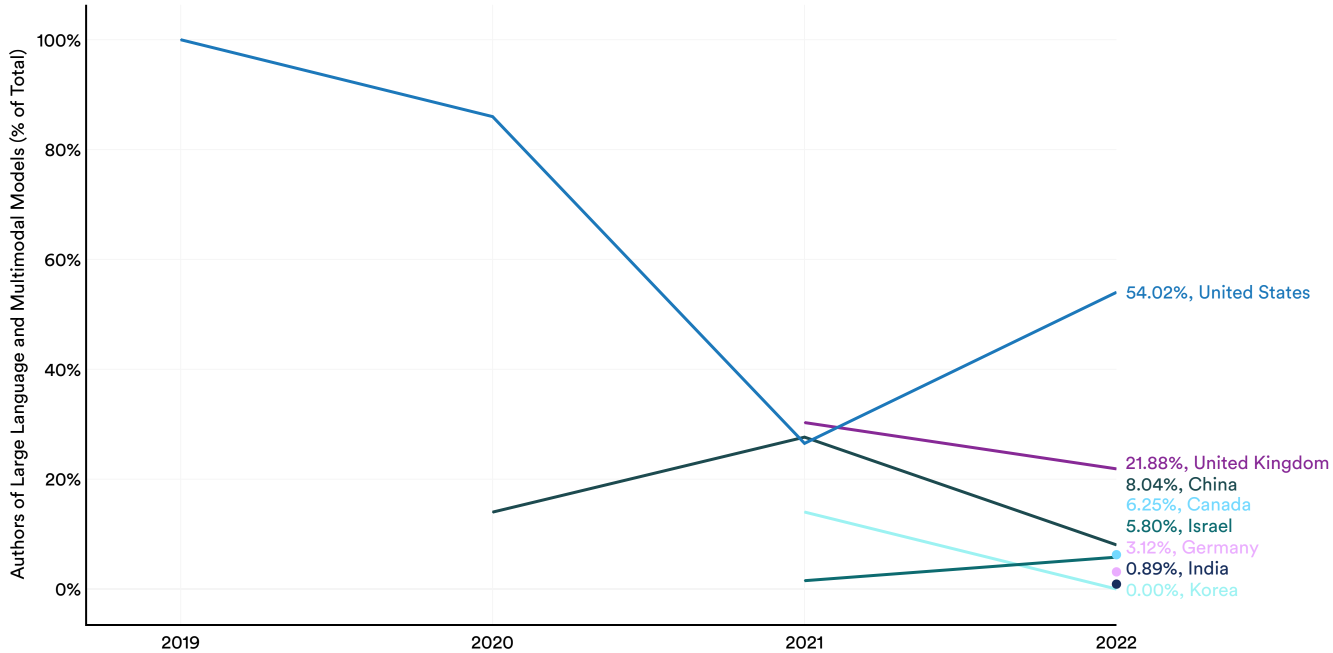

Figure 1.2.13

Figure 1.2.14 offers a timeline view of the large language and multimodal models that have been released since GPT-2, along with the national affiliations of the researchers who produced the models. Some of the notable American large language and multimodal models released in 2022 included OpenAI's DALL-E 2 and Google's PaLM (540B). The only Chinese large language and multimodal model released in 2022 was GLM-130B, an impressive bilingual (English and Chinese) model created by researchers at Tsinghua University. BLOOM, also launched in late 2022, was listed as indeterminate given that it was the result of a collaboration of more than 1,000 international researchers.

10 The AI models that were considered to be large language and multimodal models were hand-selected by the AI Index steering committee. It is possible that this selection may have omitted certain models.

Timeline and National Affiliation of Select Large Language and Multimodal Model Releases

Source: AI Index, 2022 | Chart: 2023 AI Index Report

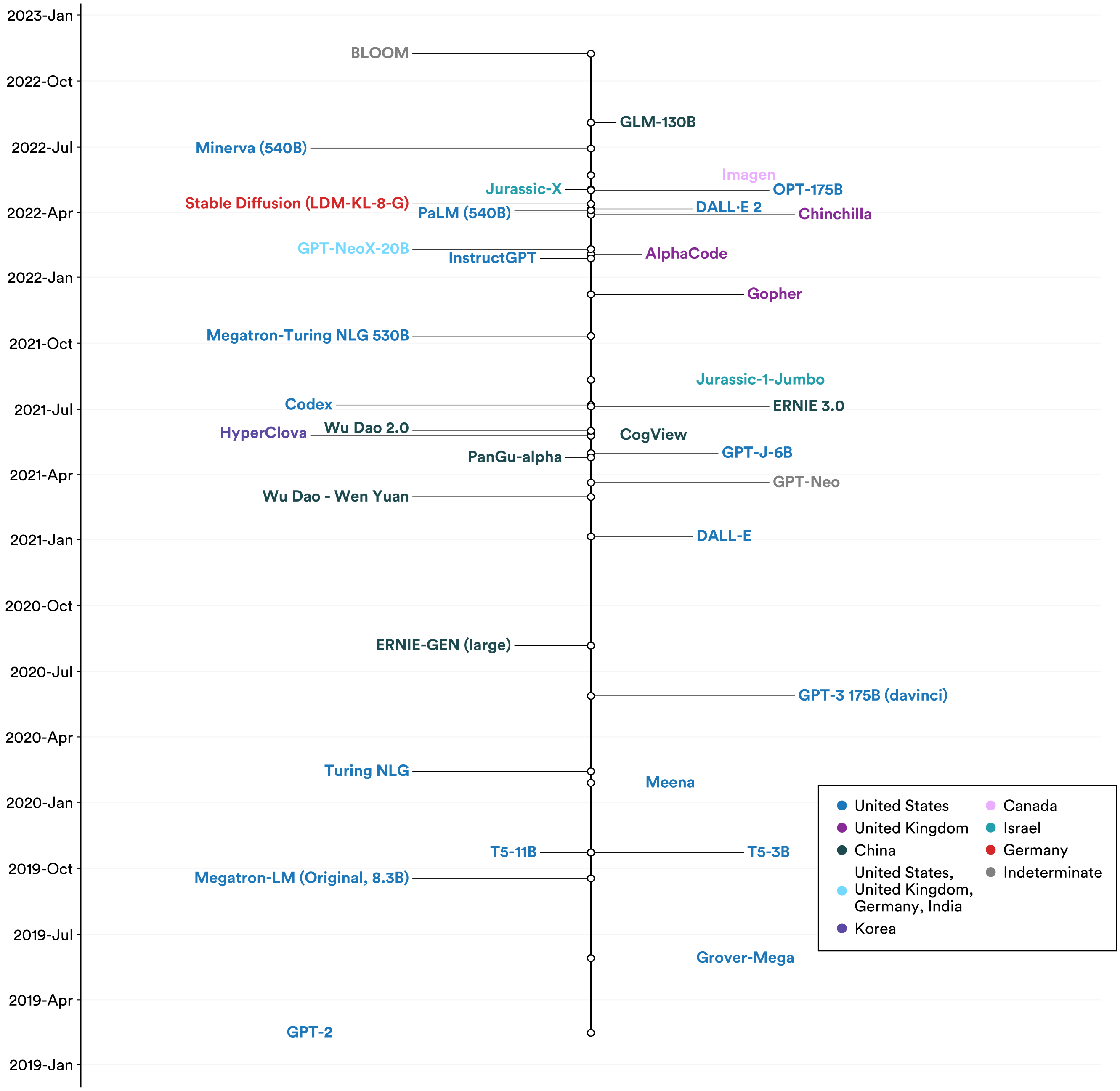

Figure 1.2.14[11]

11 While we were conducting the analysis to produce Figure 1.2.14, Irene Solaiman published a paper that has a similar analysis. We were not aware of the paper at the time of our research.

Parameter Count

Over time, the number of parameters of newly released large language and multimodal models has massively increased. For example, GPT-2, which was the first large language and multimodal model released in 2019, only had 1.5 billion parameters. PaLM, launched by Google in 2022, had 540 billion, nearly 360 times more than GPT-2. The median number of parameters in large language and multimodal models is increasing exponentially over time (Figure 1.2.15).

Number of Parameters of Select Large Language and Multimodal Models, 2019–22

Source: Epoch, 2022 | Chart: 2023 AI Index Report

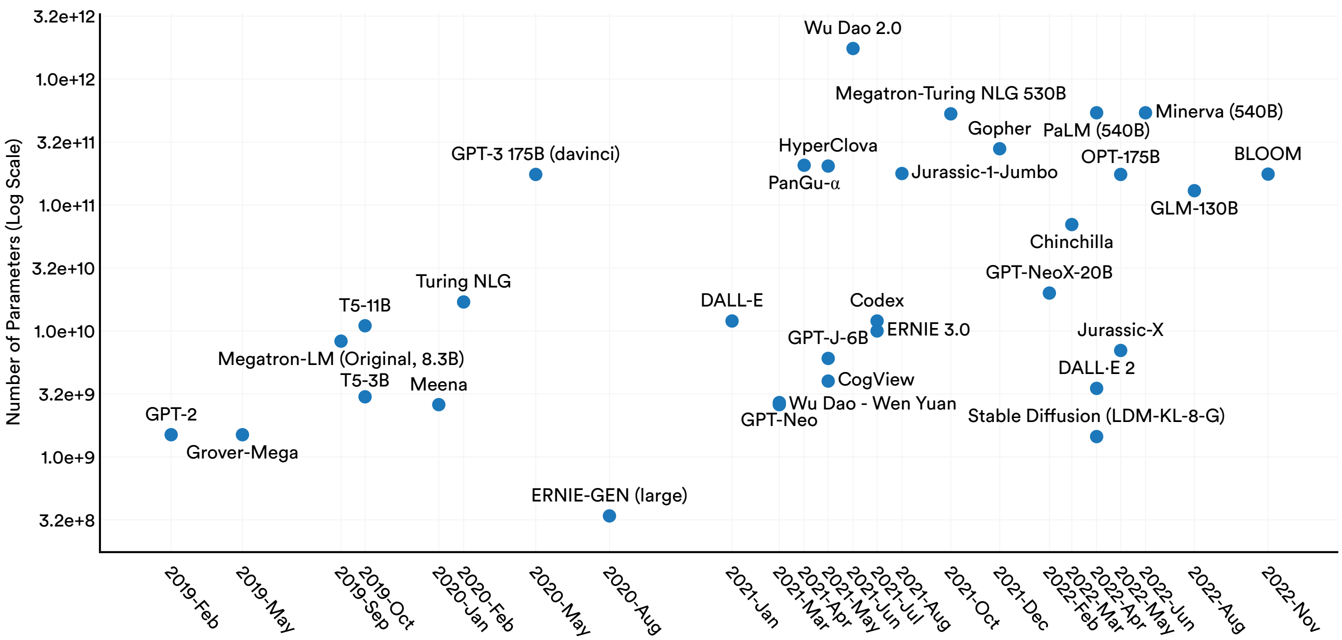

Figure 1.2.15

Training Compute

The training compute of large language and multimodal models has also steadily increased (Figure 1.2.16). The compute used to train Minerva (540B), a large language and multimodal model released by Google in June 2022 that displayed impressive abilities on quantitative reasoning problems, was roughly nine times greater than that used for OpenAI's GPT-3, which was released in June 2022, and roughly 1839 times greater than that used for GPT-2 (released February 2019).

Training Compute (FLOP) of Select Large Language and Multimodal Models, 2019–22

Source: Epoch, 2022 | Chart: 2023 AI Index Report

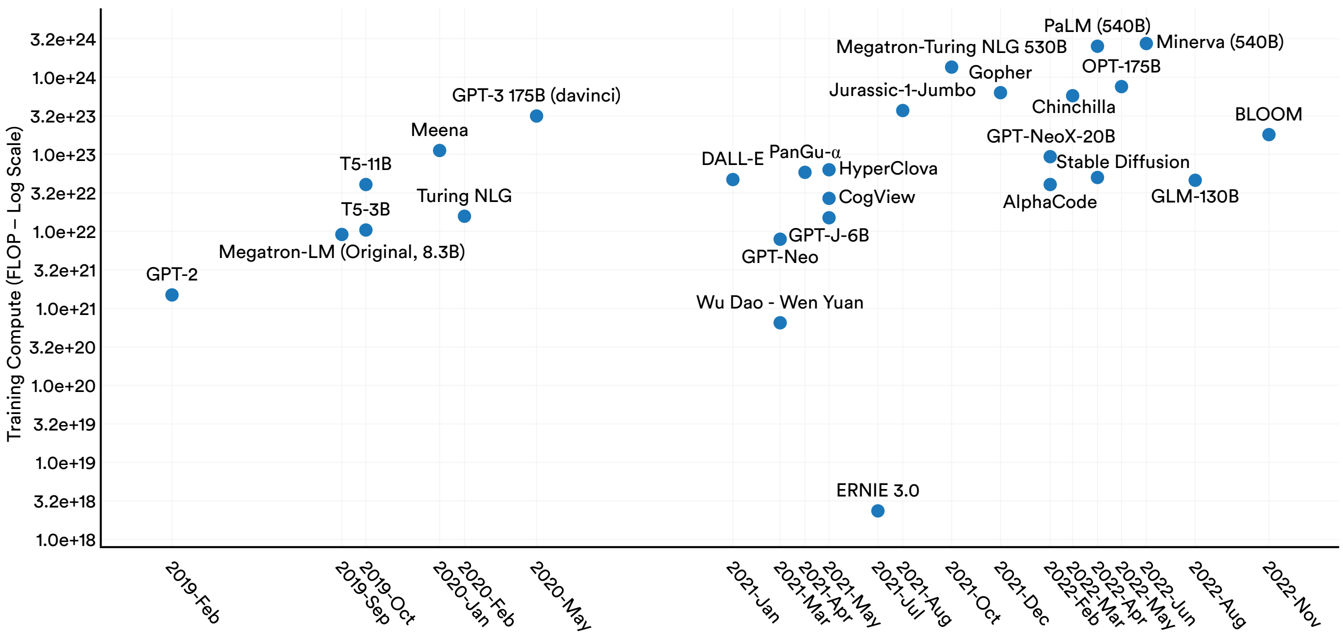

Figure 1.2.16

Training Cost

A particular theme of the discourse around large language and multimodal models has to do with their hypothesized costs. Although AI companies rarely speak openly about training costs, it is widely speculated that these models cost millions of dollars to train and will become increasingly expensive with scale.

This subsection presents novel analysis in which the AI Index research team generated estimates for the training costs of various large language and multimodal models (Figure 1.2.17). These estimates are based on the hardware and training time disclosed by the models' authors. In cases where training time was not disclosed, we calculated from hardware speed, training compute, and hardware utilization efficiency. Given the possible variability of the estimates, we have qualified each estimate with the tag of mid, high, or low: mid where the estimate is thought to be a mid-level estimate, high where it is thought to be an overestimate, and low where it is thought to be an underestimate. In certain cases, there was not enough data to estimate the training cost of particular large language and multimodal models, therefore these models were omitted from our analysis.

The AI Index estimates validate popular claims that large language and multimodal models are increasingly costing millions of dollars to train. For example, Chinchilla, a large language model launched by DeepMind in May 2022, is estimated to have cost $2.1 million, while BLOOM's training is thought to have cost $2.3 million.

Estimated Training Cost of Select Large Language and Multimodal Models
Source: AI Index, 2022 | Chart: 2023 AI Index Report

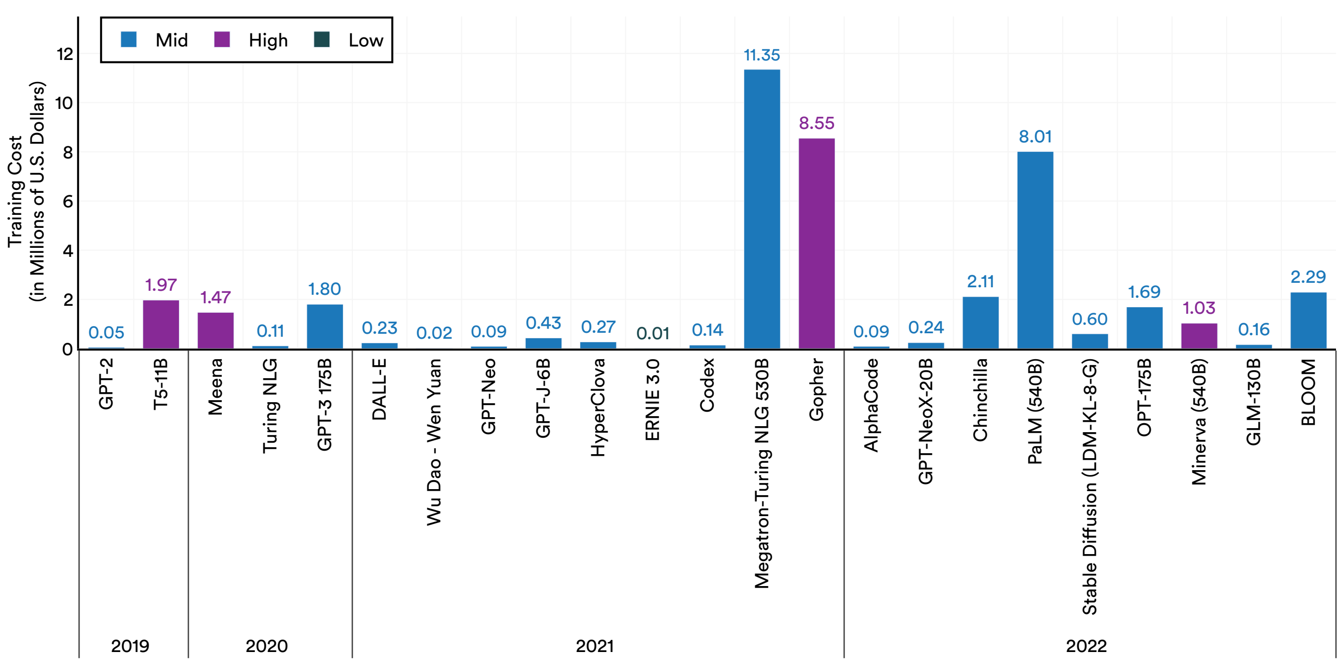

Figure 1.2.17

12 See Appendix for the complete methodology behind the cost estimates.

There is also a clear relationship between the cost of large language and multimodal models and their size. As evidenced in Figures 1.2.18 and 1.2.19, the large language and multimodal models with a greater number of parameters and that train using larger amounts of compute tend to be more expensive.

Estimated Training Cost of Select Large Language and Multimodal Models and Number of Parameters

Source: AI Index, 2022 | Chart: 2023 AI Index Report

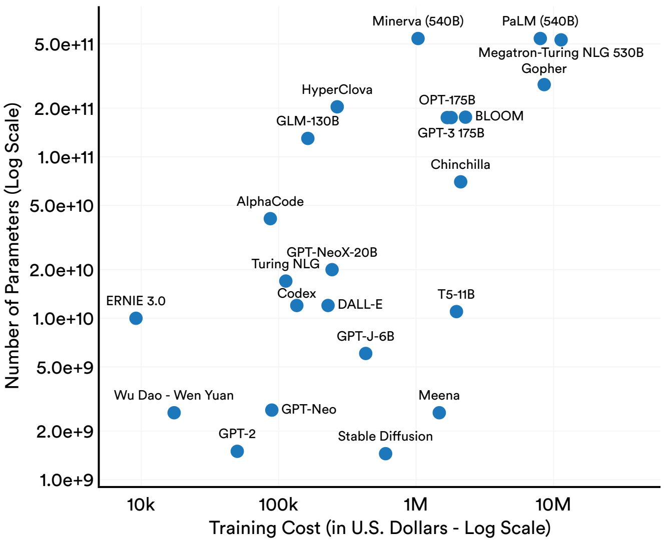

Figure 1.2.18

Estimated Training Cost of Select Large Language and Multimodal Models and Training Compute (FLOP)

Source: AI Index, 2022 | Chart: 2023 AI Index Report

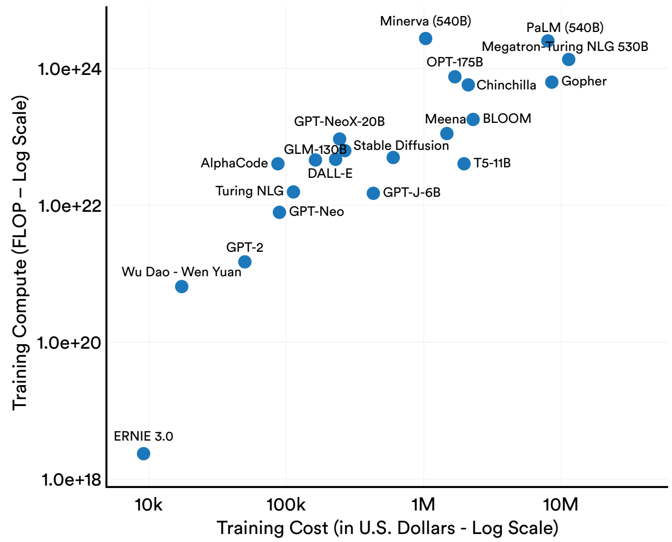

Figure 1.2.19

AI conferences are key venues for researchers to share their work and connect with peers and collaborators. Conference attendance is an indication of broader industrial and academic interest in a scientific field. In the past 20 years, AI conferences have grown in size, number, and prestige. This section presents data on the trends in attendance at major AI conferences.

1.3 AI Conferences

Conference Attendance

After a period of increasing attendance, the total attendance at the conferences for which the AI Index collected data dipped in 2021 and again in 2022 (Figure 1.3.1).[13] This decline may be attributed to the fact that many conferences returned to hybrid or in-person formats after being fully virtual in 2020 and 2021. For example, the International Joint Conference on Artificial Intelligence (IJCAI) and the International Conference on Principles of Knowledge Representation and Reasoning (KR) were both held strictly in-person.

Neural Information Processing Systems (NeurIPS) continued to be one of the most attended conferences, with around 15,530 attendees (Figure 1.3.2).[14] The conference with the greatest one-year increase in attendance was the International Conference on Robotics and Automation (ICRA), from 1,000 in 2021 to 8,008 in 2022.

Number of Attendees at Select AI Conferences, 2010–22
Source: AI Index, 2022 | Chart: 2023 AI Index Report

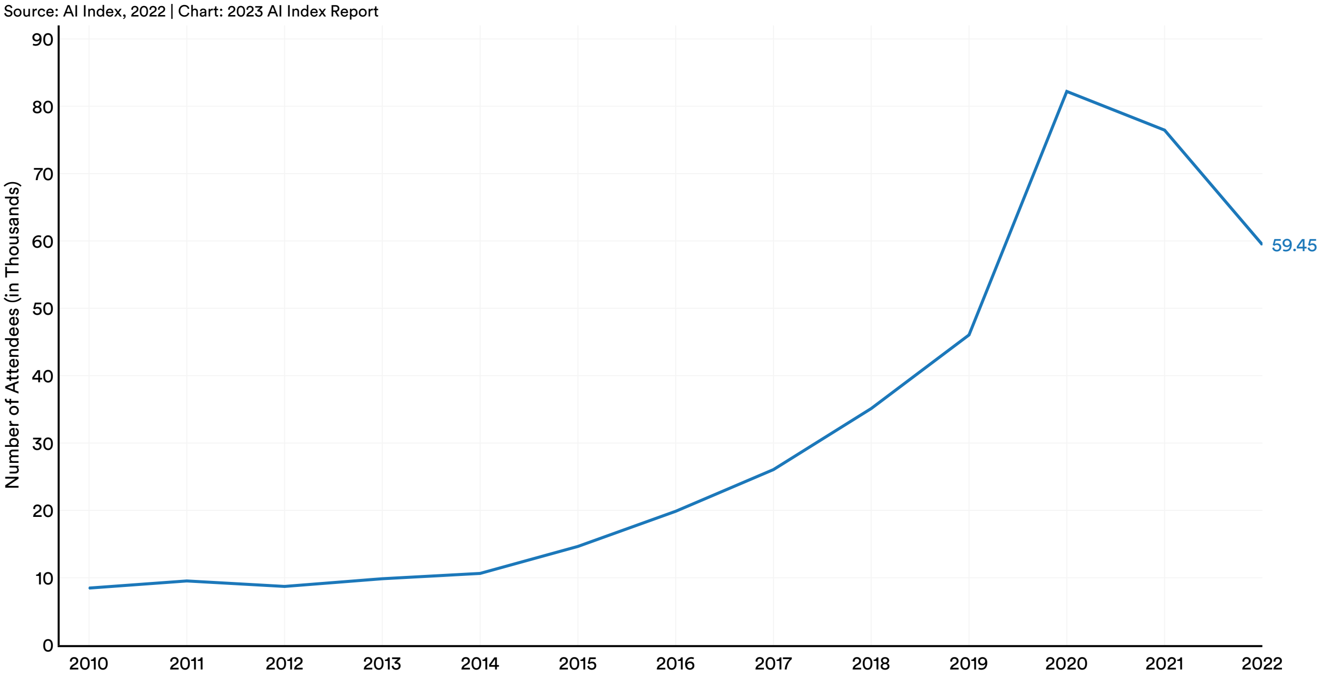

Figure 1.3.1

13 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.
14 In 2021, 9,560 of the attendees attended NeurIPS in-person and 5,970 remotely.

Attendance at Large Conferences, 2010–22
Source: AI Index, 2022 | Chart: 2023 AI Index Report

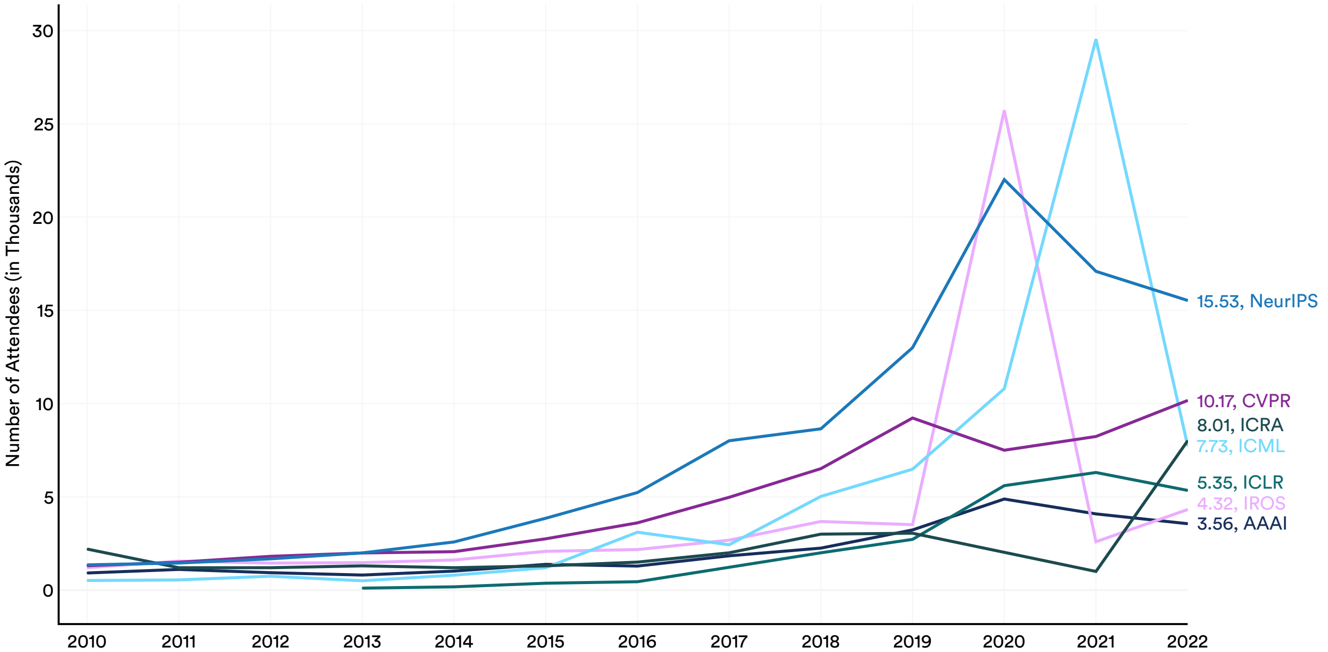

Figure 1.3.2

Attendance at Small Conferences, 2010–22
Source: AI Index, 2022 | Chart: 2023 AI Index Report

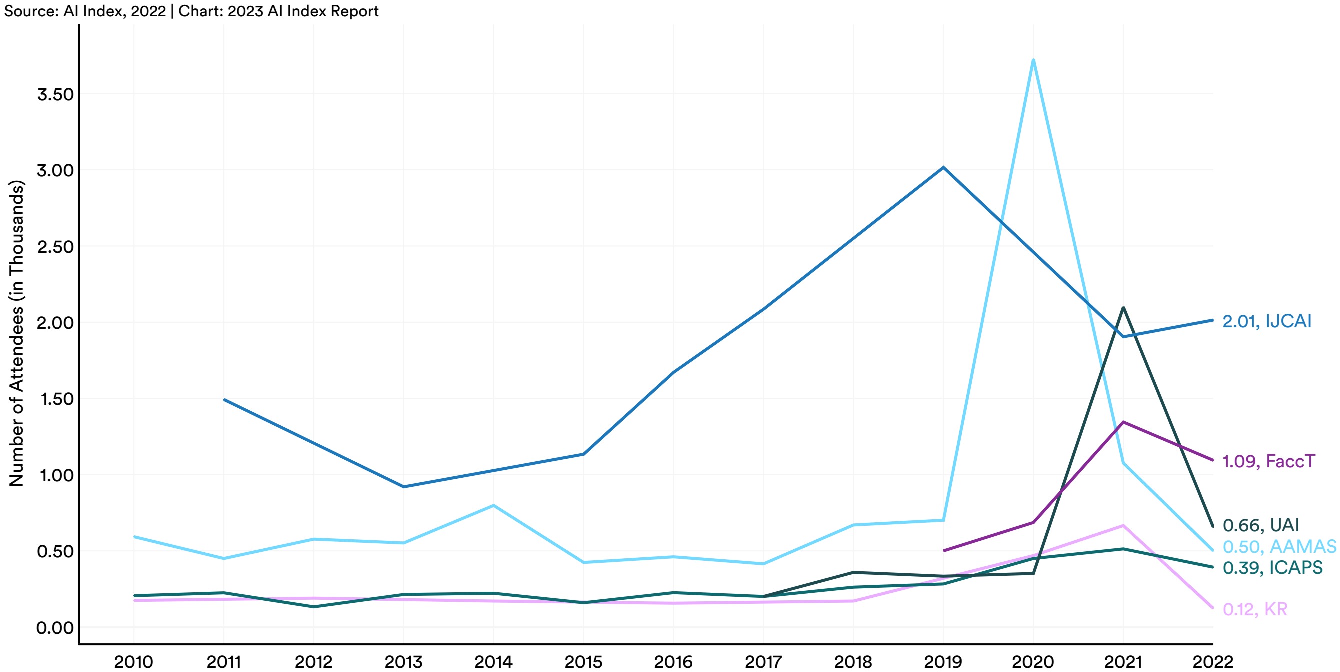

Figure 1.3.3

GitHub is a web-based platform where individuals and coding teams can host, review, and collaborate on various code repositories. GitHub is used extensively by software developers to manage and share code, collaborate on various projects, and support open-source software. This subsection uses data provided by GitHub and the OECD.AI policy observatory. These trends can serve as a proxy for some of the broader trends occuring in the world of open-source AI software not captured by academic publication data.

1.4 Open-Source AI Software

Projects

A GitHub project is a collection of files that can include the source code, documentation, configuration files, and images that constitute a software project. Since 2011, the total number of AI-related GitHub projects has steadily increased, growing from 1,536 in 2011 to 347,934 in 2022.

Number of GitHub AI Projects, 2011–22
Source: GitHub, 2022; OECD.AI, 2022 | Chart: 2023 AI Index Report

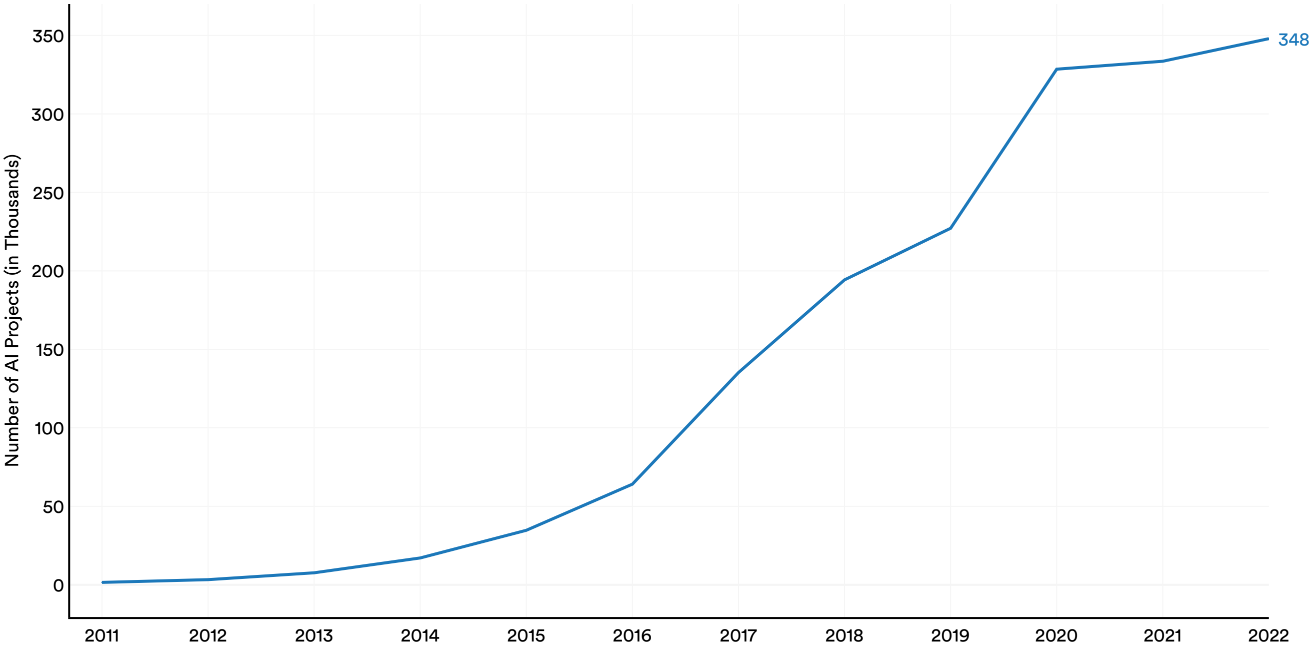

Figure 1.4.1

As of 2022, a large proportion of GitHub AI projects were contributed by software developers in India (24.2%) (Figure 1.4.2). The next most represented geographic area was the European Union and the United Kingdom (17.3%), and then the United States (14.0%). The share of American GitHub AI projects has been declining steadily since 2016.

GitHub AI Projects (% Total) by Geographic Area, 2011–22
Source: GitHub, 2022; OECD.AI, 2022 | Chart: 2023 AI Index Report

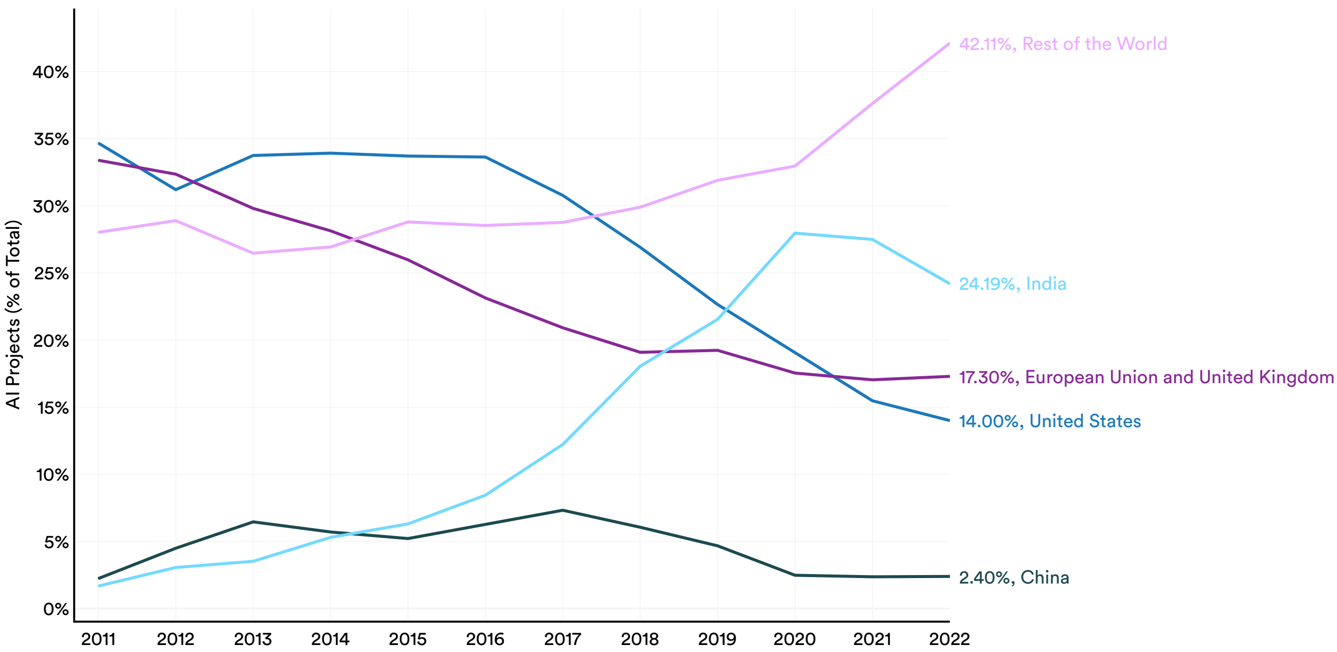

Figure 1.4.2

Stars

GitHub users can bookmark or save a repository of interest by "starring" it. A GitHub star is similar to a "like" on a social media platform and indicates support for a particular open-source project. Some of the most starred GitHub repositories include libraries like TensorFlow, OpenCV, Keras, and PyTorch, which are widely used by software developers in the AI coding community.

Figure 1.4.3 shows the cumulative number of stars attributed to projects belonging to owners of various geographic areas. As of 2022, GitHub AI projects from the United States received the most stars, followed by the European Union and the United Kingdom, and then China. In many geographic areas, the total number of new GitHub stars has leveled off in the last few years.

Number of GitHub Stars by Geographic Area, 2011–22
Source: GitHub, 2022; OECD.AI, 2022 | Chart: 2023 AI Index Report

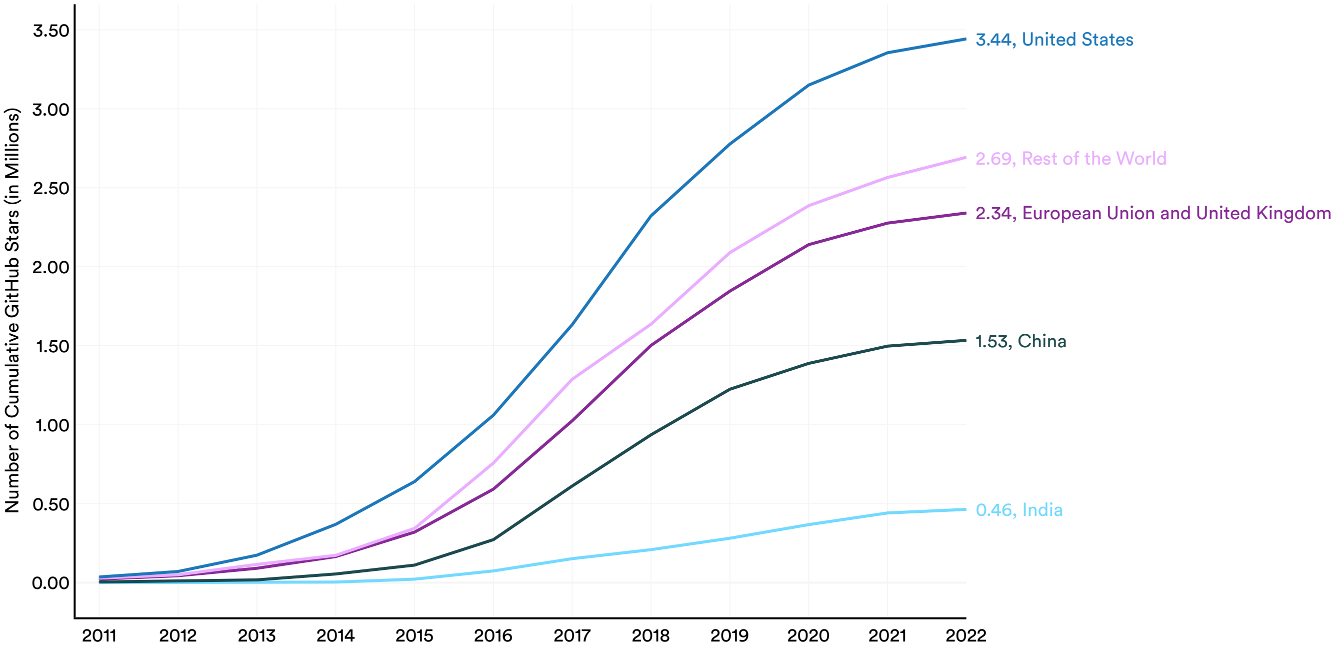

Figure 1.4.3

Appendix

Center for Security and Emerging Technology, Georgetown University

Prepared by Sara Abdulla and James Dunham

The Center for Security and Emerging Technology (CSET) is a policy research organization within Georgetown University's Walsh School of Foreign Service that produces data-driven research at the intersection of security and technology, providing nonpartisan analysis to the policy community.

For more information about how CSET analyzes bibliometric and patent data, see the Country Activity Tracker (CAT) documentation on the Emerging Technology Observatory's website.[1] Using CAT, users can also interact with country bibliometric, patent, and investment data.[2]

Publications from CSET Merged Corpus of Scholarly Literature

Source

CSET's merged corpus of scholarly literature combines distinct publications from Digital Science's Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers With Code.[3]

Methodology

To create the merged corpus, CSET deduplicated across the listed sources using publication metadata, and then combined the metadata for linked publications. To identify AI publications, CSET used an English-language subset of this corpus: publications since 2010 that appear AI-relevant.[4] CSET researchers developed a classifier for identifying AI-related publications by leveraging the arXiv repository, where authors and editors tag papers by subject. Additionally, CSET uses select Chinese AI keywords to identify Chinese-language AI papers.[5]

To provide a publication's field of study, CSET matches each publication in the analytic corpus with predictions from Microsoft Academic Graph's field-of-study model, which yields hierarchical labels describing the published research field(s) of study and corresponding scores.[6] CSET researchers identified the most common fields of study in our corpus of AI-relevant publications since 2010 and recorded publications in all other fields as "Other AI." English-language AI-relevant publications were then tallied by their top-scoring field and publication year.

CSET also provided year-by-year citations for AI-relevant work associated with each country. A publication is associated with a country if it has at

1 https://eto.tech/tool-docs/cat/
2 https://cat.eto.tech/
3 All CNKI content is furnished by East View Information Services, Minneapolis, Minnesota, USA.
4 For more information, see James Dunham, Jennifer Melot, and Dewey Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text," arXiv [cs.DL], May 28, 2020, https://arxiv.org/abs/2002.07143.
5 This method was not used in CSET's data analysis for the 2022 HAI Index report.
6 These scores are based on cosine similarities between field-of-study and paper embeddings. See Zhihong Shen, Hao Ma, and Kuansan Wang, "A Web-Scale System for Scientific Knowledge Exploration," arXiv [cs.CL], May 30, 2018, https://arxiv.org/abs/1805.12216.

least one author whose organizational affiliation(s) are located in that country. Citation counts aren't available for all publications; those without counts weren't included in the citation analysis. Over 70% of English-language AI papers published between 2010 and 2020 have citation data available.

CSET counted cross-country collaborations as distinct pairs of countries across authors for each publication. Collaborations are only counted once: For example, if a publication has two authors from the United States and two authors from China, it is counted as a single United States-China collaboration.

Additionally, publication counts by year and by publication type (e.g., academic journal articles, conference papers) were provided where available. These publication types were disaggregated by affiliation country as described above.

CSET also provided publication affiliation sector(s) where, as in the country attribution analysis, sectors were associated with publications through authors' affiliations. Not all affiliations were characterized in terms of sectors; CSET researchers relied primarily on GRID from Digital Science for this purpose, and not all organizations can be found in or linked to GRID.[7] Where the affiliation sector is available, papers were counted toward these sectors, by year. Cross-sector collaborations on academic publications were calculated using the same method as in the cross-country collaborations analysis. We use HAI's standard regions mapping for geographic analysis, and the same principles for double-counting apply for regions as they do for countries.

Epoch National Affiliation Analysis

The AI forecasting research group Epoch maintains a dataset of landmark AI and ML models, along with accompanying information about their creators and publications, such as the list of their (co)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 future 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 on November 14, 2022. This includes papers about landmark models, selected using the inclusion criteria of importance, relevance, and uniqueness, as described in the Compute Trends dataset documentation.[8]
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. The number of authors from each country represented are added up and recorded.

7 See https://www.grid.ac/ for more information about the GRID dataset from Digital Science.
8 https://epochai.org/blog/compute-trends; see note on "milestone systems."

If an author has multiple affiliations in different countries, they are split between those countries proportionately.[9]

3. Each paper in the dataset is normalized to equal value by dividing the counts on each paper from each country by the total number of authors on that paper.[10]
4. All of the landmark publications are aggregated within time periods (e.g., monthly or yearly) with the normalized national contributions added up to determine what each country's contribution to landmark AI research was during each time period.
5. The contributions of different countries are compared over time to identify any trends.

Large Language and Multimodal Models

The following models were identified by members of the AI Index Steering Committee as the large language and multimodal models that would be included as part of the large language and multimodal model analysis:

AlphaCode
BLOOM
Chinchilla
Codex
CogView
DALL-E
DALL-E 2
ERNIE 3.0
ERNIE-GEN (large)
GLM-130B
Gopher
GPT-2
GPT-3 175B (davinci)
GPT-J-6B
GPT-Neo
GPT-NeoX-20B
Grover-Mega
HyperCLOVA
Imagen
InstructGPT
Jurassic-1-Jumbo
Jurassic-X
Meena
Megatron-LM (original, 8.3B)
Megatron-Turing NLG 530B
Minerva (540B)
OPT-175B
PaLM (540B)
PanGu-alpha
Stable Diffusion (LDM-KL-8-G)
T5-3B
T5-11B
Turing NLG
Wu Dao 2.0
Wu Dao – Wen Yuan

Large Language and Multimodal Models Training Cost Analysis

Cost estimates for the models were based directly on the hardware and training time if these were disclosed by the authors; otherwise, the AI Index calculated training time from the hardware speed, training compute, and hardware utilization efficiency.[11] Training time was then multiplied by the closest cost rate for the hardware the AI Index could find for the organization that trained the model. If price quotes were available before and after the model's training, the AI Index interpolated the hardware's cost rate along an exponential decay curve.

The AI Index classified training cost estimates as high, middle, or low. The AI Index called an estimate high if it was an upper bound or if the true cost was more likely to be lower than higher: For example, PaLM was trained on TPU v4 chips, and the AI Index estimated the cost to train the model on these chips from Google's public cloud compute prices, but the

9 For example, an author employed by both a Chinese university and a Canadian technology firm would be counted as 0.5 researchers from China and 0.5 from Canada.
10 This choice is arbitrary. Other plausible alternatives include weighting papers by their number of citations, or assigning greater weight to papers with more authors.
11 Hardware utilization rates: Every paper that reported the hardware utilization efficiency during training provided values between 30% and 50%. The AI Index used the reported numbers when available, or used 40% when values were not provided.

internal cost to Google is probably lower than what they charge others to rent their hardware. The AI Index called an estimate low if it was a lower bound or if the true cost was likely higher: For example, ERNIE was trained on NVIDIA Tesla v100 chips and published in July 2021; the chips cost $0.55 per hour in January 2023, so the AI Index could get a low estimate of the cost using this rate, but the training hardware was probably more expensive two years earlier. Middle estimates are a best guess, or those that equally well might be lower or higher.

AI Conferences

The AI Index reached out to the organizers of various AI conferences in 2022 and asked them to provide information on total attendance. Some conferences posted their attendance totals online; when this was the case, the AI Index used those reported totals and did not reach out to the conference organizers.

GitHub

The GitHub data was provided to the AI Index through OECD.AI, an organization with whom GitHub partners that provides data on open-source AI software. The AI Index reproduces the methodological note that is included by OECD.AI on its website, for the GitHub Data.

Background

Since its creation in 2007, GitHub has become the main provider of internet hosting for software development and version control. Many technology organizations and software developers use GitHub as a primary place for collaboration. To enable collaboration, GitHub is structured into projects, or "repositories," which contain a project's files and each file's revision history. The analysis of GitHub data could shed light on relevant metrics about who is developing AI software, where, and how fast, and who is using which development tools. These metrics could serve as proxies for broader trends in the field of software development and innovation.

Identifying AI Projects

Arguably, a significant portion of AI software development takes place on GitHub. OECD.AI partners with GitHub to identify public AI projects—or "repositories"—following the methodology developed by Gonzalez et al.,2020. Using the 439 topic labels identified by Gonzalez et al.—as well as the topics "machine learning," "deep learning," and "artificial intelligence"—GitHub provides OECD.AI with a list of public projects containing AI code. GitHub updates the list of public AI projects on a quarterly basis, which allows OECD.AI to capture trends in AI software development over time.

Obtaining AI Projects' Metadata

OECD.AI uses GitHub's list of public AI projects to query GitHub's public API and obtain more information about these projects. Project metadata may include the individual or organization that created the project; the programming language(s) (e.g., Python) and development tool(s) (e.g., Jupyter Notebooks) used in the project; as well as information about the contributions—or "commits"—made to it, which include the commit's author and a timestamp. In practical terms, a contribution or "commit" is an individual change to a file or set of files. Additionally, GitHub automatically suggests topical tags to each project based on its content. These topical tags need to be confirmed or modified by the project owner(s) to appear in the metadata.

Mapping Contributions to AI Projects to a Country

Contributions to public AI projects are mapped to a country based on location information at the contributor level and at the project level.

a) Location information at the contributor level:

- GitHub's "Location" field: Contributors can provide their location in their GitHub account. Given that GitHub's location field accepts free text, the location provided by contributors is not standardized and could belong to different levels (e.g., suburban, urban, regional, or national). To allow cross-country comparisons, Mapbox is used to standardize all available locations to the country level.
- Top level domain: Where the location field is empty or the location is not recognized, a contributor's location is assigned based on his or her email domain (e.g., .fr, .us, etc.).

b) Location information at the project level:

- Project information: Where no location information is available at the contributor level, information at the repository or project level is exploited. In particular, contributions from contributors with no location information to projects created or owned by a known organization are automatically assigned the organization's country (i.e., the country where its headquarters are located). For example, contributions from a contributor with no location information to an AI project owned by Microsoft will be assigned to the United States.

If the above fails, a contributor's location field is left blank.

As of October 2021, 71.2% of the contributions to public AI projects were mapped to a country using this methodology. However, a decreasing trend in the share of AI projects for which a location can be identified is observed in time, indicating a possible lag in location reporting.

Measuring Contributions to AI Projects

Collaboration on a given public AI project is measured by the number of contributions—or "commits"—made to it.

To obtain a fractional count of contributions by country, an AI project is divided equally by the total number of contributions made to it. A country's total contributions to AI projects is therefore given by the sum of its contributions—in fractional counts—to each AI project. In relative terms, the share of contributions to public AI projects made by a given country is the ratio of that country's contributions to each of the AI projects in which it participates over the total contributions to AI projects from all countries.

In future iterations, OECD.AI plans to include additional measures of contribution to AI software development, such as issues raised, comments, and pull requests.

Identifying Programming Languages and Development Tools Used in AI Projects

GitHub uses file extensions contained in a project to automatically tag it with one or more programming languages and/or development tools. This implies that more than one programming language or development tool could be used in a given AI project.

Measuring the Quality of AI Projects

Two quality measures are used to classify public AI projects:

- **Project impact:** The impact of an AI project is given by the number of managed copies (i.e., "forks") made of that project.
- **Project popularity:** The impact of an AI project is given by the number of followers (i.e., "stars") received by that project.

Filtering by project impact or popularity could help identify countries that contribute the most to high quality projects.

Measuring Collaboration

Two countries are said to collaborate on a specific public AI software development project if there is at least one contributor from each country with at least one contribution (i.e., "commit") to the project. Domestic collaboration occurs when two contributors from the same country contribute to a project.