National Bank OF BELGIUM
Eurosystem

ECONOMIC POLICY

79th Economic Policy Panel Meeting

4-5 April 2024

Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT

Anton Korinek, Jai Vipra

CEPR CESifo SciencesPo.

Market concentration implications of foundation models: The Invisible Hand of ChatGPT[1]

Jai Vipra and Anton Korinek

Abstract

We analyze the technological and economic factors shaping the structure of the market for foundation models – large AI models such as those that power ChatGPT that are adaptable to downstream uses. We observe that the market for the most capable foundation models will have a tendency towards concentration, given significant economies of scale and scope, and examine the implications for competition policy and regulation. For frontier models, we discuss how antitrust authorities can address concerns about market competition and vertical integration. For models that are behind the frontier, we expect competition to be quite intense, implying a more limited role for competition policy. Moreover, we discuss how regulation can ensure a level playing field, pass data governance rules, forestall systemic risks from concentration, and internalize safety concerns.

[1] We thank Emma Bluemke, Aidan Kane, Sarah Myers West, Sanjay Patnaik, Nicholas Ritter, Max Schnidman, Eli Schrag and Rob Seamans as well as our editors, Emilio Calvano and Giacomo Calzolari, and two anonymous referees for their thoughtful comments. Any remaining errors are our own.

1. Introduction

In a post titled “Moore’s Law for Everything”, OpenAI CEO Sam Altman predicted that within the next few decades, AI technology would “do almost everything, including making new scientific discoveries that will expand our concept of ‘everything’” (Altman 2021). A paper co-authored by researchers at OpenAI estimated that most occupations are exposed to the deployment of large language models (LLMs) like ChatGPT, and that once complementary investments are made, up to 49 percent of workers could have half or more of their tasks exposed to LLMs (Eloundou et al. 2023). If AI models will indeed play such an important role in our economy, then the structure of the market in which they are offered will have first-order implications for social welfare. Policymakers, including antitrust authorities, thus need to pay close attention to the topic.

Generative AI is powered by foundation models – large AI models that use deep learning methods, are trained on vast amounts of data, and can be adapted to specific economic tasks and applications (Bommasani et al. 2021). Foundation models encompass large language models or vision language models such as OpenAI’s ChatGPT and Google DeepMind’s Gemini that can produce text and process and generate images, audio models that can transcribe or synthesize spoken text or songs, and multimodal models that combine these and other capabilities, for example producing videos and controlling robotic functions.

Building on these foundational capabilities, generative AI models have already shown great promise in performing economically useful tasks, oftentimes much faster and at lower cost than human workers. They can write computer code, find errors in code, write essays, copyedit text, summarize documents, generate ideas, interact with customers, produce images from broad instructions, recognize speech, and so on. There is already evidence that generative AI is disrupting certain categories of work (Hui et al 2023), and a survey of 1,000 US business leaders found that nearly half had replaced workers with ChatGPT (Shani 2023).

The size of the market for foundation models and generative AI was estimated to be just $3bn in 2023 (Fernandez et al. 2023) - a tiny sliver of a $100tn world economy. (In 2022, the term “generative AI” had not even been coined yet.) However, leading investment banks project that the technology may underpin 7 to 10 percent of global GDP within a decade (Hatzius et al. 2023; JP Morgan 2024), implying the potential for vast growth.

The market for foundation models is concentrated in large part because developing such models requires substantial fixed costs, primarily consisting of large expenses on computational power as well as skilled talent and large datasets. Training costs for the most advanced foundation models have, on average, grown by a factor of 3.1x/year for the past 15 years (Epoch, 2023). The most expensive publicly known foundation model trained to date is Google DeepMind’s Gemini, at an estimated cost of $630m (Epoch, 2023). Whether the trend of tripling training costs every year will go on depends on whether the economic benefits of foundation models will continue to scale with the required costs. Extrapolation comes with obvious perils, but many experts predict a

continuation of the trend for at least another 3 to 5 years (Suleyman 2023), and - given the economic promise of foundation models - possibly longer. As we elaborate in more detail below, a trillion dollar foundation model by the end of the decade is not inconceivable if the trend continues. In fact, in early 2024, OpenAI's CEO, Sam Altman, reportedly attempted to raise $7tn for the production of computer chips needed to train frontier AI models (Hagey and Fitch 2024).

The reason why technology companies are willing to spend large amounts on foundation models is that after the expensive first step, the so-called "pre-training," the model can be used as a foundation that is adapted to a wide range of downstream economic tasks. This process, called "fine-tuning," employs task-specific datasets and is significantly cheaper than pre-training (Bommasani et al. 2021).

There is an active debate in economics that discusses to what extent earlier technological advances offer useful analogies for the economic effects of artificial intelligence. Eloundou et al (2023) make the case that foundation models are general purpose technologies akin to the steam engine or electricity. The argument is that AI has the potential to transform numerous industries, demonstrates continuous improvement, and creates a powerful need for complementary technologies and societal adjustments, thereby fulfilling the core criteria of a general purpose technology.

The stated mission of several leading AI labs, including OpenAI, is even grander: that a future version of their foundation models will be able to achieve artificial general intelligence (AGI), defined as the ability to perform any cognitive task that humans can perform. If this mission were to be achieved, then their models could underpin any cognitive work and, if equipped with the necessary hardware, any human work that humans would let it perform, no matter which occupation or industry. This maximalist vision of the future role of foundation models is clearly far from a certainty, but given the rapid pace of recent advances, it may be useful to consider it as a scenario for which economists and economic policymakers should be prepared (Korinek, 2023). In such a scenario, the market for foundation models would be the entire economy.

Principal players in the market for generative AI models

Figure 1 and Table 1 show data on the market for generative AI models and the capabilities of the leading AI labs as of Feb 2024 according to the Massive Multitask Language Understanding (MMLU) benchmark, which measures the world knowledge and problem solving ability of LLMs on various subjects including sciences and the humanities across 57 tasks (Hendrycks et al 2020).

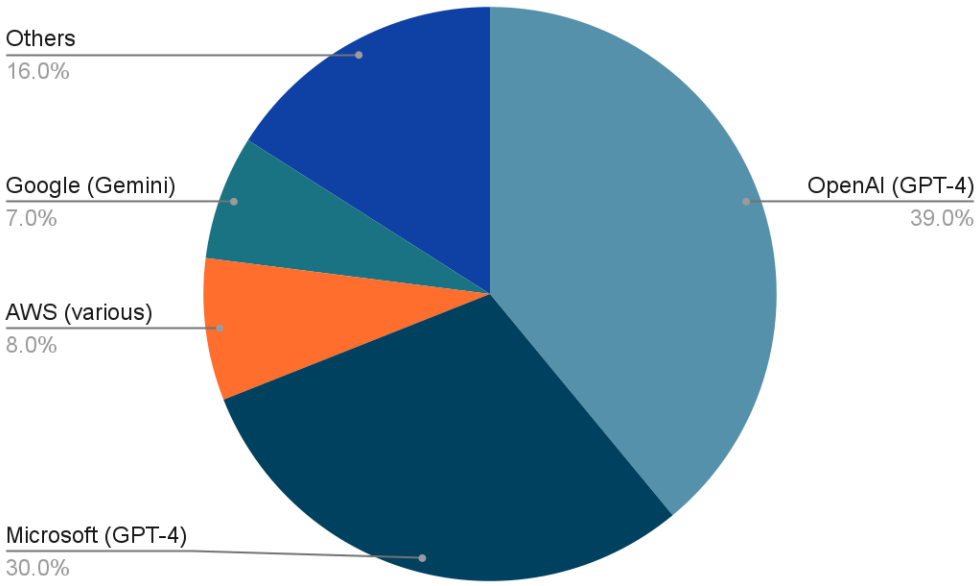

Figure 1: Principal players in the market for generative AI models (% of total spending)
Source: Fernandez et al (2023); 'Others' includes Anthropic, AI2I Labs, Cohere, Aleph Alpha, Hugging Face, Alibaba, IBM, and Baidu among others.

As illustrated in Figure 1, OpenAI's GPT-4 series was the clear market leader in 2023. OpenAI's models powered both the highly successful ChatGPT interface (39% market share) and Microsoft's offerings of generative AI (30% market share), together accounting for a 69% share of the market for generative AI. GPT-4 completed training in summer 2022, was released in March 2023 (OpenAI, 2023c), and, as indicated in Table 1, has held the title of the most powerful LLM ever since. OpenAI is structured as a non-profit, which majority-owns a for-profit entity, in which outside investors such as Microsoft have invested to help cover the tremendous cost of training frontier AI models. The investments in the for-profit are structured such that the non-profit will receive all profits once outside investors have been repaid an agreed multiple of their initial investment.[2]

The second player in the market is Google DeepMind, which is owned by Alphabet and released a foundation model called Gemini in December 2023. As illustrated in Table 1, Gemini has capabilities on par with GPT-4 (Gemini Team 2023). Google DeepMind is the result of a merger between Google Brain, Google's internal frontier AI department, and DeepMind, a British AI lab that Alphabet acquired in 2014 (Shu 2014). DeepMind had developed seminal AI models such as AlphaGo and AlphaFold, while Google Brain researchers developed the Transformer architecture

[2] For example, Microsoft's initial $1bn investment in 2019 was reportedly subject to a 100x profit cap, i.e., Microsoft will receive dividends up until its initial investment has been repaid one hundred times, before additional profits on this stake would go to the non-profit (Coldewey 2019). The profit cap is far from being reached, implying that investors still have substantial upside in the medium term.

upon which today's LLMs are based. Google entered the commercial market for generative AI systems later than OpenAI/Microsoft and currently only holds a market share of 7%.

AI firm	Country	Top AI model	Release date	MMLU
OpenAI	USA	GPT-4	Mar 2023	90.1
Google DeepMind	USA/UK	Gemini Ultra	Dec 2023	90.0
Mistral	France	Mistral Large	Feb 2023	81.2
Inflection	USA	Inflection-2	Nov 2023	79.6
01.AI	China	Yi (OS)	Nov 2023	78.8
Anthropic	USA	Claude 2	Jul 2023	78.5
Alibaba	China	Qwen 1.5 (OS)	Feb 2024	77.4

Table 1: Leading AI labs and their leading foundation models, release date, and score on the MMLU test
Source: compiled by authors

As illustrated in Table 1, the models of OpenAI and Google DeepMind are far more capable than the competition at the time of writing, with a significant gap in the MMLU score to their closest competitors. Moreover, both of the market leaders are expected to release updated versions of their models with even greater capabilities in 2024.

The market behind these two frontier models is dynamic and competitive, with several companies from the US, France, and China competing neck-on-neck. It also includes a number of open source models, indicated by "OS" in the table. Notably, Facebook/Meta released a series of models named LlaMA starting in 2023, on which outside researchers have built a large number of fine-tuned versions. Meta announced a drive towards building more advanced foundation models in early 2024, and is stockpiling hundreds of thousands of GPUs to this end (Heath 2024). Many of these models are small enough to run on laptops or cell phones. To add to the competition in this market segment behind the frontier, Google DeepMind released an open-source set of models named Gemma in Feb. 2024, which outperform other models in the same size class.

New model announcements that will make it into the ranks of Table 1 are widely expected in the coming months, for example from Meta, Amazon, x.ai, but it is unclear whether any of these models will be able to compete with the frontier models from OpenAI and Google DeepMind, especially as the latter two are themselves actively working on newer versions of their models that are expected to be released in 2024.

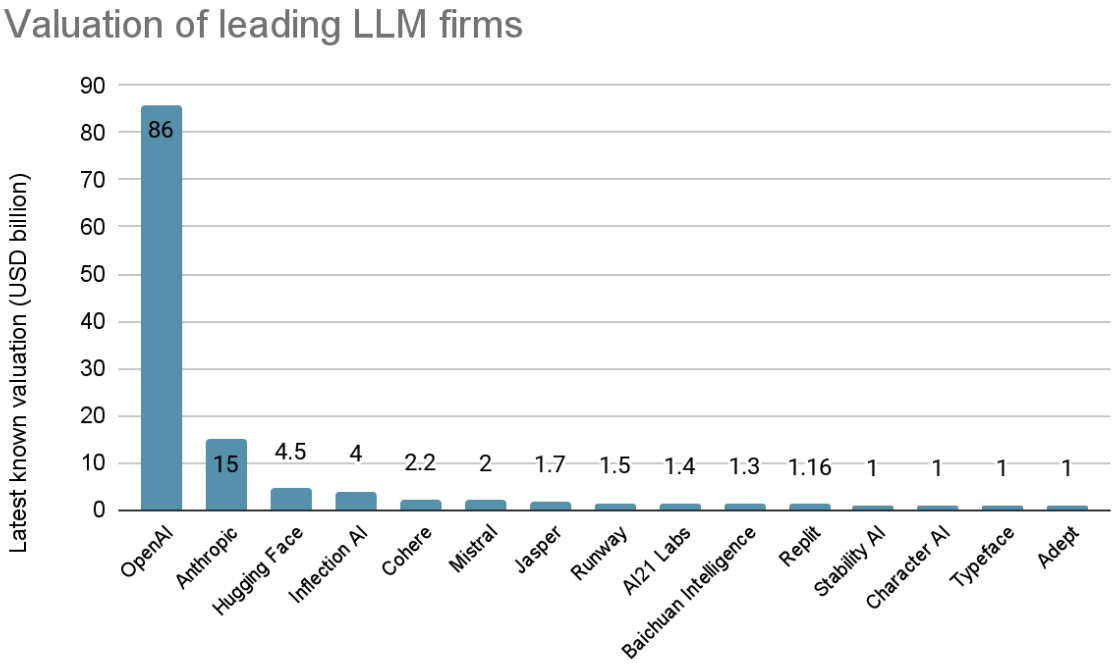

Figure 2: Most recent valuation of leading generative AI labs as of February 2023.
Source: Data collected by authors. (Latest valuation of Cohere not publicly announced.)

Figure 2 shows that there is also significant concentration when leading generative AI labs are measured by their valuation. This figure excludes Google DeepMind, which is a subunit of Alphabet for which valuation information is not publicly available. As a result, OpenAI's valuation stands head-and-shoulders above its competitors.

Company	Investments from large technology companies
OpenAI	Microsoft
Anthropic	Alphabet, Amazon, Salesforce, Zoom
Inflection AI	Microsoft, Nvidia
Hugging Face	Alphabet, AMD, Amazon, IBM, Intel, NVIDIA, Qualcomm, Salesforce
Cohere	Nvidia, Oracle, Salesforce
Mistral	Microsoft

Table 2: Investments in leading AI labs by large technology companies
Source: collected by authors

Table 2 illustrates that many of the investments in leading AI labs producing foundation models were conducted by leading large technology companies, including Microsoft, Alphabet, Amazon and Nvidia. It is important to keep track of these linkages when evaluating to what extent large technology companies have control over the market for foundation models.

An additional important consideration is that the market for the computer chips that are used to train and deploy foundation models is highly concentrated. Figure 3 shows that at the end of 2023, the market for Graphical Processing Units (GPUs) was dominated by a single company, Nvidia, which supplied chips to all the leading producers of foundation models listed in the tables above. In February 2024, a Wells Fargo report estimated that Nvidia controlled 98 percent of the data center GPU market (Norem 2024). Vipra and Myers West (2023) provide a rigorous description of the market for compute.

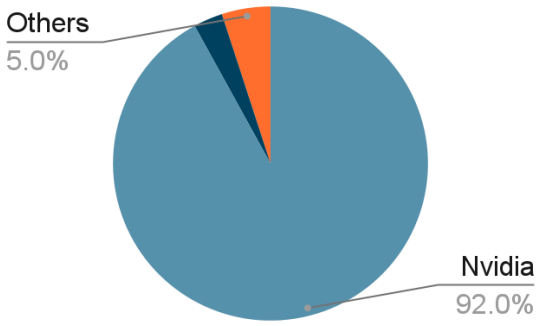

Figure 3: Market share of leading vendors in the market for GPUs
Source: Created by authors based on data from Fernandez et al (2023)

Recent attention from antitrust authorities

Given their growing importance in our economy, foundation models have already garnered the attention of antitrust regulators. In June 2023, the FTC released an assessment of competition concerns in generative AI, in which it highlighted the uneven control over its building blocks like data, talent, and computational power. It also drew attention to concerns over concentration in both generative AI markets, and other markets impacted by generative AI (Staff in the Bureau of Competition & Office of Technology 2023). In January 2024, the FTC launched an inquiry into generative AI investments and their partnerships with major cloud service providers, asking for information on partnerships between Microsoft and OpenAI, Amazon and Anthropic, and Google and Anthropic (Federal Trade Commission 2024).

The UK's Competition and Markets Authority (CMA) released a report in September 2023 proposing guiding principles to ensure competition in the market for foundation models, including access to key inputs like data and computational power, diversity of closed and open source business models, interoperability, fair dealing, and transparency, among others (Competition and Markets Authority 2023).

The European Commissioner for Competition, Margrethe Vestager, recently stated that merger control, vertical integration and algorithmic collusion are areas of interest for the EU in relation to

AI markets. (Lomas 2024a) The EU is also scrutinizing Microsoft's investment in OpenAI. (Lomas 2024b)

2. Technological characteristics and market structure

This section describes a number of technological and economic forces that characterize the market for foundation models. We observe that rapidly growing fixed costs generate significant economies of scale and scope for foundation models and observe the importance of inputs to production such as computational resources and data. Moreover, we identify strong forces toward vertical integration.

Cost structure of foundation models

Producing and operating foundation models involves three main types of costs: (i) a significant fixed cost for the pre-training of foundation models; (ii) an additional fixed cost per area of application when foundation models are adapted ("fine-tuned") for specific use cases; and (iii) low variable costs of operating the model.

Pre-training is the process of creating a foundation model, which can then be adapted for a wide range of use cases. The costs of pre-training have risen rapidly in recent years, driven primarily by growing spending on computational resources ("compute"). *The Economist* (2023) estimates the cost of training GPT-4 in the first half of 2022 at $100 million; Epoch (2023) estimates that Google DeepMind's Gemini cost around $630 million, approximately 3/4 of which was spent on compute, and the remainder covering personnel expenses. We will describe the drivers behind this rapid growth in spending on compute in the next subsection.

Fine tuning refers to the process of training a pre-trained model for a specific purpose, usually by using application-specific data. While exact figures are difficult to obtain, these costs are much lower than the fixed cost of pre-training a model, because fine-tuning requires less time, data, and compute. They include wages of in-house workers as well as the cost of outsourced workers who label data and help train models via reinforcement learning from human feedback (RLHF).

One factor that could push up these costs in the future is that fine-tuning requires data that is labeled, purpose-relevant, and may therefore be costlier to obtain. However, when the producer of a foundation model provides access to an AI application company, the purpose-relevant data is usually provided by the latter. For instance, if OpenAI provides access to GPT-4 to a healthcare provider, the healthcare provider can use its existing data on treatment patterns to fine-tune the model to its needs.

Variable costs of operation ("inference costs") are significantly lower but depend on the specific application. For example, at the time of writing, OpenAI charges $0.03 to generate 1000 syllables ("tokens") of text using its GPT-4 model – a number that is estimated to be close to the company's

cost of inference. However, one operative question for deploying foundation models in the broader economy is how many inferences per hour are needed to replace a human worker, and how much this would cost.

Since fixed costs are high and variable costs relatively low, foundation models offer a classical example of economies of scale. Moreover, the general purpose nature of foundation models also gives rise to significant economies of scope. Since one foundation model can be adapted to many different areas of application across different industries, economies of scope are expected to be very large. For instance, a single foundation model such as GPT-4 or Gemini can be used to automate copyediting, to create holiday itineraries, to check for errors in computer code, or to provide health advice. More generally, OpenAI, for instance, already allows developers to create plugins or custom GPTs to apply GPT-4 to many different applications and benefit from these economies of scope, allowing users to order groceries, search for flights, learn languages, and shop online, all by using the same foundation model (OpenAI 2023b).

Compute

The computational resources (frequently abbreviated by the neologism "compute") required to train frontier foundation models are massive. In recent years, AI researchers identified "scaling laws" for foundation models that predict how model performance increases with the amount of compute used in training the model (see e.g., Kaplan et al. 2020; Hoffman et al., 2022). To make better models, developers have to add more computational power. These regularities are important because they reduce the uncertainty that companies face when they make investment decisions.

Building on the described scaling laws, leading AI labs have increased the amount of compute deployed in frontier AI models by a factor of 4.1x per year over the past 15 years, as illustrated in Figure 4. Given reductions in chip prices, which evolved approximately in line with Moore's Law, Cottier (2023) finds that spending on compute for frontier AI systems has grown by a factor of 3.09x per year over the same period.[3]

[3] Estimates of compute costs for pre-training AI models usually take into account only the compute required for the final training run, and not the compute employed by trial training runs used to arrive at a workable architecture, which implies even greater compute costs (Heim 2021).

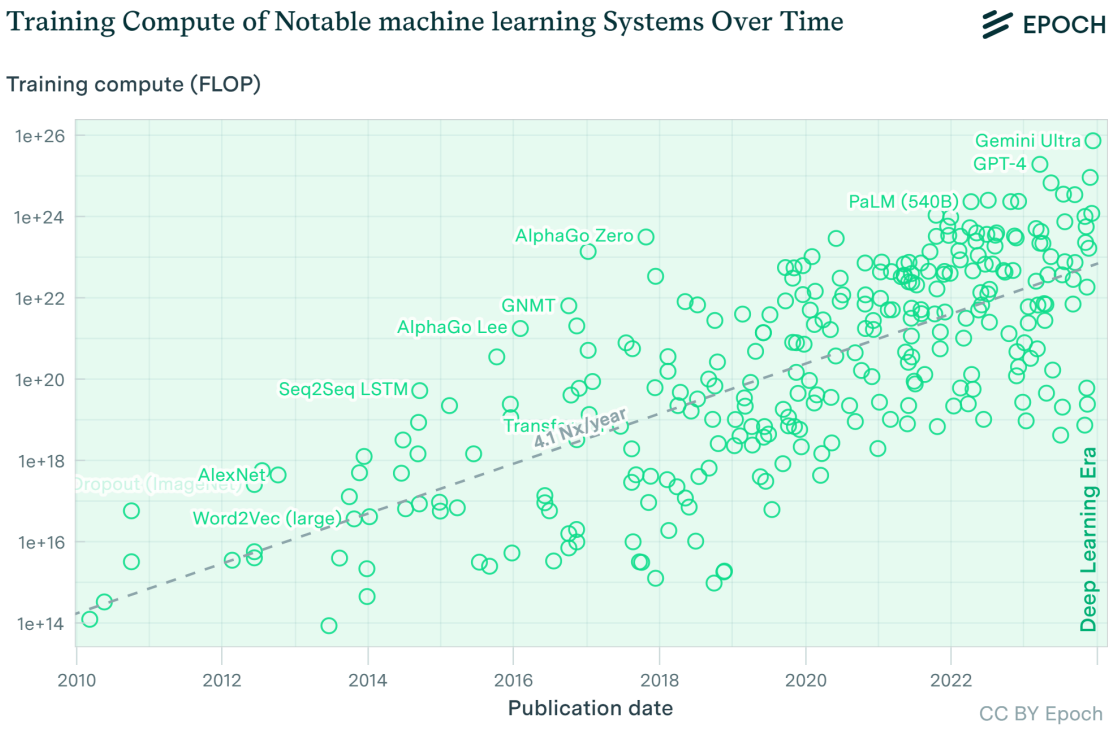

Figure 4: Training compute of notable AI systems. Copyright © by Epoch.org, reproduced under a CC-BY-4.0 license, 2024.

Whether this trend will go on will depend on whether not only the capabilities but also the economic benefits of foundation models will continue to scale with the amount of spending. Extrapolation comes with obvious perils - at first, any successful new technology grows faster than the rest of the economy as it starts from zero. Eventually, no sector can grow faster than the economy as a whole. This implies that we would expect the growth of AI training costs to eventually slow down as they become a more and more significant part of the economy. However, concurrently, economic growth may take off if advanced foundation models can replicate a growing fraction of the tasks that were traditionally performed by scarce labor (see, e.g., Aghion et al. 2019 and Trammell and Korinek 2023).

These considerations notwithstanding, technology leaders such as Suleyman (2023) and analysts predict a continuation of the described trend for at least another 3 to 5 years, and - given the economic promise of foundation models - possibly longer. We illustrate the implications in Figure 5, in which we start with the approximate compute training cost of Gemini and visualize the 3.09x/year trend growth of Cottier (2023), with lower and upper 90% confidence bounds given by growth factors of 2.34x and 3.63x per year respectively. If these rates continue, a naive extrapolation implies that it is not inconceivable that a technology company may spend more than $1tn on a single foundation model before the end of the decade. This is also consistent with reports that OpenAI's CEO, Sam Altman, attempted to raise $7tn for the production of computer chips needed to train frontier AI models in early 2024 (Hagey and Fitch 2024).

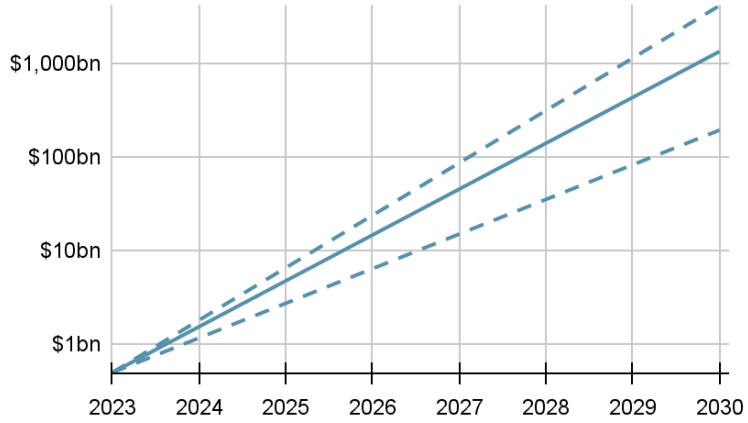

Figure 5: Extrapolation of Training compute of frontier AI systems, based on estimates from Cottier (2023).

Recent increases in demand for computational power may drive up costs even further. The costs of manufacturing additional equipment for producing computer chips are high, as they rely on highly specialized manufacturing techniques and high fixed costs (Khan, Peterson, and Mann 2021). Moreover, the semiconductor design and manufacturing industry is highly concentrated, as illustrated in Figure 3. Vipra and Myers West (2023) describe the market for compute in further detail. (We should probably cut this in the final version, but one of the world's leading technology experts, Elon Musk, proclaimed that "GPUs at this point are considerably harder to get than drugs," WSJ 2023).

There are also forces that pull in the opposite direction. The high returns generated by the scarcity of computer chips as well as government subsidies provide strong incentives for investment in additional capacity and may help to diversify the semiconductor supply chain and bring down costs. New breakthroughs in computing technology, such as neuromorphic computing, quantum computing, or improvements in memory technology, might ease concentration in the market for chips.

Overall, however, we expect that access to compute will continue to be a bottleneck for training frontier foundation models. This will likely restrict access to the compute needed for state-of-the-art models to a small number of well-funded AI labs like OpenAI and Google DeepMind. By contrast, for less advanced models, compute requirements are lower and more attainable so other companies, ranging from large tech firms to AI startups, will compete in the broader market for less advanced foundation models. This two-tiered structure may lead to some market power for the leading firms offering top-tier models, while we expect significant competition for the residual demand for lower-capability models.

Data

Foundation models are trained on large quantities of data. The most data-intensive language model at the time of writing was Google's FLAN, trained on 1.87 trillion words (Epoch 2023). Models are data-hungry and achieve higher performance when trained on higher quality text (Anil et al. 2023). Yet we are about to run out of high-quality text that is publicly available on the internet soon (Villalobos et al. 2022).

This makes proprietary datasets very useful for companies producing foundation models, which gives an advantage to large technology companies, which control a large fraction of the data generated online – including platform interactions, search, emails, photos, videos and other documents. Access to data is therefore an important reason why producers of foundation models may be interested in vertically integrating with Big Tech companies.

Over the course of 2023, a growing number of content providers on the internet have restricted free access to their content to the bots that automatically scoop up data used for the training of foundation models. For example, Fletcher (2024) finds that 79% of all news sites in the US blocked OpenAI's crawlers, as did about half of all news sites in a sample of ten advanced countries. The New York Times has even filed a copyright infringement lawsuit against OpenAI and Microsoft for using its news articles to train OpenAI's foundation models.

There are emerging techniques that can reduce the cost of acquiring data. These include simulation learning (where a simulated environment substitutes for a real training environment), self-play (where a model can interact with itself to improve its performance), and synthetic data generation (Hwang 2018; Azizi et al. 2023). Some of these techniques are domain-specific and do not work for all foundation models; some, like synthetic data generation, decrease accuracy. When such techniques are successful, they increase compute requirements, effectively substituting compute for training data.

Fine-tuning foundation models for specific applications also requires data – for instance, a manufacturer will be able to derive the most value from a foundation model by fine-tuning it on its own historical data. Such data may encompass detailed records ranging from production plans, machinery, product quality, to supply chain logistics, offering a granular view into every aspect of their operations. By leveraging this multi-dimensional dataset, the manufacturer can unlock insights for automating processes, reducing downtime, and enhancing product quality, thereby achieving a competitive edge in the industry. In many sectors, specific proprietary data held by traditional companies will be a valuable source of training data for fine-tuning purposes.

Vertically integrating economic foundations

Producers of foundation models may engage in vertical integration both upstream and downstream. Upstream vertical integration refers to integration with suppliers of inputs to

producing foundation models, the most important of which are compute and data. Downstream vertical integration refers to integration with producers or distributors of final goods and services that employ foundation models.

Vertical integration may increase consumer welfare by eliminating multiple markups in the production process, ultimately leading to lower final prices. Furthermore, it may mitigate the risks associated with specialized assets stemming from hold-up problems. An integrated firm may reduce transaction costs by maintaining better quality control over its products and streamlining operations, potentially leading to both lower prices and higher quality goods for consumers. However, by consolidating market power, vertical integration may also pose the risk of foreclosing rivals' access to essential inputs or distribution channels or enabling firms to impose vertical restraints, thereby increasing monopoly distortions. It may also reduce innovation by lowering the number of independent players competing on new technologies and ideas (see, eg, Tirole 1988, Bresnahan and Levin 2012).

Upstream Since compute has the largest input share in creating foundation models, a growing number of AI companies have vertically integrated the production of chips and foundation models. An example of this is Google DeepMind, which produces its own chips, termed Tensor Processing Units (TPUs), which are particularly efficient for AI training. Similarly, the leading chip provider for training AI models, NVIDIA, is also offering foundation models together with a training platform that allows customers to fine-tune these models for their own purposes. Amazon and Microsoft, two of the leading providers of cloud computing, are both working on in-house designed chips for AI applications.

Another type of vertical integration occurs when providers of cloud computing services and foundation models enter exclusive contracts. Microsoft's investment in OpenAI gave Microsoft the right to be the exclusive provider of cloud services for OpenAI. This provision covers both training and provision of the model (Warren 2023). Likewise, Google Cloud is the preferred provider of cloud services for Anthropic ('Anthropic Partners with Google Cloud' 2023). Many of the investments listed in Table 2 are of a similar nature.

With data an important input in pre-training foundation models, there is also significant potential for vertical integration between data-rich technology companies and producers of foundation models. Google, Meta and X/Twitter are employing data from their public platforms like (including YouTube, Facebook, Instagram, Twitter) in pre-training their foundation models (Victor 2023, Olson 2024b). At the same time, data-rich technology companies are restricting data access to webcrawlers that allow other producers of foundation models to train on their data (Verhulst, 2023), raising anti-competitive concerns.

Downstream As foundation models have a growing number of general use cases for cognitive workers, they are rapidly being integrated into office suites. For example, Google and Microsoft have both integrated generative AI capabilities into their products, including Microsoft Office,

Gmail, and Google Documents. Google also offers a beta version of search that incorporates generative AI (Reid 2023).

OpenAI is integrating its GPT-4 model in a growing number of downstream uses by allowing commercial providers to create custom GPTs that are adapted and fine-tuned for specific use cases. Since their launch in November 2023, more than a hundred thousand custom GPTs have been created that link ChatGPT to commercial applications, allowing users access to additional third-party functions that range from travel itineraries to incorporating Wolfram Mathematica in their chats. This may be the first step of turning ChatGPT into a platform.

As the capabilities of foundation models improve and as companies devise strategies for how to effectively incorporate them into their operations, downstream use cases and by extension the scope for vertical integration may continue to grow.

3. Scaling and Concentration Concerns

Competition among foundation models

At the time of writing in early 2024, startups and leading technology companies around the world are racing to develop competing foundation models and generative AI offerings as fast as possible, and fierce competition has driven the price of powerful foundation models literally to zero: Part of the strategy of firms that are attempting to catch up to market leaders, e.g., Meta, has been to open-source their model releases, i.e., to make them available to download for free to anyone who wants to use them. Although these are not models at the frontier like the ones listed in Table 1, open sourcing comes with significant economic benefits – pre-trained foundation models are non-rivalrous so free distribution of the model corresponds to the first-best price and the maximum level of consumer surplus. Moreover, open sourcing also allows researchers and other companies to build on the foundation model and fine-tune it, encouraging innovation. However, open sourcing highly capable foundation models also carries safety risks, especially if models become more powerful, as we will discuss in the last section.

OpenAI and Google DeepMind, the producers of the two leading frontier models described in Table 1, are also engaged in fierce competition, charging prices that barely allow them to cover their variable costs (Knight 2024). The competition dynamics between the two seem to be close to Bertrand competition.[4] If competition dynamics remained as fierce as they currently are, then there would be little reason for concern about market concentration.

[4] There are even some indications that inference costs may currently be priced below cost (Patel and Nishball 2023). However, determining whether pricing is below cost is tricky in nascent industries with high fixed costs and uncertain demand.

However, the investment requirements for frontier foundation models are expected to rise rapidly, making it more difficult for a large number of players to compete in the market. Moreover, as described in Section 2, the role of generative AI in the economy is expected to grow significantly from its current $3bn market size (Hatzius 2023). Moreover, the current top players enjoy significant first-mover advantages from their technological leadership, from the preemption of scarce assets, and from buyer switching costs (Lieberman and Montgomery 1988). The scarce assets accumulated by the pioneers include vast amounts of compute and datasets. Whether companies like OpenAI can maintain their first-mover advantage remains to be seen.[5] However, taken together, these trends imply that it is important for policy to ensure that the market for foundation models remains competitive.[6]

Compute concentration

Whereas the market for foundation models itself is fiercely competitive, the market for chips on which the training and inference of such models relies is highly concentrated, as illustrated in Figure 3. The creation of cutting-edge chips is a process that is highly sophisticated and involves massive R&D costs, giving rise to a complex supply chain. Moreover, from the design, the manufacturing of equipment to fabricate chips, to the fabrication itself, there are several companies that are close to monopolists in their respective functions.

Due to the complexity of the supply chain for computational power, increasing the competitiveness of computational power markets is extraordinarily difficult. Regulators can still ensure that this market is not further concentrated. Belfield and Hua (2022) discuss classical options such as disallowing some mergers and acquisitions among chip companies, seeking remedial action from entities wishing to merge; targeting abuse of dominance, including prices that are too high (although this would be difficult to monitor & implement in practice), bundling, self-preferencing, etc.; and ensuring that horizontal agreements relating to hardware standards are not anti-competitive.

Vipra and Myers West (2023) propose that structural separations in the ecosystem for computational resources can help reduce the concentration in this industry. They recommend

[5] After the release of ChatGPT, fierce technological competition also seems to have changed the open publication norms of the sector, making it more difficult for new entrants to catch up. For instance, Google changed its long-standing practice of releasing AI research papers freely, opting instead to hold back critical model details (Tiku and Vynck 2023).

[6] We should also note that some capability improvements in the future could change the prevailing competition dynamics entirely. If a lab such as OpenAI manages to develop a foundation model that has the ability to improve itself without human input, then the pace of technological progress could be rapid and sustained. This would cause first mover advantages to snowball, leave the competition ever further behind, and create a large monopoly. If such a breakthrough is achieved by a non-profit entity, then there is the potential that it will maximize social welfare rather than engaging in inefficient monopolistic behavior. However, although OpenAI is notionally owned by a non-profit, it is unclear what considerations will drive the organization's behavior.

that antitrust regulators examine the viability of separating cloud provision from chip design, compute hardware from compute software (like NVIDIA's CUDA), and model development from cloud infrastructure. They propose that these measures could reduce lock-in to compute ecosystems.

Tackling vertical integration

As the role of foundation models in the economy grows, their widespread potential applicability implies that vertical integration may become a growing concern for competition. Merger review in this area would have to pay attention to how any proposed mergers affect the cost of inputs to this market, and whether they might privilege some participants in downstream markets to the detriment of others.

Antitrust oversight will also be required even when coordination does not take the form of mergers. For instance, exclusive and preferred use contracts among computational power providers and foundation model companies are rife – as in OpenAI's deal with Microsoft that gives the latter exclusivity, or Anthropic's deal with Google to use its cloud services, raising competition concerns and potentially foreclosing the market for other entrants.

Moreover, early access provisions by foundation model producers may also raise competition concerns that may grow in importance as these models are deployed throughout the economy. For example, OpenAI offered beta access to GPT-4 to certain companies (including Duolingo, Stripe and Morgan Stanley), which privileged them in relation to their competitors.

Antitrust regulators should pay keen attention to acquisitions made by foundation model companies, especially of startups that might compete with them. Large technology companies are reportedly already wary of making acquisitions in AI, and are using investments instead (Olson 2024a). Antitrust scrutiny in investments is also growing. Governments should equip antitrust regulators with stronger ex ante powers to stop acquisitions that can be shown to significantly reduce competition, although how to deal with 'nascent competitors' is tricky in practice (Hemphill and Wu, 2020). Ex post measures in digital markets are often too little, too late as important intellectual property can be transferred before a merger or acquisition is undone.

Moreover, as foundation models become more and more integrated into the economy, they may provide intellectual infrastructure for a wide range of economic functions and play a similar role to public utilities like electricity. Regulators may need to respond to certain concerns that arise from the wide range of their uses. For example, anyone excluded from their services may be at great economic disadvantage, creating a case for instituting non-discrimination requirements for access to foundation models. In the US, public utility law prohibits undue or unreasonable price

discrimination, requiring that similar customers receiving similar services pay the same prices (Henderson and Burns 1989).[7]

When foundation models morph into platforms, for example through GPT stores, non-discrimination requirements would prevent foundation model companies from privileging their own downstream products and services over other products and services. For example, Khan (2017) argues that the use of the essential facilities doctrine – compelling a monopolist to provide easy access to competitors in an adjacent market – can be apt in such situations.

4. Regulatory concerns

In addition to antitrust concerns, the tendency of foundation models towards concentration blunts market forces and may make it desirable to impose a number of additional regulations:

Ensuring a level playing field with non-AI providers

As foundation models are deployed throughout the economy in a growing number of different functions, they are likely to compete with non-AI providers (including human providers) of different products and services. In those instances, ensuring a level playing field with non-AI providers will be essential.

The law is often silent on the liability that AI solutions carry when they engage in the real world, especially in sectors where these solutions are not already common. There is no reason why foundation models should be exempt from sectoral regulations including liability, professional licensing, and professional ethics guidelines. If these regulations are derived from various rights of consumers and citizens, they should apply in appropriate ways to foundation models as well.

For instance, governments worldwide have extensive regulations in the education sector, such as around student privacy, non-discrimination, and educational standards. AI solutions used for the classroom should explicitly be subject to the same regulatory standards. Sometimes this will require clarifications or amendments in regulation, because it is difficult to apply regulation in the same way to all technologies, or to technology and humans.

There are cases where human workers are penalized for discounting the analysis of an AI solution in their workplace, creating a lopsided liability burden. For instance, nurses in some US hospitals can disregard algorithmic assessment of a patient's diagnosis with doctor approval, but

[7] Section 205(b) of the Federal Power Act of 1920, for example, prohibits undue preference or prejudice to any person, as well as unreasonable differences in rates, charges, service and facilities. See https://www.ferc.gov/sites/default/files/2021-04/federal_power_act.pdf. In the EU, Article 10 of Directive 2002/19/EC gives the national regulatory authority power to impose access requirements on communications networks. See https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:108:0007:0020:EN:PDF.

face high risks for such disregard as they are penalized for overriding algorithms that turn out to be right. This may lead nurses to err on the side of caution and follow AI solutions even when they know they are wrong in a given instance (Bannon 2023).

To avoid situations where humans defer to AI against their better judgment, liability frameworks should be neutral to ensure that technology follows sectoral regulation and not the other way round. AI technology should not be applied in circumstances in which it does not meet regulatory standards.

Such requirements might be onerous and promote more concentration, and the burden to compensate for these effects must be borne by antitrust authorities. Moreover, such requirements would make the deployment of unsafe and substandard systems less financially attractive. It is also possible that enforcing sectoral standards promotes the development of AI systems that are more sector-specific than general, potentially leading to more competition. More importantly, we expect these requirements to protect people against the degradation of product and service standards, and the erosion of consumer rights, due to the use of AI.

Data governance

Data governance can aid in preventing undesirable concentration in foundation model markets but can also exacerbate monopoly power. Existing data protection law can institute important limits on the exclusive accumulation of data resources required to build large scale models. This may become more important as training data increasingly moves from freely available web data to proprietary datasets.

Purpose limitation laws require that data only be used for the purpose for which it was collected; a challenge to the use of proprietary data through this route is possible, given the multiple uses to which a foundation model can be put. Platforms that collect data might have to collect consent separately for the use of data through foundation models. Seen a little differently, purpose limitations can reduce the economies of scope of foundation models by limiting their legal use cases.

Many data governance regulations aim at restricting the transfer of data between firms, as exemplified by the EU's General Data Protection Regulation. Although this may be desirable for reasons such as privacy, it increases the market power of the firms with the most data, typically, the dominant firms. It also provides new incentives for vertical integration. As a solution, antitrust regulators could mandate that foundation model companies and cloud companies hold certain data in silos, i.e., not share it for other uses (including new foundation models) or with other companies within their corporate groups (Competition and Markets Authority 2020).

Data portability rights allow users to port their personal data from one platform to another at no or low cost. Some regulations and proposals also create a framework on mandatory data sharing

or interoperability to ensure a level playing field in data-driven markets. For instance the EU's Revised Payment Services Directive creates a data sharing and interoperability framework to level the playing field between banks and new payment operators. Some proposals recommend mandatory data sharing to break the monopoly in the online search market (Martens 2023).

Systemic risks from homogenisation

As foundation models are integrated into production and delivery processes for goods and services across many sectors of the economy, Bommasani et al. (2021) point out that they may give rise to systemic risks that arise from homogenisation. For example, assume that one foundation model in its fine-tuned versions is powering decision-making processes in search, market research, customer service, advertising, design, manufacturing, and many more. Widespread, cross-industrial applications mean that any errors, vulnerabilities or failures in a foundation model can threaten a significant fraction of economic activity, producing the risk of systemic economic effects that may warrant regulation.

Biases or errors in the outputs of a foundation model can be inherited by downstream models. These might not be noticed until the model is in use, and its effects might be observed throughout the economy. An example scenario is if an LLM understands demand patterns for consumer goods based on gender stereotypes, and consequently leads to over- or under-production of certain goods.

Foundation models could also be vulnerable to malicious attacks such as through data poisoning. Data poisoning refers to the practice of tampering with training data in order to influence the outputs of a model. These outputs can then be reflected in the outputs of fine-tuned models as well. Cybersecurity and data auditing methods at the pre-training level can protect against the risk of data poisoning.

Foundation models may exacerbate income inequality through automation that leads to centralization of intellectual functions. Automating tasks with foundation models can result in a single entity monopolizing entire skill sets if deployment is highly centralized, reducing competition. Meanwhile, unequal access to frontier foundation models internationally gives production advantages to countries where they are developed, disadvantaging others. This could worsen income inequality. Policies promoting decentralized access to foundation models and preventing monopolization of automated skills may mitigate these risks.

Safety considerations

Current foundation models are, for most practical purposes, quite safe to use and deploy. However, AI experts warn that future more capable AI systems could cause wide-ranging and perhaps even irreversible harm to society (Hinton et al 2023; Anderljung et al. 2023). These

concerns include the potential malfunctioning of powerful AI systems as well as the malicious use of such AI systems.

AI safety research is particularly concerned about the potential development of artificial general intelligence (AGI), which is the explicit goal of both OpenAI and Google DeepMind. AGI refers to technology that can perform all cognitive tasks that humans can perform, or at least all economically relevant tasks. Many problems could potentially arise from the further advancement of AI if it proceeds in a reckless and unsafe manner, including misuse by unscrupulous actors, catastrophes arising from errors in understanding human preferences, mass disruption of labor markets, or even disempowerment or extinction (Hendrycks et al. 2022). AI safety risk – the risk of unsafe frontier AI systems being developed and deployed – interacts with market structure and competition in several ways, which we explore in the following section.

The market structure of foundation models, whether concentrated or competitive, presents tradeoffs for AI safety efforts:

Safety risks from competition Competitive pressures in AI development may undermine safety efforts. Breakneck competition could accelerate unsafe foundation model advances before society installs safeguards. Competitive dynamics may also incentivize developers to cut corners on safety research in order to get ahead. However, competition has also encouraged open sourcing of some foundation models, which raises separate safety issues due to potential misuse.[8] Overall, unchecked competition risks rapidly advancing AI in dangerous directions without enough oversight.

Safety risks from market concentration Market concentration also poses AI safety risks. Monopolies with few competitors can amass resources more quickly to develop more powerful models, with greater potential for misuse. Concentration likewise may reduce incentives for safety precautions. However, a monopoly held by a responsible firm may also allow more controlled AI advancement. Ultimately, whether concentration or competition prevails, responsible development and strong oversight are essential to mitigate the safety hazards from increasingly capable AI systems.

Regulatory capture

A concentrated market for foundation models, combined with widespread application of foundation models, implies high financial stakes for foundation model companies. This makes it likely that both antitrust rules and the other regulations we discussed above will be subject to lobbying efforts and to growing risks of regulatory capture. While regulatory capture is typically

[8] The open-source release of Meta's Llama model has already enabled programmers to use Llama's published model weights to create novel and targeted adversarial attacks for LLMs that make it possible to circumvent the safety restrictions built into all other leading LLMs (e.g., ChatGPT, Bard, Claude), thereby seriously undermining those organizations' attempts to keep their AI systems safe (Zou et al. 2023).

viewed as the risk of firms co-opting regulators to make it more difficult for competitors to enter a market, it is also possible that lobbying ensures that certain externalities are *not* regulated, if this is in the interests of current market participants. The risk that the lobbying power of foundation model producers may rise over time is one reason why proactive regulation – before producers become too powerful to stave off regulations that they find undesirable – may be indicated.

5. Conclusions

Foundation models powered by deep learning techniques have demonstrated rapid advances in capability over the past decade. As these AI systems grow more powerful, their potential economic impact expands as well.

However, the market structure for developing and deploying these models tends toward concentration, given the economies of scale arising from large pre-training costs and bottlenecks in key inputs like data and compute. This raises competition concerns if too few players come to dominate the provision of foundation systems across industries, which policymaker should aim to counteract. It also poses new regulatory challenges, including for data governance, novel systemic risks, and safety.

If left unchecked, the trajectory of foundation models is toward one or a few dominant firms providing the AI substrate for major parts of the global economy. Preventing extreme concentration via proactive policies today is preferable to untangling monopolies after the fact. Proactive competition policies may also help to distribute the productivity gains from AI more equitably to avoid worsening inequality. With prudent competition and regulatory policies, we are hopeful that society will be able to steer this powerful new technology toward broadly shared prosperity.

Disclosures

Anton Korinek is a Professor of Economics at the University of Virginia and the Darden School of Business as well as a Nonresident Fellow at the Brookings Institution, a Research Associate at the NBER, a Research Fellow at the CEPR and the Economics of AI Lead at the Centre for the Governance of AI. Jai Vipra received financial support as a Winter Fellow from the Centre for the Governance of AI while writing this paper. The Centre for the Governance of AI is an independent research organization that is dedicated to helping humanity navigate the transition to a world with advanced AI. For further information, see https://www.governance.ai/about-us. The views expressed in this paper are the personal views of the authors and do not necessarily represent the views of the institutions they are affiliated with.

Bibliography

Altman, Sam. 2021. "Moore's Law for Everything." Blog post. https://moores.samaltman.com/

Anil, Rohan, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, et al. 2023. 'PaLM 2 Technical Report'. arXiv. https://doi.org/10.48550/arXiv.2305.10403.

'Announcing Google DeepMind'. 2023. Google DeepMind. 20 April 2023. https://www.deepmind.com/blog/announcing-google-deepmind.

'Anthropic Partners with Google Cloud'. 2023. Anthropic. 3 February 2023. https://www.anthropic.com/index/anthropic-partners-with-google-cloud.

Azizi, Shekoofeh, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. 2023. 'Synthetic Data from Diffusion Models Improves ImageNet Classification'. arXiv. https://doi.org/10.48550/arXiv.2304.08466.

Bannon, Lisa. 2023. 'When AI Overrules the Nurses Caring for You'. *The Wall Street Journal*, 15 June 2023. https://www.wsj.com/articles/ai-medical-diagnosis-nurses-f881b0fe.

Belfield, Haydn, and Shin-Shin Hua. 2022. 'Compute and Antitrust: Regulatory implications of the AI hardware supply chain, from chip design to cloud APIs'. *Verfassungsblog*, August. https://verfassungsblog.de/compute-and-antitrust/.

Boldrin, Michele, and David K. Levine. 2009. 'Does Intellectual Monopoly Help Innovation?' *Review of Law & Economics* 5 (3): 991–1024. https://doi.org/10.2202/1555-5879.1438.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2021. 'On the Opportunities and Risks of Foundation Models'. arXiv. https://doi.org/10.48550/arXiv.2108.07258.

Bresnahan, Timothy F., and Manuel Trajtenberg. 1995. 'General Purpose Technologies 'Engines of Growth'?' Journal of Econometrics 65(1), pp. 83-108.

Bresnahan, Timothy F., and Jonathan D. Levin. "Vertical Integration and Market Structure." In Handbook of Organizational Economics, edited by Robert Gibbons and John Roberts, 853-890. Princeton: Princeton University Press, 2012.

Competition and Markets Authority. 2020. 'Online Platforms and Digital Advertising Market Study: Final Report'. Competition and Markets Authority. https://www.gov.uk/cma-cases/online-platforms-and-digital-advertising-market-study.

———. 2023. 'AI Foundation Models: Initial Report'. Competition and Markets Authority. https://www.gov.uk/government/publications/ai-foundation-models-initial-report.

Corden, Jez. 2023. 'Despite Its Big OpenAI Push, Microsoft's Bing Search Market Share Decreases Year-over-Year'. *Windows Central*, 13 November 2023. https://www.windowscentral.com/microsoft/despite-its-big-openai-push-microsofts-bing-search-market-share-decreases-year-over-year.

Cottier, Ben. 2023. 'Trends in the Dollar Training Cost of Machine Learning Systems'. *Epoch* (blog). 31 January 2023. https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems.

Doctorow, Cory. 'Social Quitting'. Medium (blog), 17 November 2022. https://doctorow.medium.com/social-quitting-1ce85b67b456.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. 'An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale'. arXiv. https://doi.org/10.48550/arXiv.2010.11929.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." arXiv:2303.10130.

Epoch. 2023. 'Data Trends'. Epoch. 11 April 2023. https://epochai.org/trends.

Epoch, 2024. 'Parameter, Compute and Data Trends in Machine Learning'. Retrieved 15 Feb 2024. https://epochai.org/data/epochdb/visualization

FTC. 2023. 'FTC and DOJ Charge Amazon with Violating Children's Privacy Law.' Federal Trade Commission. 31 May 2023. https://www.ftc.gov/news-events/news/press-releases/2023/05/ftc-doj-charge-amazon-violating-childrens-privacy-law-keeping-kids-alexa-voice-recordings-forever

FTC. 2024. 'FTC Launches Inquiry into Generative AI Investments and Partnerships'. Federal Trade Commission. 24 January 2024. https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships.

Fernandez, Joaquin, Knud Lasse Lueth, and Philipp Wegner. 2023. 'Generative AI Market Report 2023–2030'. IoT Analytics. 14 December 2023. https://iot-analytics.com/product/generative-ai-market-report-2023-2030/.

Fletcher, Richard. 2024. How many news websites block AI crawlers? Blog post, Reuters Institute for the Study of Journalism, Oxford University.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, et al. 2023. 'Gemini: A Family of Highly Capable Multimodal Models'. arXiv. https://doi.org/10.48550/arXiv.2312.11805.

Hagey, Keach, and Asa Fitch. 'Sam Altman Seeks Trillions of Dollars to Reshape Business of Chips and AI'. Wall Street Journal, 8 February 2024, sec. Tech. https://www.wsj.com/tech/ai/sam-altman-seeks-trillions-of-dollars-to-reshape-business-of-chips-and-ai-89ab3db0.

Hatzius, Jan, Joseph Briggs, Devesh Kodnani, and Giovanni Pierdomenico. 2023. 'The Potentially Large Effects of Artificial Intelligence on Economic Growth'. Goldman Sachs Economic Research.

Heath, Alex. 2024. 'Mark Zuckerberg's New Goal Is Creating Artificial General Intelligence'. *The Verge*, 18 January 2024. https://www.theverge.com/2024/1/18/24042354/mark-zuckerberg-meta-agi-reorg-interview.

Heim, Lennart. 2021. 'Transformative AI and Compute - EA Forum'. 23 September 2021. https://forum.effectivealtruism.org/s/4yLbeJ33fYrwnfDev.

———. 2023. 'This Can't Go on(?) - AI Training Compute Costs'. *Blog.Heim.Xyz* (blog). 1 June 2023. https://blog.heim.xyz/this-cant-go-on-compute-training-costs/.

Hemphill, C. Scott, and Tim Wu. 'Nascent competitors.' University of Pennsylvania Law Review 168, no. 7 (June 2020): 1879-1910.

Henderson, J. Stephen, and Robert E. Burns. 1989. 'An Economic and Legal Analysis of Undue Price Discrimination'. The National Regulatory Research Institute.

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 'Measuring Massive Multitask Language Understanding'. arXiv, September 2020. https://doi.org/10.48550/arXiv.2009.03300.

Hinton, Geoffrey et al. 2023. "Statement on AI Risk." Center for AI Safety (CAIS), accessed February 15, 2024, https://www.safe.ai/statement-on-ai-risk.

Hjelm, Max. n.d. 'CoreWeave Becomes NVIDIA's First Elite Cloud Services Provider for Compute — CoreWeave'. Accessed 30 June 2023. https://www.coreweave.com/blog/coreweave-becomes-nvidias-first-elite-cloud-services-provider-for-compute.

Hoffmann, Jordan; Borgeaud, Sebastian; Mensch, Arthur; Buchatskaya, Elena; Cai, Trevor; Rutherford, Eliza; Casas, Diego de Las; Hendricks, Lisa Anne; Welbl, Johannes; Clark,

Aidan; Hennigan, Tom; Noland, Eric; Millican, Katie; Driessche, George van den; Damoc, Bogdan. 2022. "Training Compute-Optimal Large Language Models". arXiv:2203.15556.

Holmes, Aaron. 2023. 'How Microsoft Swallowed Its Pride to Make a Massive Bet on OpenAI'. The Information. 25 May 2023. https://www.theinformation.com/articles/how-microsoft-swallowed-its-pride-to-make-a-massive-bet-on-openai.

Holmes, Aaron, and Anissa Gardizy. 2023. 'AI Developers Stymied by Server Shortage at AWS, Microsoft, Google'. *The Information*, 7 April 2023. https://www.theinformation.com/articles/ai-developers-stymied-by-server-shortage-at-aws-microsoft-google.

Hui, Xiang, Oren Reshef, and Luofeng Zhou. 'The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market'. SSRN Scholarly Paper. Rochester, NY, 31 July 2023. https://doi.org/10.2139/ssrn.4527336.

Hwang, Tim. 2018. 'Computational Power and the Social Impact of Artificial Intelligence'. SSRN Working Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3147971.

JP Morgan Research. 'Is Generative AI a Game Changer?', 14 February 2024. https://www.jpmorgan.com/insights/global-research/artificial-intelligence/generative-ai.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. 'Scaling Laws for Neural Language Models'. arXiv. https://doi.org/10.48550/arXiv.2001.08361.

Kerin, Roger A., P. Rajan Varadarajan, and Robert A. Peterson. 1992. 'First-Mover Advantage: A Synthesis, Conceptual Framework, and Research Propositions'. *Journal of Marketing* 56 (4): 33–52. https://doi.org/10.1177/002224299205600404.

Khan, Lina M. 2017. "Amazon's Antitrust Paradox." *The Yale Law Journal* 126 (3): 710–805.

Khan, Saif M., Dahlia Peterson, and Alexander Mann. 2021. 'The Semiconductor Supply Chain'. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/the-semiconductor-supply-chain/.

Korinek, Anton. 2023. Scenario Planning for an A(G)I Future, *IMF Finance & Development Magazine* 60(4), pp. 30-33.

Knight, Will.. 2024. 'Amazon's Cloud Boss Likens Generative AI Hype to the Dotcom Bubble'. *Wired*, 7 February 2024. https://www.wired.com/story/amazons-cloud-boss-selipsky-generative-ai-hype/.

Li, Wendy Y. 2023. 'Regulatory Capture's Third Face of Power'. *Socio-Economic Review* 21 (2): 1217–45. https://doi.org/10.1093/ser/mwad002.

Lieberman, Marvin B., and David B. Montgomery. 1988. 'First-Mover Advantages'. *Strategic Management Journal* 9 (S1): 41–58. https://doi.org/10.1002/smj.4250090706.

Lomas, Natasha. 'Big Tech AI Infrastructure Tie-Ups Set for Deeper Scrutiny, Says EU Antitrust Chief'. TechCrunch, 20 February 2024. https://techcrunch.com/2024/02/20/eu-merger-control-ai/.

———. 'EU Checking If Microsoft's OpenAI Investment Falls under Merger Rules'. TechCrunch, 9 January 2024. https://techcrunch.com/2024/01/09/openai-microsoft-eu-merger-rules/.

Martens, Bertin. 'What Should Be Done about Google's Quasi-Monopoly in Search? Mandatory Data Sharing versus AI-Driven Technological Competition', 6 July 2023. https://www.bruegel.org/working-paper/what-should-be-done-about-googles-quasi-monopoly-search-mandatory-data-sharing-versus.

Mattioli, Dana. 2020. 'Amazon Scooped Up Data From Its Own Sellers to Launch Competing Products'. *Wall Street Journal*, 23 April 2020, sec. Tech. https://www.wsj.com/articles/amazon-scooped-up-data-from-its-own-sellers-to-launch-com

peting-products-11587650015.
Norem, Josh. 'Analysts Estimate Nvidia Owns 98% of the Data Center GPU Market'. ExtremeTech (blog), 1 February 2024. https://www.extremetech.com/computing/analysts-estimate-nvidia-owns-98-of-the-data-center-gpu-market.
Olson, Parmy. 2024a. 'Google, Microsoft Will Dominate AI as Computing Costs Surge'. Bloomberg.Com, 19 February 2024. https://www.bloomberg.com/opinion/articles/2024-02-19/artificial-intelligence-microsoft-google-nvidia-win-as-computing-costs-surge.
———. 2024b. 'Zuckerberg's Secret Weapon for AI Is Your Facebook Data'. Bloomberg.Com, 6 February 2024. https://www.bloomberg.com/opinion/articles/2024-02-06/zuckerberg-s-plan-for-ai-hinges-on-your-facebook-and-instagram-data.
OpenAI. 2023a. 'Our Structure'. 2023. https://openai.com/our-structure.
———. 2023b. 'ChatGPT Plugins'. OpenAI. 23 March 2023. https://openai.com/blog/chatgpt-plugins.
Patel, Dylan, and Daniel Nishaball. 'Inference Race To The Bottom - Make It Up On Volume?' SemiAnalysis (blog), 19 December 2023. https://www.semianalysis.com/p/inference-race-to-the-bottom-make.
Reid, Elizabeth. 2023. 'Supercharging Search with Generative AI'. *Google* (blog). 10 May 2023. https://blog.google/products/search/generative-ai-search/.
Salesforce. 2023. 'Top Generative AI Statistics for 2024'. Salesforce. 1 September 2023. https://www.salesforce.com/news/stories/generative-ai-statistics/.
Shani, Inbal. 2023. 'Survey Reveals AI's Impact on the Developer Experience'. *The GitHub Blog* (blog). 13 June 2023. https://github.blog/2023-06-13-survey-reveals-ais-impact-on-the-developer-experience/.
Shu, Catherine. 2014. 'Google Acquires Artificial Intelligence Startup DeepMind For More Than $500M'. *TechCrunch*, 27 January 2014. https://techcrunch.com/2014/01/26/google-deepmind/.
Staff in the Bureau of Competition & Office of Technology. 2023. 'Generative AI Raises Competition Concerns'. *Federal Trade Commission* (blog). 29 June 2023. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns.
Suleyman, Mustafa. The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma. First edition. New York: Crown, 2023.
The Economist. 2023. 'Large, Creative AI Models Will Transform Lives and Labour Markets', 22 April 2023. https://www.economist.com/interactive/science-and-technology/2023/04/22/large-creative-ai-models-will-transform-how-we-live-and-work.
Tiku, Nitasha, and Gerrit De Vynck. 2023. 'Google Shared AI Knowledge with the World — until ChatGPT Caught Up'. *The Washington Post*, 4 May 2023. https://www.washingtonpost.com/technology/2023/05/04/google-ai-stop-sharing-research/.
Tirole, Jean. 1988. The Theory of Industrial Organization. Cambridge, MA: MIT Press.
Trammell, Philip, and Anton Korinek. 'Economic Growth under Transformative AI'. Working Paper. Working Paper Series. National Bureau of Economic Research, October 2023. https://doi.org/10.3386/w31815.
Varadarajan, Rajan, Manjit S. Yadav, and Venkatesh Shankar. 2008. 'First-Mover Advantage in an

Internet-Enabled Market Environment: Conceptual Framework and Propositions'. *Journal of the Academy of Marketing Science* 36 (3): 293–308. https://doi.org/10.1007/s11747-007-0080-y.

Verhulst, Stefaan G. 'Are We Entering a Data Winter? On the Urgent Need to Preserve Data Access for the Public Interest'. Frontiers Policy Labs (blog), 2024. https://policylabs.frontiersin.org/content/commentary-are-we-entering-a-data-winter.

Victor, Jon. 2023. 'Why YouTube Could Give Google an Edge in AI'. *The Information*, 14 June 2023. https://www.theinformation.com/articles/why-youtube-could-give-google-an-edge-in-ai.

Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. 'Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning'. arXiv. https://doi.org/10.48550/arXiv.2211.04325.

Vipra, Jai, and Sarah Myers West. 'Computational Power and AI'. AI Now Institute, 27 September 2023. https://ainowinstitute.org/publication/policy/compute-and-ai.

Warren, Tom. 2023. 'Microsoft Extends OpenAI Partnership in a "Multibillion Dollar Investment"'. *The Verge*, 23 January 2023. https://www.theverge.com/2023/1/23/23567448/microsoft-openai-partnership-extension-ai.

Zhou, Viola. 2023. 'AI Is Already Taking Video Game Illustrators' Jobs in China'. *Rest of World*, 11 April 2023. https://restofworld.org/2023/ai-image-china-video-game-layoffs/.