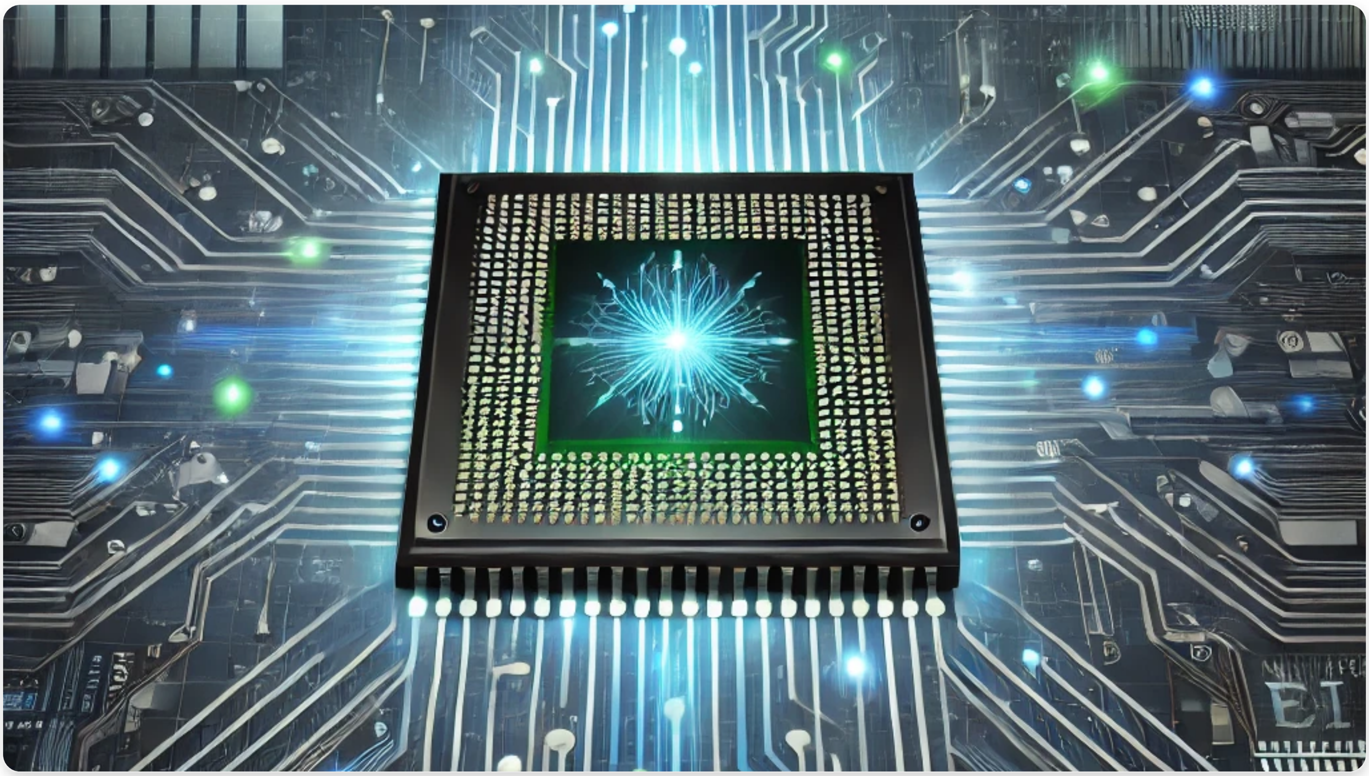


What drives progress in AI? Trends in Compute

January 3, 2025

Peter Slattery, Tal Roded, Emanuele Del Sozzo, and Haoran Lyu | MIT FutureTech

In this article, we provide a high-level overview of another key trend in AI models: that progress in hardware underpins further improvements in AI systems. This is the third in a series of articles on AI trends. We also have related articles on data and algorithms.

Background

Progress in Artificial Intelligence (AI) is underpinned by advances in three areas: compute, data, and algorithms. **Compute** refers to the computational resources – the physical hardware (e.g., CPUs and GPUs) which are used to run calculations and process data. **Algorithms** are the procedures or formulas that computer systems use to solve problems or complete tasks. **Data** refers to the information being processed or produced by a computer, for instance, to train and validate AI models. This article focuses on progress in **Compute** and the impacts on AI development.

How do we measure compute?

Compute is most often measured as the number of operations that the processor can perform in a given unit of time. This is done using various metrics that are frequently combined to provide a comprehensive view of a system's performance, with their relevance varying depending on the specific use case or application:

- FLOPS (Floating Point Operations Per Second) quantifies a processor's ability to handle complex mathematical calculations on real numbers
- MIPS (Million Instructions Per Second) counts how many basic machine instructions have been executed.
- Clock speed, measured in Hertz, indicates how many cycles a processor completes per second and is combined with other metrics to calculate performance like peak FLOPS.
- Storage performance is assessed using IOPS (Input/Output Operations Per Second), a measure of the overall computing system.
- Memory speed is measured by bandwidth in GB/s.

An interesting example of this is Berkeley's Roofline Model, which combines memory bandwidth and computing power to define the potential performance of a system when executing a given workload.

Why progress in compute matters

To provide a broader perspective, it is important to recognize how compute has evolved historically to overcome the limitations imposed by the end of Dennard scaling and the slowing of Moore's Law. Over the years, researchers have explored a variety of approaches to boost computing power, such as parallel processing, specialized hardware like Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), and distributed computing systems (see this paper for more detail).

These innovations have enabled significant strides in AI, especially as the complexity of models and datasets has grown exponentially. In particular, the transition from general-purpose CPUs to more specialized processors has helped address bottlenecks in training AI models, enabling faster and more efficient computation that directly contributes to the performance gains seen in modern machine learning applications.

As AI and machine learning models are further implemented, we will care more and more about their performance - and compute is one of the main drivers of this performance (Thompson, et al (2022)). For instance, Ho et al (2024) find that compute scaling has contributed roughly twice as much as algorithmic progress to more effective compute.

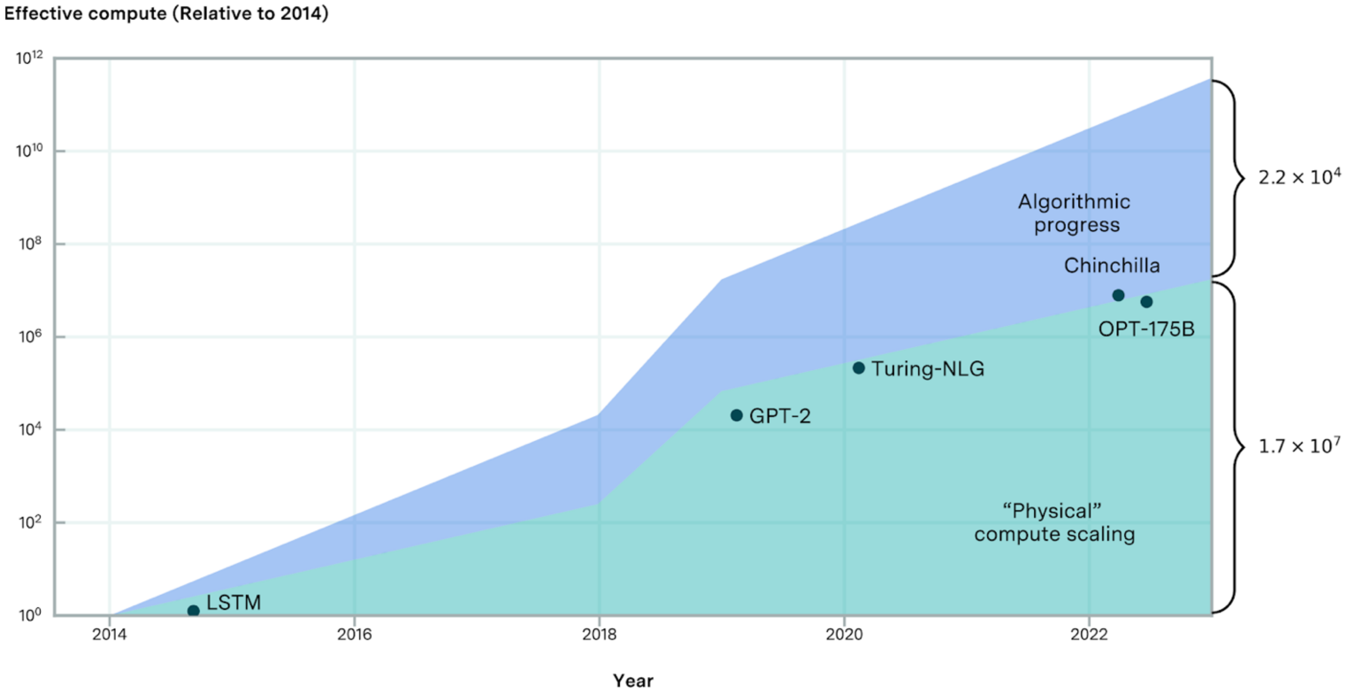
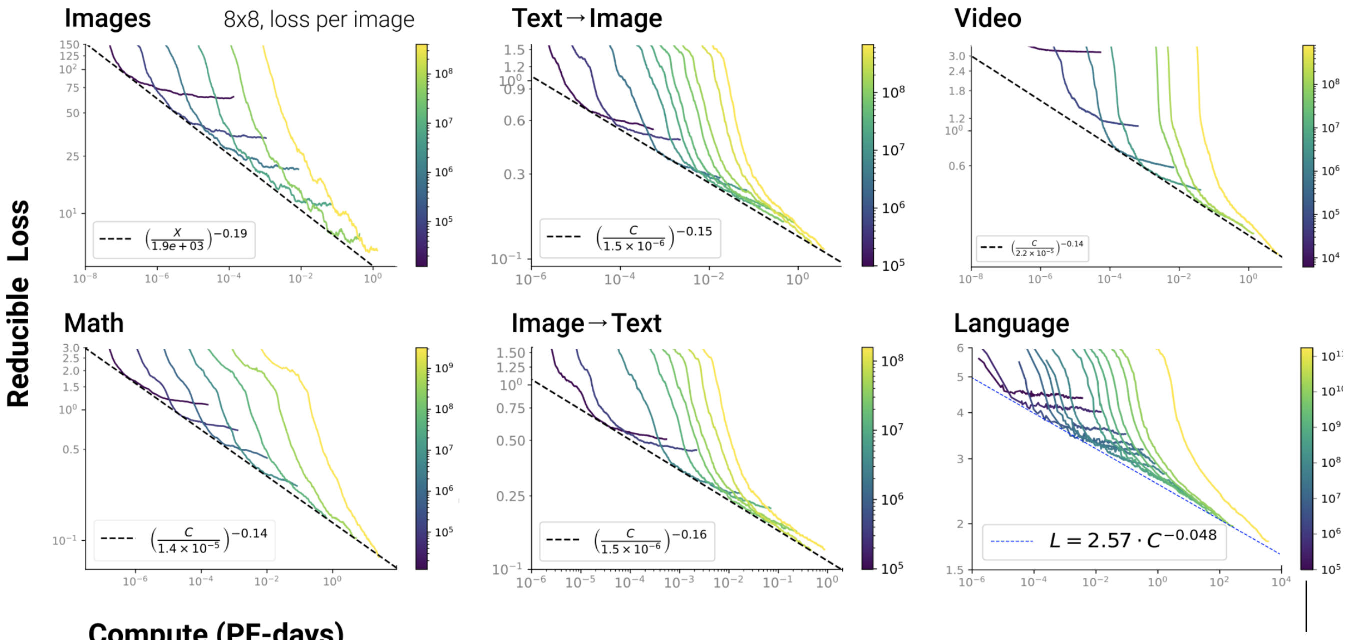


Figure 1: A stylized illustration of the relative contribution of compute scaling and algorithmic progress to effective compute [source]

Higher computing power allows models to execute algorithms and process data at a much faster rate, reducing run time and improving their learning and output performance. For instance, watch the video below to see how increased computation dramatically affects the performance of Sora, a text-to-video model made by OpenAI.

In essence, increasing the computation of the base Sora model sixteen times is the difference between almost incomprehensible and relatively realistic ouput.

We see similar effects of scaling computation in other AI domains. As Figure 2 below shows, as you increase the number of “Compute PF-days" (A measure of the computational resources used to train AI models) the loss (or error) of the model decreases at different rates across different application domains.



Line color denotes model size

Figure 2: Comparing Reducible Loss to Compute across AI domains [source]

Precisely understanding the connection between computation and performance is important because it helps us to understand the likely development of AI, where to invest, and what to watch out for.

For instance, say we know that you get a 2x increase in computation per dollar every year, and that every 2x increase in computation will lead to a 10% performance in the performance of text-to-video models.

We can use that information to estimate the year when performance meets different thresholds, for instance, thresholds that might make text-to-video commercially viable for replacing certain roles, or supporting certain antisocial activities.

Notably, these relationships will differ across domains; increased computation makes some types of AI performance improve more quickly than other types. For instance, the image shared above suggests that you can approximately double performance at creating images (on one metric) with ~40x more compute. However, the same progress in language production requires ~1,900,000x as much compute - a huge difference with many important implications for anyone investing in, or working in, those areas.

Progress in compute and its effect on AI

As Figure 3 shows, the computing power used by AI models has increased dramatically over time.

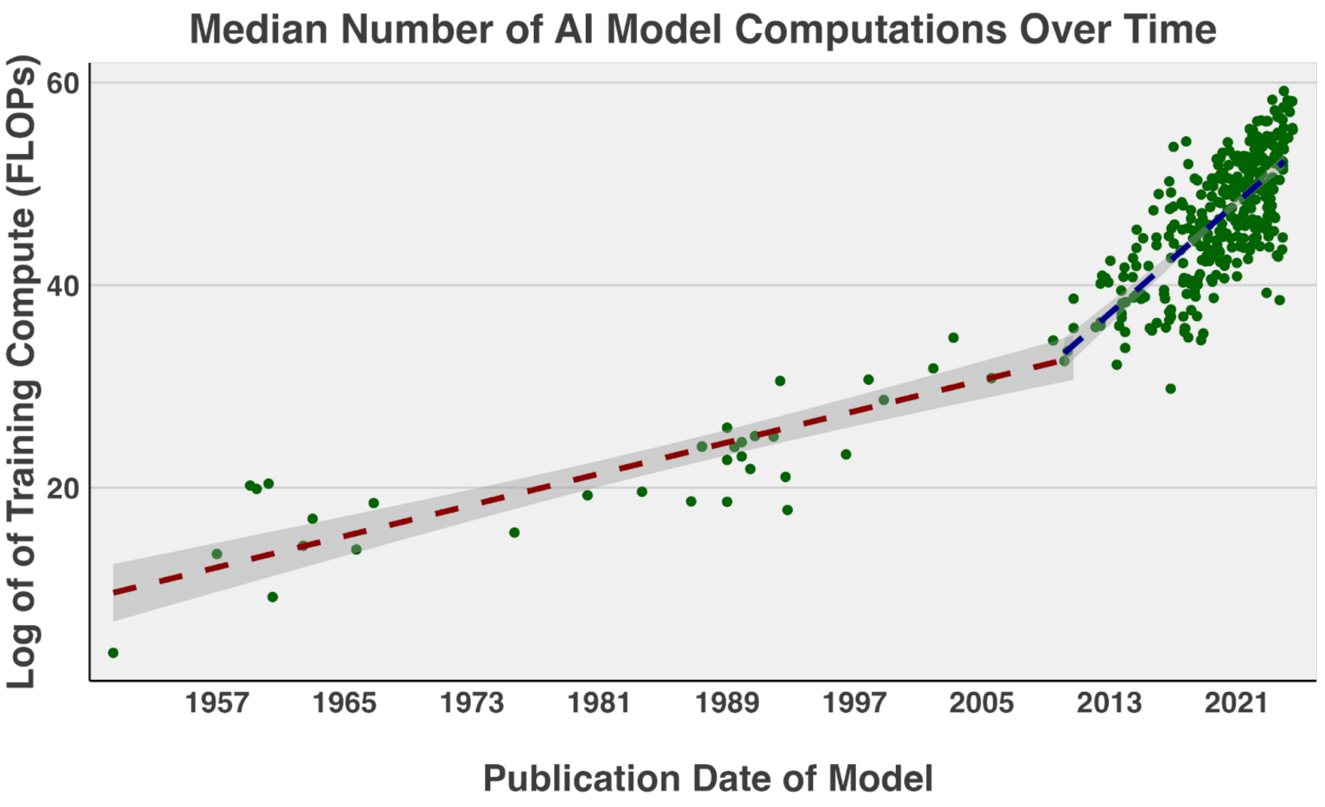


Figure 3: Trend of training compute, measured in FLOPs and taken in log, over time for AI models. A linear fit is added, represented by the dotted red line. [source]

The horizontal axis represents the years, from 1957 to 2021, when different AI models were published. The vertical axis shows the amount of computing power these models use, measured in "FLOPs". These are shown in Logs: Log 0 is 1 FLOPs (1 calculation per second), Log 10 is 10 billion FLOPs, Log 20 is 100 quintillion FLOPs.

The green dots represent individual AI models. As we move to the right (more recent years), these dots climb higher and higher, showing that newer models generally require much more computational resources.

What is striking about this graph is the sharp upward trend. In the early years, from the 1950s to around 2000, the increase in computing power was steady but relatively modest. However, starting around 2010, there was a dramatic surge upward where progress accelerated even faster.

To understand how this surge in computing power is possible, it is essential to consider the key hardware innovations that have driven these advancements. Central to the rise in AI model performance are specialized processors like GPUs and TPUs, which are specifically designed to handle parallel computations. GPUs, in particular, originally developed for graphics, have become the backbone of AI development, offering a significant performance boost over traditional CPUs by allowing many computations to be processed simultaneously.

TPUs, developed by Google specifically for AI, further optimize the performance of machine learning workloads, enabling faster training times and more efficient use of compute resources. These hardware advancements have played a crucial role in making the growth rate shown in Figure 3 possible.

Figure 4 shows that advances over the past decade have been made steadily across both industry and academia.

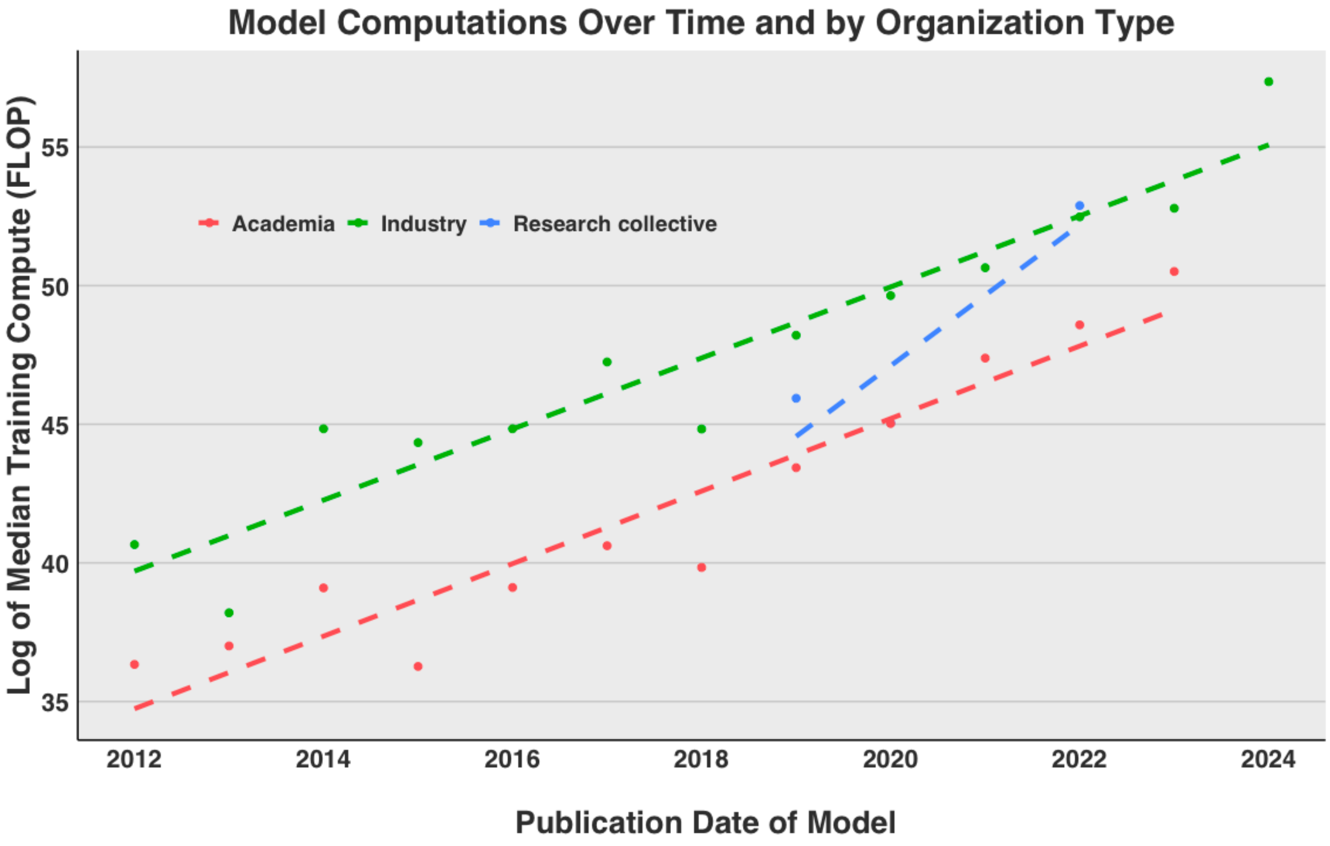


Figure 4: Trend of training computer measured in FLOPs and taken in log, over time for AI models and separated by organization category. A linear fit is added for each organization category, represented by the dotted lines. [source]

This graph illustrates the growth in AI model computations from 2012 to 2024, categorized by three types of organizations: academia, industry, and research collectives. The vertical axis shows the logarithmic scale of computational power (measured in FLOPs), while the horizontal axis represents the publication date of the AI models.

All three sectors show a significant increase in computational power over time, indicating rapid advancement in AI capabilities across the board. However, there are notable differences in their trajectories. The green line representing industry consistently sits at the top, suggesting that companies are consistently using the most computational resources for their AI models.

Interestingly, the blue line for research collectives starts later but shows a steep upward trajectory, nearly matching the industry level by 2024. This suggests a rapid ramp-up in collaborative research efforts. Meanwhile, the red line for academia shows consistent growth but generally uses less computational power compared to industry and research collectives.

The most recent major releases in AI, such as the transformative Large Language Models (LLMs) of the past several years, have been led by industry actors. This reflects the high costs of building and training AI models and represents a potential risk to the future development of both AI and compute as research is done by ever-fewer academic and public institutions.

Could AI contribute to hardware progress?

Recent developments in AI have both been caused by computing power improvements and have increased the pace of computing power growth. As these models grow in complexity and usefulness, we can expect them to also help accelerate the growth of compute. For instance, AI algorithms can analyze and optimize chip layouts, potentially discovering more efficient designs than human engineers. At the same time, the need for more computing power reshaped the microarchitecture of various computing devices due to the introduction of AI-oriented compute units, such as Tensor Cores in the NVIDIA GPUs.

AI progress may therefore significantly contribute to its own development. Such a scenario could lead to a feedback loop of escalating progress, where AI discoveries drive further algorithmic improvements and hardware optimizations: an acceleration which could have far-reaching implications.

Conclusion

Compute progress means that AI models can process more information and perform more complex tasks with increasing efficiency. This is crucial for AI development: As computational power grows, we can train larger and more capable models, and explore innovative approaches.

In addition to the significant gains in compute during the training phase, advances in specialized hardware like GPUs and TPUs also have a profound impact on inference performance and accessibility during deployment. These hardware innovations not only improve the efficiency of training large, complex models but also make the real-time operation of these models faster and more cost-effective, which is crucial for broad AI adoption.

For example, GPUs accelerate inference by parallelizing operations, allowing AI systems to make predictions or generate outputs in real time. Furthermore, the improved efficiency of hardware reduces the costs of deploying AI models, enabling smaller or less powerful systems to be built at a fraction of the cost compared to earlier models.

As a result, the same computational power that once trained massive models like GPT-3.5 is now more affordable, allowing companies and researchers to experiment with smaller-scale systems and enabling wider use cases in industries with less computational resources.

Our analysis shows that compute progress has been dramatic and consistent, particularly in recent years, with industry leading the charge. For AI-related advancements, we find that while algorithmic improvements play a vital role, the majority of performance gains have come from increased computational power.

Resources

For further reading on progress in compute and implications of compute trends, see Thompson, Neil C., Shuning Ge, and Gabriel F. Manso. "The importance of (exponentially more) computing power, our computer progress database and work from Epoch AI.

References and data

Epoch AI: Data on Notable AI Models

Epoch AI: Training Compute of Frontier AI Models Grows by 4-5x per Year

Brookings: What should be done about the growing influence of industry in AI research?

FutureTech
THE ECONOMIC AND TECHNICAL FOUNDATIONS
OF PROGRESS IN COMPUTING

MIT CSAIL

MIT MANAGEMENT
SLOAN SCHOOL

Find us

FutureTech
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Office 386, Cambridge, MA 02139
617-258-5030

MIT Accessibility