NBER WORKING PAPER SERIES

CONCENTRATING INTELLIGENCE:
SCALING AND MARKET STRUCTURE IN ARTIFICIAL INTELLIGENCE

Anton Korinek
Jai Vipra

Working Paper 33139
http://www.nber.org/papers/w33139

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2024

This is a revised version of a paper presented at the 79th Economic Policy panel meeting in Brussels in April 2024. We thank Susan Athey, Emma Bluemke, Claire Dennis, Avi Goldfarb, Aidan Kane, Pia Malaney, Sarah Myers West, Sanjay Patnaik, Nicholas Ritter, Max Schnidman, Eli Schrag, Rob Seamans and Joseph Stiglitz as well as our editors, Emilio Calvano and Giacomo Calzolari, our discussant Doh-Shin Jeon, and two anonymous referees for thoughtful comments and conversations. Any remaining errors are our own. Korinek gratefully acknowledges financial support from the Center for Innovation, Growth and Society of the Institute for New Economic Thinking (INET-CIGS, grant no INO21-00004). Jai Vipra gratefully acknowledges financial support as a Winter Fellow from the Centre for the Governance of AI. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research or of the other institutions that the authors are affiliated with. An earlier version of this paper was circulated under the title "Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT."

Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence
Anton Korinek and Jai Vipra
NBER Working Paper No. 33139
November 2024
JEL No. D43, K21, L4, L86, O33

ABSTRACT

This paper examines the evolving structure and competition dynamics of the rapidly growing market for foundation models, with a focus on large language models (LLMs). We describe the technological characteristics that shape the AI industry and have given rise to fierce competition among the leading players. The paper analyzes the cost structure of foundation models, emphasizing the importance of key inputs such as computational resources, data, and talent, and identifies significant economies of scale and scope that may create a tendency towards greater market concentration in the future. We explore two concerns for competition, the risk of market tipping and the implications of vertical integration, and we evaluate policy remedies that aim to maintain a competitive landscape. Looking ahead to increasingly transformative AI systems, we discuss how market concentration could translate into unprecedented accumulation of power, highlighting the broader societal stakes of competition policy.

Anton Korinek
Department of Economics
University of Virginia
Monroe Hall 246
248 McCormick Rd
Charlottesville, VA 22904
and NBER
anton@korinek.com

Jai Vipra
Department of Science and
Technology Studies
Cornell University
415 Morrill Hall
Ithaca, NY 14853
jv474@cornell.edu

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. 2024). 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.

Recent advances in AI have derived from the growing use of so-called foundation models – large AI models that use deep learning methods, are pre-trained on vast amounts of data, and can then be fine-tuned and adapted to specific economic tasks and applications (Bommasani et al. 2021). The most visible category of foundation models have been large language models (LLMs) such as OpenAI’s GPT-4 and Google DeepMind’s Gemini, which can process and generate text on any topic, and which have been at the center of the generative AI revolution. LLMs 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, and so on. In the domain of economics, Korinek (2023a, 2024) demonstrates how LLMs can enhance research productivity, from ideation and background research to writing and data analysis. Moreover, there is already evidence that generative AI is disrupting certain categories of work (Hui et al 2023), and a survey of US business leaders found that nearly half had replaced workers with ChatGPT (Shani 2023). Still, the current generation of LLMs clearly also has its limitations.

At the time of writing in November 2024, the market for frontier foundation models was very dynamic, in a state of flux, and characterized by fierce competition. This is in stark contrast to when we circulated the first draft of this paper in September 2023, in which we described a market in which a single model, GPT-4, had a significant technological lead, and its producer, OpenAI, accordingly commanded the lion’s share of the market (Vipra and Korinek, 2023). A little over a year later, 16 different AI labs have produced models that surpass the capabilities of the original GPT-4, according to the popular LMSYS leaderboard benchmark (Chiang et al., 2024), and several of these models have been distributed in an open source fashion, allowing anyone in the world to freely download and operate them. The technological landscape may change quickly – for example, if OpenAI or another lab makes a significant breakthrough and releases a foundation model that exhibits capabilities that put them far ahead of the competition. However, when it comes to user numbers and monetization, OpenAI currently has a significant lead, given its first-mover status.

The size of the market for generative AI was estimated to be just $3bn in 2023 (Fernandez et al. 2023) – a tiny sliver of the $100tn world economy. In 2022, the term "generative AI" had not even been coined. Yet the large number of entrants into the market suggests that many actors are willing to make significant investments because they expect large technological advances in the future, with commensurately large monetary rewards down the road. This is consistent with projections by leading investment banks and consultancies that the technology may underpin 7 to 10 percent of global GDP within a decade, implying the potential for significant growth (Hatzius et al. 2023; JP Morgan Research 2024). With sufficient technological advances, optimists believe that the growth effects of AI could be far higher still (see, e.g., Trammell and Korinek, 2023). Under that view, the future rewards may be astronomic, and it is understandable that companies are racing to pursue incumbency.

Given that foundation models may play such an important role in the economy of the future, competition authorities around the world are playing close attention to the dynamics of the market (see, e.g., Vestager, 2024). Moreover, there are technological and economic forces that provide strong reasons to be concerned about competition in the market for AI, as we examine in this paper.

For balance, let us observe that there is also a camp that sees the future economic benefits of AI in a much dimmer light. A recent Goldman Sachs report was entitled "GenAI: Too much spend, too little benefit?" (Goldman Sachs, 2024). In a nutshell, this pessimistic camp believes that AI will not significantly advance beyond current capabilities and will have only minor economic impacts. For example, Acemoglu (2024) estimated that the growth effects of generative AI will be only 0.07% per year over the coming decade. In that case, the market concentration implications will be a footnote in economic history. Investors are currently attempting to determine which view is more appropriate.

Meanwhile, the stated mission of several leading AI labs, including OpenAI, is even grander: that a future version of their foundation models will achieve artificial general intelligence (AGI), defined as the ability to perform any cognitive task that humans can perform. If this mission is achieved, then their models could underpin any cognitive work and, if equipped with the necessary hardware, any 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 speculative, 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, 2023b; Korinek and Suh, 2024). In such a scenario, the market for foundation models would be the entire economy.

Moreover, the structure of the market for AI systems may also carry important implications for power dynamics in this scenario, as market concentration would likely translate into an unprecedented accumulation of power by the entities controlling AGI systems. This power would extend far beyond traditional economic domains, affecting the social and political landscape globally. Proactive measures aimed at preventing excessive market concentration in the AI sector could then serve a dual purpose: maintaining economic competitiveness and acting as a vital check against excessive power consolidation, which could help ensure that this transformative technology's potential is better aligned and harnessed for broader societal benefit.

The remainder of this paper is structured as follows. Section 2 provides a snapshot of the current market for generative AI, including the principal players and recent attention from antitrust authorities. Section 3 analyzes the technological characteristics and market structure of foundation models, examining the cost structure and the importance of key inputs such as compute, data, and talent. Section 4 delves into concerns regarding market concentration, discussing the risks of market tipping and vertical integration, as well as the importance of ensuring a level playing field with non-AI providers and addressing safety considerations. This section also reviews the regulatory frameworks in the EU. Finally, Section 5 concludes.

2. A snapshot of the market for generative AI

This section describes the dynamics and fierce level of competition in the market for generative AI as well as the characteristics of the major players at the time of writing. We start by examining how the capabilities of the most powerful LLMs compare to each other and then evaluate how this shapes the evolution of market shares.

Fierce competition among the principal players

Table 1 lists a snapshot of the top four AI labs together with their most highly ranked publicly available LLM as of November 4, 2024, ranked by their LMSYS score, which constitutes a widely used benchmark for model performance. LMSYS (Language Model System) is a rating system for LLMs that is based on the Elo rating system originally developed for chess (Chiang et al., 2024). It evaluates models by letting them compete against each other in responding to real user requests, with human users rating which model produced the higher-quality response. The models' scores are continually adjusted to reflect the models' relative probabilities of winning.[2]

As the table illustrates, the leading systems are clustered quite closely together in their LMSYS scores, indicating that they are very similar in their capabilities and, from an economic perspective, close substitutes. For instance, the two top-ranked systems, OpenAI's ChatGPT-4o and Google DeepMind's Gemini 1.5 Pro 002, differ by only 37 points on the LMSYS scale, which translates to ChatGPT-4o having approximately a 55.3% chance of outperforming Gemini in a given user request. Even in competition with the fourth-ranked model in the table, Claude 3.5 Sonnet (New), ChatGPT-4o has only a 57.7% chance of winning. Other ranking systems, such as the MMLU benchmark developed by Hendrycks et al. (2020) that measures the world knowledge

[2] The LMSYS probability formula, based on the Elo system, is $P(\text{A beats B}) = 1 / (1 + 10\text{^}(-\Delta/400))$, where Δ is the difference in LMSYS scores between the two competing models A and B.

and problem solving ability of LLMs on scientific subjects, show a similar pattern of bunching among the top-ranked models.

Lab	Country	Top Model	Released	LMSYS
OpenAI	USA	ChatGPT-4o-latest	2024-09-03	1340
Google DeepMind	USA/UK	Gemini-1.5-Pro-002	2024-09-24	1303
xAI	USA	Grok-2	2024-08-13	1290
Anthropic	USA	Claude 3.5 Sonnet (New)	2024-10-22	1286

Table 1: Top foundation model providers according to LMSYS arena score of best model
Source: LMSYS leaderboard at https://lmarena.ai/?leaderboard. Accessed on Nov 4, 2024.

Another notable reflection of the fierce competition in the sector is that every single one of the LLMs in the table has been released or updated within the past three months. Given the fast pace of advances in AI and the level of competition, the leading AI labs have made it a habit to regularly release new versions of their models in an attempt to outdo each other. For example, the market leader, OpenAI, has released seven model updates since the first version of GPT-4 was made public in March 2023. These updates have significantly improved the quality of the model's responses (the current LMSYS score of the original GPT-4 is only 1186), increased its speed more than 3-fold, enhanced the amount of text that it can process 16-fold, while reducing the cost of generating a given amount of output by 92% – and all this within less than 20 months. Perhaps not coincidentally, OpenAI has repeatedly released updates shortly after another model displaced them from the top spot of the LMSYS leaderboard to regain its pole position – for example, when Google DeepMind displaced a previous version of GPT-4o in early August 2024. Many observers note that the prices charged by the leading AI labs barely allow them to cover their variable costs (Knight 2024) – the competition dynamics seem to be close to Bertrand competition.[3]

Leading foundation model providers

Some background on the leading players is useful to understand their motivations and strategies. OpenAI was the first player betting on the success of large language models. It has been the leader in the field since it released GPT-1 in 2018. OpenAI was founded in 2015, with the stated mission "to ensure that artificial general intelligence benefits all of humanity." In 2019, OpenAI founded a for-profit subsidiary to attract funding for the growing cost of computational resources necessary to train cutting-edge AI models. Its main investor is Microsoft, which has provided

[3] 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.

close to $14bn of funding to OpenAI so far. The investments in the for-profit subsidiary were structured such that the non-profit will receive all profits once outside investors have been repaid an agreed multiple of their initial investment.[4] This structure suggests that OpenAI is not acting solely as a profit-maximizing entity, although it needs to generate sufficient returns to fund its expenditure on compute. This tension between "serving humanity" and generating profits is visible in a number of actions by OpenAI, including some boardroom turmoil experienced in November 2023. After closing a $6.6bn funding round at a $157bn valuation in October 2024, OpenAI committed to restructure itself into a public benefit corporation (PBC), in which the non-profit would only hold a minority stake.

The second player in the market is Google DeepMind, owned by Google parent Alphabet, with its Gemini series of models. 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). Google Brain researchers developed the Transformer architecture upon which today's LLMs are based, while DeepMind developed seminal AI models such as AlphaGo and AlphaFold and has been the leader in advanced AI development in other areas over the past decade. Still, both entities were caught by surprise about the rapid rise of LLMs and have only recently made it to the top of the LLM league. DeepMind also has a wide-ranging portfolio of other cutting-edge AI capabilities, which may prove useful as it competes with OpenAI to develop LLMs that have better reasoning capabilities. For example, DeepMind AlphaProof and AlphaGeometry (2024) recently performed at the level of Silver Medalists at the International Math Olympiad.

xAI was founded in March 2023 by Elon Musk, who originally was a co-founder of OpenAI but left their board in 2018. The rapid ascent of xAI's Grok-2 into the top-3 within 17 months of the lab's founding illustrates how contestable the market for frontier AI systems is for someone willing to spend the requisite amounts. xAI also benefits from its close relationship with X, formerly Twitter, which Elon Musk took over in 2022.

Fourth-ranked Anthropic was founded by ex-OpenAI employees who disagreed with the increasingly market-oriented direction taken by OpenAI. It was originally created as a public benefit corporation in 2021 (PBC) with a similar mission as OpenAI. It briefly held the top spot on the LMSYS leaderboard in March 2024. All four proprietary LLM providers described so far offer their models via chatbots that charge their users subscription fees. There is intense competition by other LLM providers that rank very closely to the ones listed in Table 4, including by the Chinese labs 01 AI, Zhipu AI, and Qwen.

Meta, formerly Facebook, also ranks among the top-10 model providers, with an LMSYS score of 1267 for its Llama 3.1 model, but pursues a different business strategy from labs offering

[4] 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).

proprietary models: it distributes the Llama series of models on an open-source basis, implying that anyone can freely download the model, fine-tune or otherwise modify it, and operate it on their own computers.[5] Meta's strategy was driven by the two main factors. First, Meta's main business depends on advertising traffic on their social media platforms rather than selling access to LLMs so giving away Llama for free did not undercut their main source of revenue; if anything, it helped Meta's content creators. Second, despite employing AI godfather Yann Lecun as its chief AI scientist, Meta was a late-comer to LLMs and at first not able to play in the tier of frontier AI models listed in Table 1. When it created Llama 1.0 – clearly below the frontier at its release in February 2023 – the only way to be a relevant player in the market was to offer the model open source.

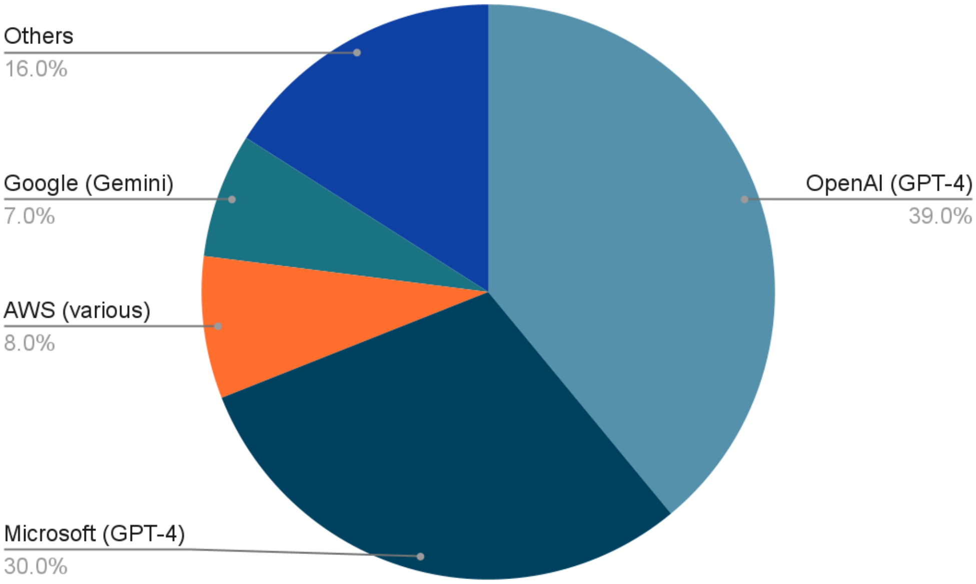

Figure 1: Principal players in the market for generative AI models (% of total spending)
Source: Fernandez et al (2023); 'Others' includes Anthropic, AI2I Labs, Cohere, Aleph Alpha, Hugging Face, Alibaba, IBM, and Baidu among others.

Economically speaking, competition had pushed the price of the model all the way to zero. This strategy successfully created a following for Llama. 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 may at some point also carry safety risks, which we will discuss further below. Meta is

[5] One limitation of Meta's license is that it only applies to companies that have fewer than 700 million monthly users - a provision meant to prevent other social networks competing with Meta from benefiting.

reportedly also stockpiling hundreds of thousands of GPUs to train the next generation of AI models (Heath 2024), suggesting that it will continue to play a role among the frontier models.

All in all, the LMSYS leaderboard lists more than a hundred notable LLMs that have been released since early 2023, reflecting the dynamic nature of the market. Some of them compete on capabilities like the ones listed in the table; some of them differentiate themselves by their smaller size, designed to run on laptops or cell phones.

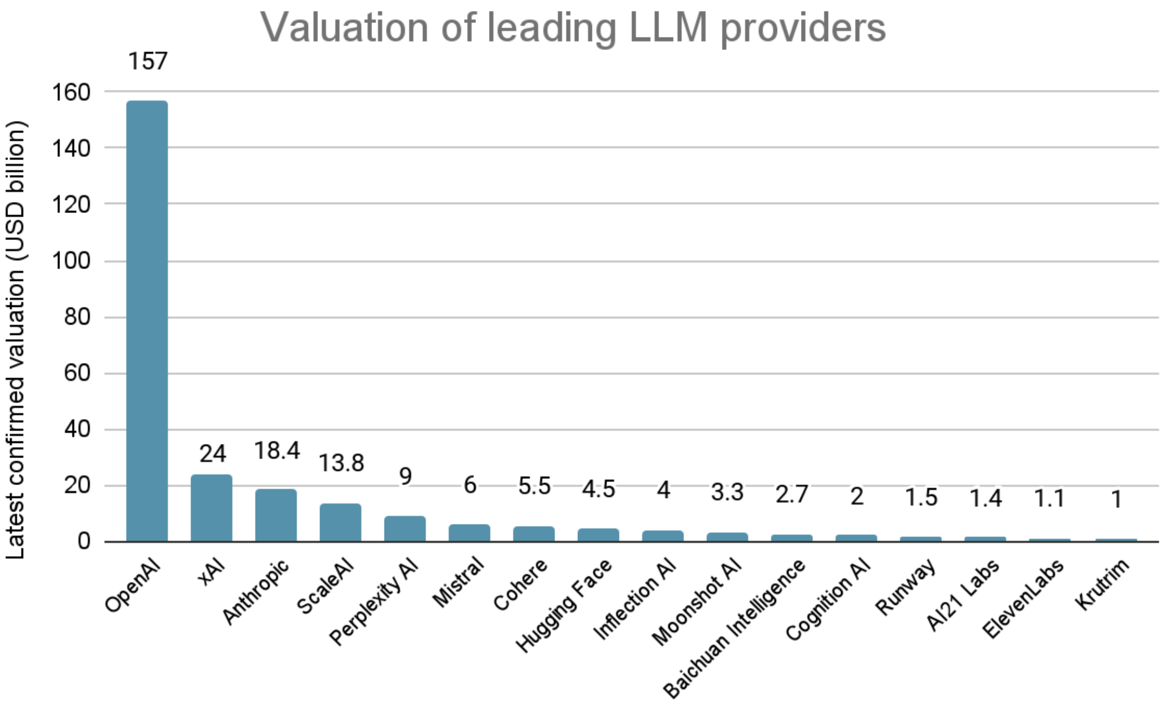

Figure 2: Most recent valuation of leading generative AI labs as of October 2024.
Source: Data collected by authors.

Market shares Figure 1 displays the market share of different generative AI providers at the end of 2023. Given its first-mover advantage, OpenAI's GPT-4 series was the clear market leader, powering both OpenAI's offerings such as ChatGPT (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. Google DeepMind was a distant second with only 7% market share. All others, including Anthropic's Claude, are included in "Others" or "AWS" if accessed via Amazon web services. Note that Meta's Llama does not earn revenue since it's open source.

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.

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. Notably, OpenAI insisted that its investors – including NVIDIA, the chipmaker critical for modern AI systems – agree to an exclusivity clause whereby they are not allowed to invest in OpenAI's rivals in its latest funding round in October 2024 (Hammond and Morris 2024). 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.

Company	Investments from large technology companies
OpenAI	Microsoft, Nvidia
Anthropic	Alphabet, Amazon, Salesforce, Zoom
Scale AI	Amazon, Meta, Nvidia, Intel, AMD
Perplexity	Nvidia, Amazon
Inflection AI	Microsoft, Nvidia
Hugging Face	Alphabet, AMD, Amazon, IBM, Intel, NVIDIA, Qualcomm, Salesforce
Mistral	Microsoft, Nvidia, Salesforce, Samsung, IBM
Baichuan	Alibaba, Tencent, Xiaomi
Moonshot	Alibaba, Tencent

Table 2: Investments in leading AI labs by large technology companies
Source: Collected by authors (does not include indirect funding through VCs)

Over the course of 2024, there have been multiple "acquisitions-by-proxy" of smaller AI firms listed in Figure 2 and Table 2: In March 2024, Inflection AI's co-founders and a majority of its staff were hired by Microsoft in exchange for a $650m licensing deal with the startup. This is expected to lower its valuation and has drawn attention from antitrust authorities since it may represent a take-over in disguise (Devin 2024). Similar deals were conducted between Amazon and the startup Adept as well as by Google DeepMind and Character.ai.

Market for compute

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. 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). 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. Vipra and Myers West (2023) provide a rigorous description of the market for compute.

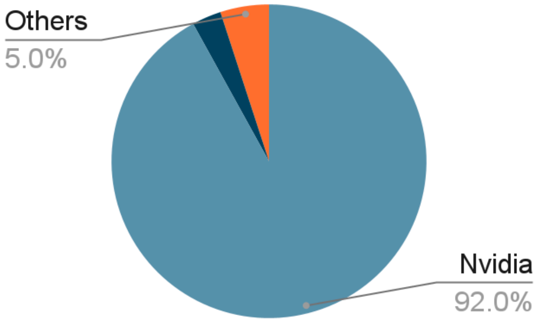

Figure 3: Market share of leading vendors in the market for GPUs
Source: Created by authors based on data from Fernandez et al (2023)

Recent attention from antitrust authorities

Although still only a tiny fraction of the overall economy, foundation models have already garnered significant attention from 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 US Department of Justice (DoJ) is also investigating Nvidia's acquisition of Run:ai, a startup that tries to optimize the use of GPUs (Palma 2024). The DoJ is reportedly concerned about allocative decisions made by Nvidia as well as about its software platform CUDA.

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). In a 2024 update, the CMA outlined three key risks to competition in this market: that firms controlling critical inputs may risk access to them, that incumbents in business or consumer-facing markets could distort the foundation models market, and that partnerships among key players could extend market dominance (Competition and Markets Authority 2024).

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)

In July 2024, the FTC, the DoJ, the CMA and the Competition Commissioner for the EU took the rare step of issuing a joint statement on AI competition issues, recognising AI as a technological inflection point and raising concerns about concentrated control of inputs, the extension of market power in AI-related markets, and potentially anticompetitive inter-firm partnerships (Vestager et al. 2024).

Some argue that competition authorities were not sufficiently proactive when the current generation of digital platforms formed – including companies such as Alphabet, Amazon, Apple, or Facebook (see e.g., Stigler Committee on Digital Platforms, 2019). The described actions suggest that competition authorities seem keen not to let these developments repeat themselves in the space of AI. However, this needs to be weighed against the risk of fighting the previous war. An appropriate policy response requires clarity and careful analysis of the similarities and differences between existing digital platforms and AI foundation models. This is what our paper aims to contribute to.

3. Technological characteristics and market structure

This section describes 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 the main inputs to production such as computational resources, talent, and data.

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 fine-tuned, i.e., adapted for specific use cases; and (iii) low variable costs of operating the model. Together, these characteristics generate significant economies of scale.

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”). Epoch (2024) estimates the compute cost of training Gemini Ultra in 2023 at $130 million. We will describe the drivers behind this rapid growth in spending 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 a company aims to employ a foundation model in its business area, it usually already possesses purpose-relevant data. For instance, if a healthcare provider aims to integrate GPT-4 in its operations, it 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 $1 per 100,000 syllables (“tokens”) of text generated using its GPT-4o 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.

In the following, we analyze the main inputs in producing foundation models – compute, data, and talent – in turn.

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 need 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,

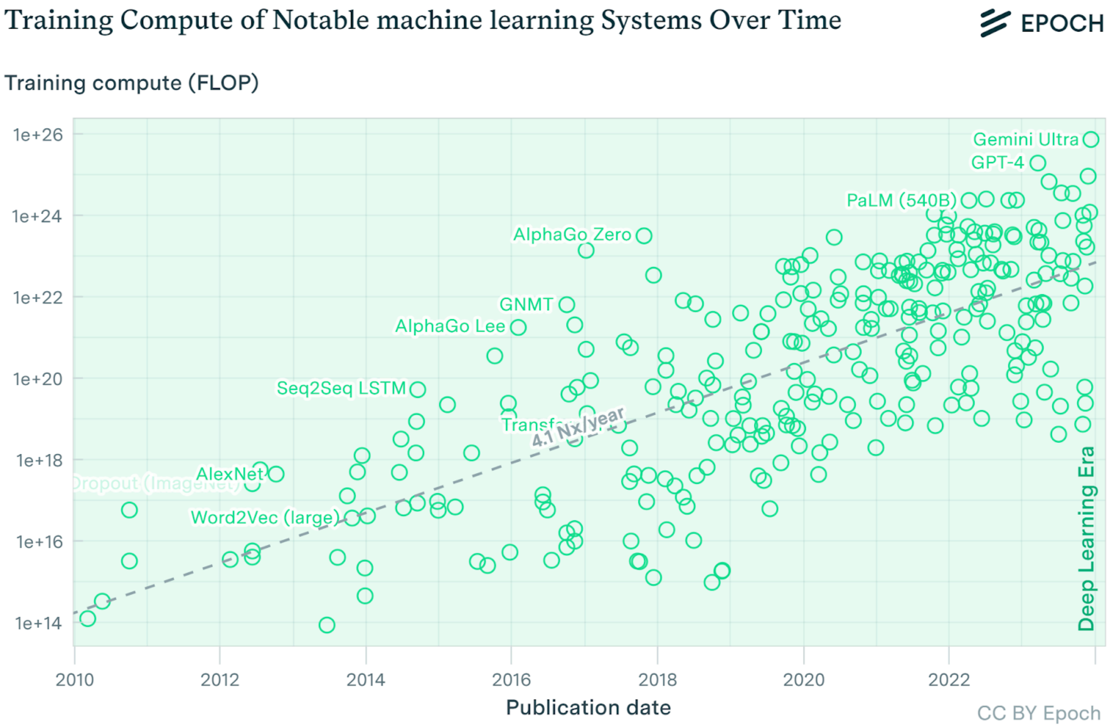

Figure 4: Training compute of notable AI systems. Copyright © by Epoch.org, reproduced under a CC-BY-4.0 license, 2024.

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.[6]

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.

[6] 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).

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).

Technology leaders such as Suleyman (2023) and other 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. Simple extrapolation of these trends may set us on a path to reach trillion-dollar foundation models by the end of the decade.

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.

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 in the coming years. Governments across the globe have taken notice of this situation and of the strategic importance of securing a domestic chip supply. In the US, the CHIPS and Science Act of 2022 aims to bolster the US semiconductor industry by allocating more than $50 billion to domestic semiconductor research and manufacturing. The Act aims to strengthen American supply chain resilience and maintain US technological leadership in AI. It also includes an initiative to explore AI's role in improving semiconductor manufacturing processes and the potential for export controls on advanced AI.

The European Chips Act of 2023 is central to the EU's semiconductor strategy, allocating €43 billion to developing more semiconductor plants and increasing Europe's global market share in the sector. The initiative also includes funding for research and development, workforce training, and fostering partnerships between government, industry, and academia. Just like the US, the EU is focused on strengthening its semiconductor supply chain and reducing dependency on foreign suppliers.

China has implemented a multi-pronged approach to boost its semiconductor capabilities in response to US export controls. This includes massive state-backed investments, with the latest fund valued at $47.5 billion, rapid expansion of manufacturing capacity, and accelerated development of domestic design capabilities. China is also focusing on building up its upstream supply chain, including chipmaking equipment and materials. Despite making significant progress, China still faces challenges in closing the technology gap with global leaders and reducing dependence on foreign inputs (UC IGRC, 2023).

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 that 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.

Moreover, since the beginning of the generative AI era in 2023, a growing number of content providers on the internet have restricted access to their data to the bots that automatically scoop up data used for the training 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.

Open sourcing and other forms of providing free public access to data could be useful avenues to reducing the risk of market concentration as it allows new entrants to train models with minimal data acquisition costs.

There are also new techniques to reduce the cost of acquiring training 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), which effectively substitutes training data with compute. These mechanisms are growing in importance for the training of frontier foundation models. For example, Thompson (2024) estimates that about 70% of the data used for training GPT-5 would be synthetic 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.

Gans (2024) emphasizes the importance of distinguishing between training data and input data for AI models. He argues that training data is primarily a driver of market entry, while input data more directly impacts the cost and quality of AI offerings in existing markets.

Talent

A McKinsey survey revealed that even in 2021, most organizations found it difficult to hire for AI-related roles in general (Chui et al. 2022). For instance, 32 percent of all survey respondents found it very difficult to hire AI data scientists, while 46 percent found it somewhat difficult. Interestingly, companies that saw the most returns from AI use found it easier than other organizations to hire AI talent, but they still faced difficulties and attempted to close the gap through upskilling (Chui et al. 2022).

Given the large number of players aiming to build foundation models, the demand for researchers and engineers who can build frontier foundation models and who can build the server farms necessary for such models is rising rapidly.

Complementarities between talent and compute imply that workers generally find it most attractive to work with employers who have the computational resources to enable them to work on cutting-edge foundation models. In fact, access to compute is often used as a hiring mechanism. Moreover, talented workers at frontier AI labs also experience significant synergies from working with each other. As a result, new entrants to the market often have to hire people who already work with market leaders and pay a premium to induce them to switch, increasing the cost of entry. Hiring talented engineers is also worthwhile because they can improve algorithmic and software efficiency, thereby cutting down on the exorbitant cost of compute requirements. In the short run, the supply of engineers and researchers with expertise in frontier foundation models is price-inelastic because it takes a number of years to acquire the requisite expertise. This has led to exorbitant salary levels for frontier AI developers.

An additional factor complicating the development of a talent pipeline is that the competition for talent and the high cost of compute make it comparatively less attractive for researchers to remain at universities where they would teach future generations of students. Henshall (2024) reports that the proportion of AI Ph.D.s going into industry has increased from 21% in 2004 to 73% in 2022. Greater public funding for university-based AI researchers, such as the US National Science Foundation’s National Artificial Intelligence Research Resource (NAIRR) may address this problem.

To conclude this section, the market for foundation models exhibits characteristics that generate significant economies of scale and scope. The production of foundation models relies heavily on three key inputs: computational resources, data, and talent. The rapid growth in compute requirements has led to substantial increases in costs and potential supply chain challenges. Concurrently, the scarcity of high-quality training data and intense competition for AI talent further compound the challenges faced by new entrants.

These technological and economic factors contribute to substantial first-mover advantages for current industry leaders. Companies like OpenAI, Google, and others have gained technological leadership through early investments in foundation models. They have also preempted scarce assets crucial for AI development, including vast amounts of compute resources, proprietary datasets, and top-tier AI talent. Whether they can maintain their first-mover advantage remains to be seen.

4. Concentration Concerns

If competition dynamics remained as fierce as they currently are, then there would be little reason for concern about concentration in the market for foundation models. However, in this section, we explore two economic mechanisms that may give rise to significant market concentration and that antitrust authorities may want to analyze, the risk of market tipping and the risks emanating from vertical integration with companies that have significant market power. These two risks are of course not entirely independent of each other, but they capture two complementary plausible pathways through which antitrust concerns in the market for foundation models may materialize.

Risk of market tipping

One plausible concern is that the market for foundation models may follow a similar playbook as digital platforms did in the early 2000s. Similar to generative AI, the markets of many digital platforms started out with a first mover and then experienced rapid entry and fierce competition that often lasted for many years and led to significant losses among the players involved. Ultimately, however, the losses became unsustainable and led to a shake-out whereby a single platform or a small number of platforms survived, became the dominant players, and started to earn significant monopoly rents. The expectation of these profits was what led to the large number of entrants, who had all hoped that their specific offerings would allow them to tip the market and switch to earning monopoly rents. The economic forces that allowed for this phenomenon included (i) significant economies of scale and scope, (ii) network effects, (iii) relatedly, data feedback loops, and (iv) inertia in user behavior (Jeon, 2023). Each of these four forces represents a distinct form of increasing returns.

The technological characteristics of the market for foundation models that we described before feature both some similarities and differences from classic digital platforms. The cost structure of foundation models gives rise to similarly large economies of scale and scope. However, in the market for foundation models, both the costs of frontier models and their capabilities are rising far more rapidly. The implication of the growing investment requirements for state-of-the-art foundation models is that the number of players that a market of a given size can support is shrinking fast, creating a growing force towards natural monopoly.[7] However, offsetting this force is that the capabilities and by extension the market for generative AI are expected to rise as well. It is unclear whether investment requirements or expectations of future profits will rise faster in coming years.

The role of network effects is significantly smaller than for platforms – a given ChatGPT user does not derive significant direct benefits from others joining ChatGPT. The exception is that data feedback loops are material for both platforms and foundation models: for platforms, they are typically about improving matching efficiency; for foundation models, they are about generating additional training data to improve model performance, although this type of user data is not the main source of quality gains. Gans (2024) observes that the existence of data feedback loops does not necessarily guarantee market dominance. He shows that data-enabled learning functions can favor lagging firms if the returns to additional data are decreasing sufficiently quickly. For both platforms and foundation models, user inertia makes it vastly more difficult for a competitor to attract away users and increases the stability of a dominant market position.

A separate factor that is exclusive to foundation models is that there is also a form of increasing returns to intelligence that give rise to an intelligence feedback loop. Foundation models are starting to play a significant role in cognitive work, esp. in programming. If an AI lab has better internal foundation models available than its competitors, then it will be faster at making progress on the next model version. This is already reflected in the current rapid pace of AI progress.[8] As long as the difficulty of making further progress does not rise too fast, this may allow the leading lab to separate itself further and further from the competition, ultimately developing a technology that is so far advanced compared to its competitors that the lab can dominate the market. In the limit, if a lab develops a foundation model that has the ability to improve itself without human input, then the pace of technological progress could be further accelerated, giving rise to what Good (1965) described as an “intelligence explosion.” This would cause first-mover advantages to snowball, leave the competition ever further behind, and create a large monopoly.

Analysis of policy remedies The described factors giving rise to market concentration also represent natural anchor points for potential policy measures that lean against monopolization:

Policies that promote data access and sharing can mitigate the effects of data feedback loops. This could include creating public datasets for AI training, ensuring access to training data for all market participants, and in some instances mandating data sharing. We note that such provisions

[7] This realization seems to have played an important role in the recent take-overs of Inflection AI, Adept, and Character.ai.

[8] For example, Google recently announced that more than 25% of the new code generated at the company comes from AI (Pichai, 2024).

clash with many data governance regulations that aim to restrict the transfer of data between firms, as exemplified by the EU's General Data Protection Regulation. Although this may help to protect users' privacy, it may increase the market power of the firms with the most data, typically, the dominant firms. It also provides new incentives for vertical integration, as we will discuss further below. 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, as mandated, e.g., for gatekeepers under the EU's Digital Markets Act. Encouraging federated learning and privacy-preserving AI techniques could also help reduce the data advantage of large incumbents by promoting methods that allow AI models to learn from distributed datasets without centralizing the data.

Other policies that are under discussion aim to combat inertia in user behavior by reducing switching costs. Subscription fees that are independent of usage discourage users from signing up for multiple providers; this can be remedied by aggregators that provide access to several providers for the same flat fee. Common standards for APIs may make it easier for business users to switch between competing models. Additionally, transparency and user education campaigns can help inform users about the importance of trying different AI services and the potential risks of over-reliance on a single provider, encouraging more active decision-making.

The intelligence feedback loop that we identified could be addressed by mandatory sharing of research results and by ensuring that outsiders gain swift access to the leading companies' most intelligent tools. Before the release of ChatGPT in November 2022, most AI labs shared their research findings about model architectures quite freely. The fierce technological competition has since changed this open publication norm, making it more difficult for new entrants to catch up (Tiku and Vynck 2023). Mandatory disclosures about model architectures as well as public investments in research could help to maintain a more level playing field in terms of technological capabilities. Encouraging open-source AI development is another potential policy measure that may counteract the concentration of AI capabilities in a few dominant players. Both of these strategies are subject to safety caveats that we cover below.

All of these measures may be employed in addition to the traditional pillars of competition policy such as merger reviews. The lessons learned from digital platforms serve as a reminder that proactive responses may be an important element of competition authorities fulfilling their missions.

Risks from vertical integration

Another risk to a competitive landscape in the market for generative AI is that large tech companies with significant market power vertically integrate with producers of foundation models. By consolidating market power, vertical integration may pose the risk of foreclosing rivals' access to essential inputs or important distribution channels, as well as the risk of 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). However, despite these concerns, let us note that vertical integration can also bring significant benefits. It may increase consumer welfare by eliminating multiple markups in the production process, ultimately leading to lower final prices. Furthermore, it may mitigate the risk of hold-up problems associated with specialized assets. 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.

The producers of foundation models may integrate 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.

Upstream vertical integration: Since compute represents 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 made Microsoft the exclusive provider of cloud services for OpenAI, which gives them significant leverage over OpenAI (Warren 2023). Now that Microsoft is developing competing foundation models in-house, it could use this leverage to strangle OpenAI. Likewise, Google Cloud is the preferred provider of cloud services for Anthropic (Anthropic 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 and Meta are employing data from their public platforms (including YouTube, Facebook, Instagram) in pre-training their foundation models, and Elon Musk's xAI is training on data from X/Twitter (Victor 2023, Olson 2024b). At the same time, data-rich technology companies are restricting data access to web crawlers that allow other producers of foundation models to train on their data (Verhulst, 2023), raising anti-competitive concerns.

Downstream vertical integration: 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.

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, some of which 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.

Analysis of policy remedies As the role of foundation models throughout the economy grows, their widespread applicability implies that vertical integration may become a growing concern for competition. Antitrust regulators may want to pay keen attention to acquisitions made by foundation model companies, especially of startups that might compete with them. Dealing with 'nascent competitors' is tricky (Hemphill and Wu, 2020), and antitrust authorities would have more tools available if they are given stronger ex-ante powers to stop acquisitions that can be shown to significantly reduce competition. Ex-post measures in digital markets are often less effective, in part because important intellectual property or business secrets can be transferred before a merger or acquisition is undone.

Given the growing scrutiny, large technology companies are reportedly already wary of making acquisitions in AI (Olson 2024a). Antitrust oversight could also be applied when coordination does not take the form of explicit mergers. For instance, investments and exclusive or 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. Such deals give large technology companies significant power over AI startups, which could be abused for anti-competitive purposes.

Unequal 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 early access to GPT-4 to certain companies (including Duolingo, Stripe and Morgan Stanley), which privileged them in relation to their competitors.

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. Several insights from public utility regulation may therefore be relevant. For example, non-discrimination requirements for access to foundation models could avoid the problem of anyone being excluded from AI services, which could place them at great economic disadvantage. 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).[9]

When foundation models morph into platforms, as demonstrated for example by OpenAI's GPT store, 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. In the EU, such service providers could fall under the gatekeeper provisions of the Digital Markets Act, requiring service providers to meet non-discriminatory service standards.

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 more and more with non-AI providers (including human providers) of different products and services. Policy could promote a level playing field between AI and non-AI providers.

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 strong economic reason why foundation models should be exempt from sectoral regulations including liability, professional licensing, and professional ethics guidelines. If these regulations are derived from the rights of consumers and citizens, a level playing field requires that they apply to foundation models as well.

For instance, governments worldwide have extensive regulations in the education sector, e.g., around student privacy, non-discrimination, and educational standards. A level playing field would require that AI solutions used for the classroom be explicitly subject to the same regulatory standards. Sometimes this may 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 face high risks for such disregard as they are penalized for overriding algorithms that turn out to

[9] 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.

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).

Liability frameworks that remain neutral could avoid situations where humans defer to AI against their better judgment, so that technology follows sectoral regulation and not the other way round.

Safety considerations

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. The further advancement of AI could potentially give rise to many problems, including drastic misuse by unscrupulous actors, catastrophes arising from errors in understanding human preferences, or even human disempowerment or extinction (Hendrycks et al. 2022).

Although current foundation models seem quite safe to use and deploy for most practical purposes, it may be useful for antitrust authorities to pay attention to safety risks in their analysis as such risks may interact with the sector's market structure in significant ways. The reason is that attempts to make the market more competitive may squarely collide with efforts to regulate AI safety. For example, releasing models in an open-source manner, which is highly useful from a competition perspective, would no longer be feasible if such models pose severe safety risks. More generally, competitive dynamics may incentivize developers to cut corners on safety research in order to get ahead. If AI reaches the point where it poses serious safety risks, then considerations about maximizing consumer welfare by keeping the market competitive may have to be sacrificed in order to keep humanity safe from greater harm.

Regulatory frameworks in the EU

The European Union finds itself in a challenging position as it seeks to regulate the rapidly evolving artificial intelligence landscape. Unlike the United States, the EU lacks a homegrown AI powerhouse among the leading labs listed in Table 1. This absence places the EU in a delicate situation: while it can craft laws, regulations, and antitrust rules, it risks driving away foreign companies if these rules become too burdensome. The delayed releases in Europe of the first version of AI models by both Google and Anthropic underscore this risk. Moreover, overly stringent regulations could further hinder domestic AI producers in their efforts to catch up with global leaders. As a result, there is a looming risk of an "intelligence divide," whereby Europe could increasingly fall behind the capabilities of countries at the technological frontier. With this caveat in mind, the EU has introduced two main pieces of legislation that have implications for market structure and competition in the AI industry, the AI Act, and the Digital Markets Act (DMA).

The AI Act of 2024 is particularly noteworthy for its provisions on General Purpose AI (GPAI) systems, which cover foundation models such as LLMs that were trained using a particular compute threshold. These provisions may help increase competition by requiring GPAI providers to be more transparent about their models' capabilities and limitations, thereby allowing smaller companies and new entrants to better understand and potentially compete with established models. Furthermore, the Act requires GPAI providers to collaborate with downstream users in high-risk applications, which could foster a more diverse ecosystem of AI applications. On the other hand, the stringent risk management requirements and the need for extensive documentation might pose a significant burden that reduces competition and increases the costs of AI labs, which may especially affect smaller companies or startups.

The Digital Markets Act, which entered into force in November 2022, complements the AI Act by focusing on ensuring fair competition in digital markets. It introduces the concept of "gatekeepers" – large online platforms with significant market power – and imposes obligations on interoperability and data access on them to prevent anti-competitive practices. These provisions may be quite suitable for large providers of foundation models. However, under the rules of the DMA, a company needs to meet the criteria for gatekeepers for three years to be designated as such. Given the rapid developments in the field of AI, it is questionable whether this threshold is sufficiently timely to meet the Act's goals. The DMA may be useful in reducing barriers to entry in AI services and mitigate some of the data advantages held by larger firms. However, the complexity of designating AI companies as gatekeepers in a rapidly changing technological landscape presents a significant challenge for regulators.

5. Conclusions

The rapid advancement of foundation models has ushered in a new era of artificial intelligence with far-reaching economic implications. As these AI systems grow more capable, they have the potential to reshape entire industries and fundamentally alter the structure of our economy.

The market structure for developing and deploying these models exhibits a strong tendency toward concentration, driven by significant economies of scale and scope. The growing costs of pre-training, coupled with bottlenecks in critical inputs like data and talent, are continually raising the stakes for new entrants. This poses a pressing question about the future of competition in AI: Will we see a winner-take-all scenario where a single dominant firm provides the AI substrate for major parts of the global economy? Or will we experience a diverse ecosystem of AI providers?

In the short term, policymakers face the challenge of how to best implement proactive competition policies to address this concern. These could include mandating data sharing to mitigate the effects of data feedback loops, promoting interoperability and common API standards to reduce switching costs, and encouraging more sharing of research results to counteract the concentration of capabilities. Antitrust authorities could also scrutinize vertical integration attempts, particularly between large tech companies and AI labs, to prevent the leveraging of market power across different sectors of the AI value chain.

Moreover, as foundation models become increasingly integrated into the economy, they may come to resemble public utilities. Non-discrimination requirements for access to these models, similar to those applied in other essential industries, could avoid shutting out potential users.

Yet, in the medium term, the challenge for policymakers is even more complex in the face of AI's rapid evolution. Efforts to promote competition may have to be carefully balanced against other vital considerations, particularly AI safety. As these systems become more powerful, the risks associated with their misuse or malfunction grow exponentially. How can we strike the right balance between fostering innovation, ensuring safety, and maintaining a competitive landscape?

As we grapple with these challenges, another crucial consideration is how to ensure AI technologies benefit society broadly rather than concentrating their advantages among a small subset of stakeholders (Korinek and Balwit, 2024). This extends beyond just distributing economic gains to encompass equitable access to AI-enabled tools, opportunities, and resources that could reshape economic participation in fundamental ways.

Looking ahead, the governance of AI may become one of the defining challenges of our time. It will require unprecedented collaboration between technologists, policymakers, economists, and ethicists. It may force us to grapple with fundamental questions about the distribution of the gains from AI, and ultimately, the role we want these powerful systems to play in shaping our society.

As we stand at this technological inflection point, the decisions we make today will have profound implications for the future. With prudent competition policies, thoughtful regulation, and a commitment to steering this powerful technology toward broadly shared prosperity, we have the opportunity to harness the immense potential of AI while mitigating its risks. With prudent competition policies and thoughtful regulation, AI technology could lead to broadly shared progress and opportunity, aligned with our collective values and aspirations for a more equitable world.

Bibliography

Acemoglu, Daron. 2024. “The Simple Macroeconomics of AI.” Forthcoming, *Economic Policy*.

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.

Anthropic. ‘Anthropic Partners with Google Cloud’. 2023. 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/.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. ‘On the Dangers of Stochastic Parrots’. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021. https://dl.acm.org/doi/10.1145/3442188.3445922.

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

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’. https://www.gov.uk/government/publications/ai-foundation-models-initial-report. ———-. 2024. ‘AI Foundation Models: Update Paper’. https://www.gov.uk/government/publications/ai-foundation-models-update-paper.

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.

———-, Robi Rahman, Loredana Fattorini, Nestor Maslej, David Owen. 2024. How Much Does It Cost to Train Frontier AI Models? *Epoch* (blog). Jun 03, 2024. https://epochai.org/blog/how-much-does-it-cost-to-train-frontier-ai-models.

The cost of training frontier AI models has grown by a factor of 2 to 3x per year for the past eight years, suggesting that the largest models will cost over a billion dollars by 2027.

Coldewey, Devin. ‘After Raising $1.3B, Inflection Is Eaten Alive by Its Biggest Investor, Microsoft’. TechCrunch, 19 March 2024. https://techcrunch.com/2024/03/19/after-raising-1-3b-inflection-got-eaten-alive-by-its-biggest-investor-microsoft/.

DeepMind AlphaProof and AlphaGeometry Teams. 2024. `AI achieves silver-medal standard solving International Mathematical Olympiad problems.’ 25 July 2024. https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/

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.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. `GPTs are GPTs: Labor market impact potential of LLMs.’ *Science* 384(6702): 1306-1308.

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

——— 2024b. ‘Parameter, Compute and Data Trends in Machine Learning’. Retrieved 15 Feb 2024. https://epochai.org/data/epochdb/visualization.

European Commission. 2023. `European Chips Act.’ Regulation (EU) 2023/1781 of the European Parliament and of the Council, September 20, 2023. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32023R1781.

Federal Trade Commission. 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.

Gans, Joshua. 2024. “Market Power in Artificial Intelligence.” NBER Working Paper w32270.

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.

Goldman Sachs. 2024. “GenAI: Too much spend, too little benefit?” Top of Mind, Global Macro Research 129. June 25, 2024. https://www.goldmansachs.com/insights/top-of-mind/gen-ai-too-much-spend-too-little-benefit.

Hagey, Keach, and Asa Fitch. 2024. ‘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.

Hammond, George, and Stephen Morris. 2024. ‘OpenAI Asks Investors Not to Back Rival Start-Ups Such as Elon Musk’s xAI.’ Financial Times, October 2, 2024. https://www.ft.com/content/66e0653e-c446-47b2-8a7f-baa54ccbfb9a.

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.
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.

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.

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/.

Korinek, Anton. 2023a. Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature* 61(4), pp. 1281-1317.

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

Korinek, Anton. 2024. Generative AI for Economic Research: LLMs Learn to Collaborate and Reason. Dec. 2024 Update of "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature*.

Korinek, Anton and Avital Balwit. 2024. "Aligned with Whom? Direct and Social Goals for AI Systems." Oxford Handbook of AI Governance, pp. 65-85.

Korinek, Anton and Donghyun Suh. 2024. Scenarios for the Transition to AGI. NBER Working Paper w32255.

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.

Narayanan, Arvind, and Sayash Kapoor. 2024a. AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton: Princeton University Press.

———-- 2024b. 'AI Scaling Myths'. AI Snake Oil (blog), 27 June 2024. https://www.aisnakeoil.com/p/ai-scaling-myths.

Nathan, Alison, Jenny Grimberg, and Ashley Rhodes. 'Gen AI: Too Much Spend, Too Little Benefit?' Top of Mind. Goldman Sachs Global Macro Research, 25 June 2024. https://www.goldmansachs.com/insights/top-of-mind/gen-ai-too-much-spend-too-little-benefit.

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. 2023. 'ChatGPT Plugins'. Mar 23 2023. https://openai.com/blog/chatgpt-plugins.
OpenAI. 2024. 'OpenAI and Elon Musk'. Mar 5, 2024. https://openai.com/index/openai-elon-musk

Palma, Stefania. 'US Probes Nvidia's Acquisition of Israeli AI Start-Up'. Financial Times, 2 August 2024, sec. Nvidia. https://www.ft.com/content/666ee087-975f-459e-8444-1f9b1d94a008.

Patel, Dylan, and Daniel Nishball. '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.

Pichai, Sundar. 2024. 'Alphabet Q3 2024 Earnings Call.' October 29, 2024. https://blog.google/inside-google/message-ceo/alphabet-earnings-q3-2024/

Reid, Elizabeth. 2023. 'Supercharging Search with Generative AI'. *Google* (blog). 10 May 2023. https://blog.google/products/search/generative-ai-search/.

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.

Stigler Committee on Digital Platforms. 2019. Final Report. https://research.chicagobooth.edu/stigler/media/news/committee-on-digital-platforms-final-report.

Suleyman, Mustafa. The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma. First edition. New York: Crown, 2023.

Thompson, Alan. 2024. `What's in GPT-5? A Comprehensive Analysis of Datasets Likely Used to Train GPT-5.' LifeArchitect.ai. August 2024.

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.

UC Investment Geopolitical Risk Committee (IGRC). 2023. `China's Evolving Semiconductor Strategy.’ University of California, August 11, 2023. https://ucigcc.org/blog/chinas-evolving-semiconductor-strategy/.

U.S. Congress. 2022. `CHIPS and Science Act.’ Public Law 117-167. 117th Cong., August 9, 2022. https://www.congress.gov/117/plaws/publ167/PLAW-117publ167.pdf.

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.

Vestager, Margrethe, Sarah Cardell, Jonathan Kanter, and Lina Khan. ‘Joint Statement on Competition in Generative AI Foundation Models and AI Products’. Federal Trade Commission, 22 July 2024. https://www.ftc.gov/legal-library/browse/joint-statement-competition-generative-ai-foundation-models-ai-products.

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.

Vipra, Jai and Anton Korinek. 2024. ‘Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT’. Brookings Center on Regulation and Markets Working Paper #9. https://www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt/.

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.