RAND

Research Report

JON SCHMID, TOBIAS SYTSMA, ANTON SHENK

Evaluating Natural Monopoly Conditions in the AI Foundation Model Market

For more information on this publication, visit **www.rand.org/t/RRA3415-1**.

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.
© 2024 RAND Corporation
RAND® is a registered trademark.

Cover: cemagraphics/Getty Images and Anastasiia/Adobe Stock.

Limited Print and Electronic Distribution Rights

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, please visit www.rand.org/pubs/permissions.

About This Report

Because of the wide variety of tasks they can be used to perform, *foundation models*—a class of artificial intelligence (AI) models trained on large and diverse datasets and capable of performing many tasks—have the potential to have a large effect in shaping the economic and social effects of AI. The authors of this report examine the economic and production attributes of pre-trained foundation models to answer a single question: Is the market for foundation models a natural monopoly? To assess whether foundation models are a natural monopoly, we establish a set of empirical criteria for classifying a market as a natural monopoly and then apply these criteria to the status quo foundation model market and to four hypothetical scenarios set in 2027 to understand possible future market dynamics. The criteria set forth and the results of the analysis should be of interest to anyone concerned with the set of important and thorny questions that sit at the intersection of AI and economic policy.

Technology and Security Policy Center

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Technology and Security Policy Center, which explores how high-consequence, dual-use technologies change the global competition and threat environment, then develops policy and technology options to advance the security of the United States, its allies and partners, and the world. For more information, contact tasp@rand.org.

Funding

Funding for this work was provided by gifts from RAND supporters.

Acknowledgments

The authors would like to thank Benjamin Sperisen and Nicholas Brown for their insightful comments on a draft of this report and Emily Haglund for her overall help with the document. Emma Westerman and Jeff Alstott were critical to conceptualizing the research question. We are grateful to our reviewers, Edward Parker and Anton Korinek, for their detailed reviews; their comments significantly improved this report.

Summary

The authors of this report examined the economic and production attributes of pre-trained artificial intelligence (AI) foundation models to answer the following questions: Does the market for foundation models have the characteristics of a natural monopoly, and, if so, is regulation of that market needed? A *natural monopoly* refers to a market in which the total cost of serving the full range of demand is lower for a single firm than for multiple firms. Unlike a conventional monopoly, in a natural monopoly, competition and traditional antitrust policy cannot be assumed to alleviate the problems associated with concentrated market power (e.g., elevated prices, poor product quality, etc.). Therefore, identifying whether the market for foundation models is a natural monopoly is critical to selecting an appropriate policy response.

To assess whether foundation models are a natural monopoly, we establish a set of theoretically derived[1] empirical criteria for classifying a market as a natural monopoly and then apply these criteria to the status quo foundation model market and to four hypothetical scenarios set in 2027 to understand possible future market dynamics.

Application of the natural monopoly criteria to the status quo foundation language model market (as of January 2024) indicates that the current case for a natural monopoly is relatively strong. This conclusion is based on the observations that the current generation of foundation models is reasonably homogeneous, economies of scale are high, costs are largely sunk, and network effects and economies of scope are present. Table S.1 presents the application of the natural monopoly criteria to the status quo system.

[1] Paul L. Joskow, "Regulation of Natural Monopoly," in A. Mitchell Polinsky and Steven Shavell, eds., *Handbook of Law and Economics*, North-Holland, 2007.

Table S.1. Application of Natural Monopoly Criteria to the Status Quo Market

Natural Monopoly Criteria	Foundation Model Variable	Status Quo Market
Homogeneous good	Limited product variation (at pre-trained foundation model level)	Yes
Economies of scale	Cost of training relative to variable costs	High
Sunk cost	Resale value of initial compute, data, labor, and foundation model	High
Network effects	Community-led development (open-source models)	Moderate
Economies of scope	Firms have multiple products using common foundation model	High
Case for a natural monopoly		Strong

To consider how market structure may change in the future, we vary two technology variables critical to determining the future competitive structure of the market for foundation models: the scaling hypothesis and the cost of compute technology.[2]

In scenarios in which the scaling hypothesis is assumed to hold—i.e., the relationship between performance and model size/compute expenditure persists—the case for a natural monopoly relative to the status quo market is stronger. This is largely due to the increase in economies of scale and risk of sunk costs associated with pre-training very large foundation models. In contrast, in scenarios in which the scaling hypothesis breaks down, the argument for a natural monopoly decreases relative to the status quo market. Table S.2 summarizes these findings and indicates the likely market structure to emerge in the four scenarios considered.

[2] These two technology variables were selected based on their expected role in determining the future production costs characteristics of foundation models. In Chapter 5, we explain the logic by which each variable can be expected to affect production costs for foundation model developers.

Table S.2. Summary of Scenario Outcomes (relative to status quo)

Natural Monopoly Criteria Relative to Status Quo	Scenario 1 (scaling hypothesis fails, cost low)	Scenario 2 (scaling hypothesis fails, cost high)	Scenario 3 (scaling hypothesis holds, cost low)	Scenario 4 (scaling hypothesis holds, cost high)
Homogeneous good	Lower	Slightly lower	Same	Same
Economies of scale	Lower	Slightly lower	Higher	Much higher
Sunk cost	Lower	Slightly higher	Higher	Much higher
Network effects	Lower	Lower	Same	Higher
Economies of scope	Lower	Slightly lower	Same	Uncertain
Likely market structure	Monopolistic competition	Oligopoly	(weak) Monopoly	(strong) Monopoly
Case for a natural monopoly relative to status quo	Weaker	Weaker	Stronger	Stronger

While we find that the status quo market has characteristics of a natural monopoly, we believe that the rationale for natural monopoly regulation is weak. This is due to the low observed social cost associated with the current market structure. Potential social costs of concentration in the market for AI foundation models include pricing above marginal cost, low product quality, costs associated with market concentration in the market for compute, the environmental impact of large training runs, and systemic risk. If evidence of significant social costs emerges, the question of regulation should be reconsidered.

Contents

Figures and Tables

Figures

Tables

Chapter 1. Introduction

The effect of artificial intelligence (AI) on the global economy can be expected to be large. Recent research estimates that nearly all occupations already have some job task exposure to AI applications, and up to 15 percent of occupations are highly exposed to AI.[3] Another recent study estimates that, for 19 percent of U.S. workers, more than half of their typical work activities could be substituted by large language models (LLMs), a subset of AI applications.[4]

Because of the wide variety of tasks they can be used to perform, *foundation models*—a class of AI model trained on large and diverse datasets and capable of performing many tasks—have the potential to have an outsized effect in shaping the economic and social effects of AI.[5] Foundation models are a core technology upon which a growing ecosystem of applications and products is being developed.[6] For example, foundation models can be fine-tuned for chatbots,[7] medical image analysis,[8] autonomous driving,[9] and many other applications.

[3] Tobias Sytsma and Éder M. Sousa, *Artificial Intelligence and the Labor Force: A Data-Driven Approach to Identifying Exposed Occupations*, RAND Corporation, RR-A2655-1, 2023.

[4] Certain occupations, such as interpreters and translators, financial quantitative analysts, tax preparers, proofreaders and copywriters, legal secretaries, and administrative assistants have task exposures that exceed 75 percent. The cited study does not make predictions about the adoption timelines for LLMs (Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," arXiv, arXiv:2303.10130, 2023).

[5] A technology's *generality* refers to the extent to which it has broad application, often as a critical input to enabling downstream technological innovation. Highly general technologies, often referred to as *general-purpose technologies*, have been observed to spur waves of multi-sector innovation and economic growth (Timothy F. Bresnahan and M. Trajtenberg, "General Purpose Technologies 'Engines of Growth'?" *Journal of Econometrics*, Vol. 65, No. 1, January 1995).

[6] To illustrate this distinction, consider the well-known application ChatGPT. ChatGPT is a fine-tuned model—via instruction fine-tuning and reinforcement learning from human feedback—based on the generative pre-trained transformer (GPT) model called GPT-4. The direct focus of this report is on entities such as GPT-4, not ChatGPT.

[7] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Tasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv, arXiv:2307.09288v2, July 19, 2023.

[8] Akmalbek Bobomirzaevich Abdusalomov, Mukhriddin Mukhiddinov, and Taeg Keun Whangbo, "Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging," *Cancers*, Vol. 15, No. 16, August 2023.

[9] Anoli Marathe, Deva Ramanan, Rahee Walambe, and Ketan Kotecha, "WEDGE: A Multi-Weather Autonomous Driving Dataset Built from Generative Vision-Language Models," *Computer Vision Foundation*, 2023.

Industrial structure and the chosen regulatory response to non-competitive markets have been observed to affect the quality of regulated products and services,[10] as well as their prices.[11] For example, a natural monopolist may cause social harms by charging excessive prices, limiting production, or underinvesting in research and development (R&D).[12] Therefore, the structural dynamics, industrial arrangement, and state of competition within the market for foundation models is of first-order concern for anyone interested in the economic and social impacts of AI.

In the report that follows, the authors consider the economic and production attributes of foundation models to answer a single question about industry composition: Is the market for foundation models a natural monopoly?[13] A *natural monopoly* refers to a market in which the total cost of serving the full range of demand is lower for a single firm than for multiple firms. In a natural monopoly, as in a monopoly more generally, a single producer can charge excessive prices or sell low-quality goods relative to a competitive market. However, as opposed to a conventional monopoly, in a natural monopoly, competition, and thus traditional antitrust policy, cannot be assumed to ameliorate the problems associated with concentrated market power. To select an appropriate policy response, it is therefore critical to identify whether a given industry is a natural monopoly.

In addition to addressing the narrower question of whether foundation models are a strong candidate for natural monopoly designation, we believe the framework provided here may be of utility in addressing broader questions related to market concentration in the foundation model market, and within the AI ecosystem more generally. Specifically, market concentration in the foundation model market will be determined, in part, by variables such as the character of model production costs, the extent of product differentiation, network effects, and the presence of economies of scope and scale. In describing how these variables manifest in the specific case of foundation models and proposing preliminary means for measuring these variables, the authors of this report seek to inform future investigation into market concentration, market power, and industry structure in the emerging AI ecosystem.

To assess whether foundation models are a natural monopoly, we follow a four-step research design. First, we establish a set of generic criteria for classifying a market as a natural monopoly.

---

[10] Chunrong Ai and David E. M. Sappington, "The Impact of State Incentive Regulation on the US Telecommunications Industry," *Journal of Regulatory Economics*, Vol. 22, No. 2, 2002.

[11] William J. Baumol, Janusz A. Ordover, and Robert D. Willig, "Parity Pricing and Its Critics: A Necessary Condition for Efficiency in the Provision of Bottleneck Services to Competitors," *Yale Journal on Regulation*, Vol. 14, No. 145, 1997.

[12] Paul L. Joskow, "Regulation of Natural Monopoly," in A. Mitchell Polinsky and Steven Shavell, eds., *Handbook of Law and Economics*, North-Holland, 2007.

[13] Narechania (2021) and Vipra and Korinek (2023) consider questions related to the one considered here. We recommend those studies to any reader interested in additional discussion of the relationship between market structure and artificial intelligence (Tejas N. Narechania, "Machine Learning as Natural Monopoly," *107 Iowa Law Review*, Vol. 1543, 2021; Jai Vipra and Anton Korinek, "Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT," Brookings, Center on Regulation and Markets Working Paper #9, September 2023).

Cost subadditivity, the necessary and sufficient condition for natural monopoly, is unobservable, and, therefore, our criteria are based on economic features that leave empirical markers.[14] Second, we describe the production process for foundation models. Third, we adapt the natural monopoly criteria to the particular production process for foundation models. This step illustrates how the generic natural monopoly criteria would manifest in the particular case of foundation models. Finally, we apply the adapted natural monopoly criteria to the status quo foundation model market, as well as to four hypothetical scenarios set three years in the future.

The focal market of this study is that of pre-trained foundation models at the technological frontier (e.g. OpenAI's GPT, Anthropic's Claude, and Google's Gemini). We have scoped the study to pre-trained models for several reasons. First, because pre-trained models can be fine-tuned or adapted to specific domains or datasets, the final consumer-facing output of a foundation model is not a single product, but a variety of potential products, spanning multiple markets.

In general, public users of AI models interact with fine-tuned models. For instance, chatbots like ChatGPT and Claude allow users to interact with a version of a foundation model that has been fine-tuned for chat. However, even users who access models through an application programming interface (API) do not typically interact with a pre-trained version of the model. For example, OpenAI performs additional tuning of its GPT models using InstructGPT, an additional training process that refines the raw pre-trained model to produce "more truthful and less toxic" responses.[15] We include this type of alignment training as fine-tuning and distinguish pre-trained models as those that directly result from the initial training process. Focusing on pre-trained models allows us to focus on a single market, the scope appropriate for analysis of market competition.

Second, pre-trained models reflect the large upfront investments in compute required to develop a foundation model. By focusing on pre-trained models, we can assess how these fixed (or sunk) costs influence market dynamics and barriers to market entry. Third, much of the unique capability of foundation models comes from the pre-training process (e.g., transfer learning), so understanding pre-trained models provides insights into the core value-add technology of foundation models. Fourth, fine-tuned models, due to the low barriers to entry and the differentiated products resulting from each fine-tuning process, do not present a particularly strong case for being a natural monopoly. Finally, we focus on pre-trained models because not

[14] A cost function is subadditive when a single firm's total cost of delivering a good or service over the full range of market demand is lower than the combined total cost incurred by multiple firms when delivering the same good or service.

[15] OpenAI, "Aligning Language Models to Follow Instructions," webpage, January 27, 2022; Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, et al., "Training Language Models to Follow Instructions with Human Feedback," arXiv, arXiv:2203.02155, March 4, 2022.

all foundation model developers (e.g., Meta) host their own models for inference and therefore do not incur the same cost as firms that do.

Chapter 2. Natural Monopoly Criterion

Monopoly, Natural Monopoly, and Policy Intervention

A *monopoly* refers to a market with a single supplier—i.e., a market characterized by the absence of competition.[16] The absence of competition allows the monopolist to set prices, product quality, and production levels in a way that can be socially suboptimal. Empirically, the negative welfare effects of monopoly can be substantial.[17] As a result, monopolies are often regulated through such policies as price controls, licensing, or (in extreme cases) state ownership, to mitigate potential harms.

Monopolies emerge dynamically and can arise for various reasons. For example, patents provide a firm with a temporary legal right to a monopoly on the patented product. In this case, restriction of competition is argued to be justified based on the incentive it provides to innovation.[18] Alternatively, a monopoly can arise due to anti-competitive business practices. Antitrust laws seek to prevent or punish such behavior to promote a competitive market. Finally, monopolies can emerge naturally, based on the production characteristics of the product or service in question. A natural monopoly exists when a single firm can produce a good or service for the entire market at a lower total cost than would be incurred by a combination of multiple firms.

In practice, there is no clear distinction between natural monopolies and other forms of imperfect competition, such as monopolistic competition or oligopoly. For example, subtle differences in how product substitutability is measured can determine whether a market is deemed to have natural monopolistic characteristics. As a result, in reality, there is a continuum between textbook natural monopolies and perfect competition, with most industries falling somewhere in between the two. Nevertheless, identifying whether an industry is a natural monopoly matters for two reasons. The first relates to general characteristics of monopolies and the resultant potential for economic inefficiency and consumer harm. The second relates to the particular characteristics of natural monopolies and the implications of these characteristics for appropriate policy intervention.

While natural monopolies do not typically arise from anti-competitive behavior, firms in this position can still take advantage of market power by charging excessive prices. This behavior

---

[16] Irving Fisher, *Elementary Principles of Economics*, The Macmillan Company, 1912.

[17] James A. Shmitz Jr., *New and Larger Costs of Monopoly and Tariffs*, Federal Reserve Bank of Minneapolis, Research Department Staff Report 468, September 11, 2012.

[18] Kenneth Arrow, "Economic Welfare and the Allocation of Resources," in National Bureau of Economic Research, *The Rate and Direction of Inventive Activity: Economic and Social Factors*, Princeton University Press, 1962.

can have negative welfare effects. For example, if a natural monopolist sets prices above the marginal cost of production, allocative efficiency is sacrificed, and welfare will fall.[19] Furthermore, a natural monopolist may have little incentive to innovate, leaving consumers with low-quality products. There are also inefficiencies that arise when a natural monopolistic market has too much competition. For instance, there are inefficiencies that arise from the duplication of facilities, where multiple firms invest in redundant infrastructure rather than sharing a common network. This duplication can result in increased average costs across firms and higher prices for consumers.

While the inefficiencies and consumer harms associated with natural monopolies may mirror those of monopolies more generally, the appropriate policy intervention for these two conditions differs. Because in a natural monopoly a single firm is cost dominant to multiple firms, the conventional antitrust intervention to monopoly—i.e., promotion of competition via new market entry—may fail to redress consumer harms. In fact, in a natural monopoly, the existence of more than one market participant may exacerbate consumer harms. In sum, in a natural monopoly, antitrust policy levers may not always be appropriate, leaving the government to redress consumer harms by regulating price, profits, product quality standards, availability, and market entry.

Proposed Indicators of Natural Monopoly

Defining Natural Monopoly

A natural monopoly exists when a single firm can provide a homogeneous good or service for the full market at a lower total cost than would be incurred by more than one firm. A homogeneous product is one that is perfectly substitutable: A consumer will be indifferent to the exchange of one version for another one and will base purchasing decisions purely on the price of the good. More formally, the necessary and sufficient condition for a single-product natural monopoly is cost function subadditivity.[20] A cost function is subadditive when a single firm's total cost of delivering a good or service over the full range of market demand is lower than the combined total cost incurred by multiple firms when delivering the same good or service.[21] Such conditions can prevent market entry because the cost structure realized by the incumbent firm is not attainable by a would-be entrant.

Consider a market where *K* firms each produce a homogeneous output q^k, where *k = 1, 2, …, K*. In this market, total output for a product is represented by $Q = \sum_{k=1}^{K} q^k$. $C(q)$ is the dollar

[19] Tanga McDaniel Mohr, "Natural Monopoly," in Timothy C. Haab and John C. Whitehead, eds., *Environmental and Natural Resource Economics: An Encyclopedia*, Greenwood, 2014.

[20] Multi-product cost subadditivity requires both economies of scope and economies of scale to be present (Joskow, 2007).

[21] Joskow, 2007.

cost for a firm to produce a certain amount q of produced output. In this market, a cost function C is said to be subadditive if $C(Q) < C(q^1) + \cdots + C(q^k)$. That is, if the inequality holds for all possible multi-firm disaggregations of output vector q, *C(q)* is subadditive, and the market is a natural monopoly at q.[22]

Critically, the existence of potential entrants (i.e., market contestability) has the same effect as the existence of actual competitors: They limit the incumbent's ability to set prices. In fact, contestability may even obviate the case for a natural monopoly's regulation, as the market-power-constraining effect of a potential entrant, in theory, produces a welfare result that equals price regulation.[23]

Empirical Assessment of Natural Monopoly

Cost subadditivity—the necessary and sufficient condition for defining a natural monopoly—is not directly observable. This is because defining subadditivity requires comparison to an alternative hypothetical market structure—i.e., it requires comparison to an unobserved counterfactual. This measurement challenge is amplified in an emerging market such as foundation models, in which (1) firms may price below cost (and thus not reveal their true cost structure) to gain market share, and (2) detailed historical input cost estimates are not available.

However, in practice, a subadditive production cost structure typically arises due to product homogeneity, economies of scale, network effects, sunk costs, and economies of scope. These features, while not trivial to measure, in some cases leave empirical markers. We therefore propose a natural monopoly identification criterion based on the accumulation and strength of indicators. The criterion is as follows: For a single homogeneous product, the presence of economies of scale, sunk costs, and network effects strengthens the case for a natural monopoly existing. In a multi-product setting, the case is based on these factors and economies of scope. As the magnitude of these effects increases, so does the case for natural monopoly. The remainder of this section presents the assessment criteria (summarized in Table 2.1).

[22] William W. Sharkey, *The Theory of Natural Monopoly*, Cambridge University Press, 1982.

[23] Joskow, 2007.

Table 2.1. Indicators of Potential Natural Monopoly

Criteria	Definition
Product homogeneity	A homogeneous product is one that is perfectly substitutable: A consumer will be indifferent to the exchange of one version for another one and will base purchasing decisions solely on price.
Economies of scale	Economies of scale exist when the average cost of production decrease when its production increases.
Sunk cost	Sunk costs exist when the value of an investment in an asset is worth less than an alternative use of the investment.
Network effects	Network effects exist when the utility of a good or service increases as the number of users grows (i.e., with scale).
Economies of scope	Economies of scope exist when the total cost of producing two or more distinct products within a single firm is less than the total costs of producing those products within more than one firm.

Criterion 1: Product Homogeneity

In theory, a natural monopoly requires that the products within the focal market be homogeneous. A homogeneous product is one that is perfectly substitutable: A consumer will be indifferent to the exchange of one version for another one and will base purchasing decisions purely on the price of the good. In contrast, non-homogeneous (i.e., differentiated) products are imperfect substitutes for each other and thus compete, at least in part, to satisfy the same market demand. In practice, with the exception of commodities such as soybeans or oil, true product homogeneity is rare; products within a market are typically, to some degree, imperfect substitutes. For the purposes of establishing our natural monopoly assessment criteria, we will simply posit that the more homogeneous (i.e., less differentiated) the products under consideration are, the stronger will be the case for natural monopoly.

Product homogeneity is dependent on how the focal market is defined.[24] To illustrate the relationship between market definition, product homogeneity, and natural monopoly, consider the market for multiplayer online video games. Video game developers operate in an industry characterized by high fixed costs and low variable costs in their production process. It costs millions of dollars to develop a new online video game, most of these initial costs are sunk, and it cost very little to distribute one. Furthermore, multiplayer online games have network effects; their appeal increases as more players join the game. If the market in this case is defined broadly as multiplayer online video games, the market is not a strong candidate for natural monopoly

[24] Joskow contends that market definition is the most thorny step in natural monopoly determination related to defining the scope of the product, citing product substitutability and defining the geographic scope of the market as the critical components of determining natural monopoly and stating, "Whether an industry is judged to have classical natural monopoly characteristics inevitably depends on judgments about the set of substitute products that are included in the definition of the relevant product market . . . and the geographic expanse over which the market is regulated" (Joskow, 2007, p. 1248).

status because there are many other games to which players could switch without incurring a full loss of utility. Defined broadly, the market for video games instead displays characteristics of monopolistic competition, where each video game producer sells a unique variety of a game. Under this definition, the possibility for product differentiation in the video game market allows many firms to compete and limits the ability for a single firm to take control of the entire market. However, if the boundaries of the market are drawn more narrowly, the extent of assessed competition may change. For example, if the market were defined as multiplayer online video games with a female protagonist set in a medieval European setting, the number of entrants would decrease. Whether the narrow or the broad market boundaries are appropriate should be determined by product substitutability—i.e., by the character of demand.

Criterion 2: Economies of Scale

Economies of scale occur when a firm's average cost of production decrease as its production output increases.[25] To illustrate the relationship between economies of scale and natural monopoly, consider an industry in which a firm faces a production function characterized by high fixed costs and low variable costs across the full range of production.[26] In this case, increases in production spread the fixed costs over a larger number of units, and the average cost of production decreases indefinitely.[27]

Canonical cases of natural monopoly matching this high fixed-cost/low variable-cost structure include railways, electricity provision, water provision, and wired telecommunications. In each of these cases, the very high initial fixed costs associated with creating the necessary service-provision infrastructure (e.g., rail networks, power lines, water service lines, and telephone lines), coupled with the low cost of servicing an additional consumer, along with the homogeneity of the product in the eyes of consumers, establish a fundamental piece of the argument for defining these markets as natural monopolies.

Criterion 3: Network Effects

Network effects, or network externalities, are, in essence, demand-side economies of scale. A network effect refers to the phenomenon whereby the utility of a good or service increases as the number of customers grows. For example, the utility of a telephone connected to a network of two phones is low; the phone's owner can call just one other phone. As more phones are added to the network, the phone's owner can call more people, and the phone's utility increases.

---

[25] Economies of scale can be extended to multiproduct firms as economies of scope, where production costs decline as the number of products the firm produces increase.

[26] *Fixed costs* are costs that are independent to the level of production. Examples of fixed costs include R&D expenditure and provision of infrastructure (e.g., railroad tracks). *Variable cost* refers to the cost that increases as output increases.

[27] While declining average cost due to scale economies is one of the primary ways that a firm may have subadditive costs, a single product firm can have a subadditive cost function during a portion of production when costs are rising, so long as the firm's total cost is less than that of that of multiple firms.

The effect of network externalities on natural monopolies is to amplify the extent to which a firm benefits from scale. Network effects prevent market entry because, when present, they can increase the demand attained by the incumbent beyond that attainable by a new entrant.

Criterion 4: Sunk Cost

Fixed costs are often sunk costs in markets within natural monopolies. By definition, a *sunk cost* refers to a cost that has already been incurred and cannot be recovered. For a given investment, sunk costs exist when the asset purchased is worth less than some alternative use for that asset. If the investment has no alternative use, sunk costs are complete. Within the context of natural monopolies, sunk costs are typically associated with upfront infrastructure investments that cannot be easily reappropriated once they have been made. For example, the cost of developing rail routes is partially sunk because their value outside of rail service is minimal. While all sunk costs are fixed, not all fixed costs are sunk. For example, if a firm purchases industry-standard or highly modular manufacturing equipment that can be resold, the cost of this equipment would be fixed but not sunk.

A market is contestable to the extent that a new entrant can quickly enter the market and offer a competing product at a low cost. High sunk costs act as a barrier to contestability, limiting potential entrants by increasing the expected cost of entry (absent sunk costs, the cost of entry is zero because fixed entry costs can be fully reappropriated in the event of failure). Thus, sunk costs can limit the contestability of a market by dissuading entrants from making the initial investments required to produce.

Additionally, the risk of sunk costs being lost can affect a firm's entry decision and thus market contestability. If the probability that a project may fail is high (e.g., as they might be for an R&D process or a technically challenging project) and the costs are sunk, the deterrent to enter is also high, and contestability falls.

Criterion 5: Economies of Scope

Economies of scope occur when the total cost of producing two or more distinct products within a single firm is less than the total costs of producing those products within multiple firms. In a multiproduct setting, subadditivity means that a single firm's total cost to produce any combination of these products is less than the cost of producing the same combination of the products by more than one firm. Just as economies of scale let firms realize efficiency by spreading fixed cost over a larger number of units, economies of scope let firms spread their fixed cost over multiple distinct product lines. Economies of scope occur when a firm has developed some resource that serves as a critical input to more than one product line. For example, by developing a common operating system for use across phones, tablets, laptops, and workstations, Apple spreads its development cost across each of these product lines.

Technological Change in Natural Monopoly

The discussion above illustrates that a market's status as a natural monopoly is fundamentally about the economics of production of the focal product and the contestability of the market. These variables, in turn, are highly susceptible to technological change. To illustrate how technological change relates to natural monopoly, below we briefly consider the effects of technological change on a firm's production cost and market contestability.

Production Economics

For a given market, technological innovation in the production process can change cost conditions and thus cost subadditivity. If diseconomies of scale exist at some point of output, a technological or process innovation that decreases fixed cost will shift a firm's minimum efficient scale downward, which will increase the range of output on which more than one firm can be supported. For example, the case for water supply and distribution being a natural monopoly depends on the high initial cost associated with building water treatment facilities and distribution networks. However, the advent of new water treatment technology, such as decentralized treatment modules, reduces the fixed cost requirement, eroding the case for natural monopoly.[28]

Technological change can also convert fixed sunk cost to recoverable costs. If a modular system—i.e., one that relies on standardized components or interfaces—replaces a bespoke system, it may be possible to recover a higher proportion of costs by reusing standardized or modular components beyond their original use.[29]

Market Contestability

Technological change can also affect market contestability.[30] In the language of our natural monopoly assessment criteria, technological change can decrease product homogeneity within a market by introducing substitutes.

The case of landline telephones illustrates how technological innovation can create contestability. In the early market for landline telephones, the fixed costs to establish telephone infrastructure were very high, these costs were mostly sunk, network effects were large, variable costs were low, and product differentiation was minimal. These conditions made a strong case that landline telephony was a natural monopoly. However, in 1995, voice over internet protocol (VoIP) technology allowed a user to make voice calls over their internet connection. By 2003,

---

[28] Peter Debaere and Andrew Kapral, "The Potential of the Private Sector in Combating Water Scarcity: The Economics," *Water Security*, Vol. 13, No. 1, August 2021.

[29] Lynne Kiesling, "Economic Foundations: Natural Monopoly Theory II," *Knowledge Problem*, February 9, 2023.

[30] A market is contestable if a would-be market entrant is able to enter or exit the market while facing low barriers of entry or exit.

roughly 25 percent of calls were made using VoIP.[31] The introduction of a substitute good, via technological change, increased competition and decreased the strength of the argument that the landline telephone market was a natural monopoly.

When to Regulate a Natural Monopoly

The technical question of whether a market is a natural monopoly should be separated from normative questions about whether to regulate it. [32] This is because the case to regulate depends on the existence of social cost via market failures.[33] However, not all markets that match the technical definition of a natural monopoly impose significant social cost, and the cost (financial or political) of regulation may be larger than the cost borne from market inefficiencies.[34] Additionally, the social and practical costs of natural monopolies and their regulation can change over time. As a result, some industries have gone through periods in which they were strictly regulated as natural monopolies, only to later become unregulated markets of imperfect competition. For instance, the cable television industry has undergone several periods of regulation and deregulation since the 1960s.[35]

Sources of social cost in a natural monopoly include high prices (i.e., pricing below marginal cost), low product quality, limited product availability, and low levels of innovation within the market. Social costs may also manifest as the opportunity cost of resources invested in redundant infrastructure or in the redundant use of highly skilled labor. The large scale at which natural monopolies operate both precludes competition and, when the social cost of the market inefficiency is extremely high, can create justification for government ownership out of public interest. Public utilities are an example of government-owned natural monopolies.

---

[31] Johnson Hur, "The History of VoIP," *BeBusinessed*, webpage, undated.

[32] Joskow, 2007.

[33] A second necessary condition for regulation is the existence of policy instruments that would plausibly ameliorate market failures such that ex-post social costs are lower social costs associated with the natural monopoly.

[34] Joskow, 2007.

[35] Adam M. Zaretsky, "I Want My MTV . . . and My CNN and My ESPN and My TBS and . . . ," *Regional Economist*, July 1995.

Chapter 3. The Market for and Development of Foundation Models

In this chapter, we provide an overview of the key technological and economic factors driving the development and adoption of foundation models. First, we discuss the concepts of pre-training and fine-tuning that enable the versatility of these models. Then, we analyze the market structure surrounding both closed- and open-source models. Additionally, we provide an overview of the computing, data, algorithms, labor, and other inputs necessary for creating pre-trained foundation models. Lastly, we categorize the various cost factors associated with foundation model development as either fixed or variable costs.

Defining Foundation Models

The landscape of AI has undergone an evolution—transitioning from basic algorithms to sophisticated neural networks.[36] The most recent surge of AI progress has been propelled by *foundation models*—a term defined and popularized by Stanford's Center for Research on Foundation Models (CRFM). CRFM defines a *foundation model* as any model trained on a broad set of data that can be adapted for various downstream uses.[37] Some prominent examples of foundation models include OpenAI's GPT, Meta's Llama, Anthropic's Claude, and Google's Gemini models.

Pre-Training and Adaptation

Unlike narrow AI models designed to perform specific tasks, the broad capabilities of foundation models make them general-purpose technologies.[38] Foundation models achieve their generalizability using two key steps: pre-training and adaptation.

In pre-training, the model is exposed to a diverse variety of datasets. At this time, the model identifies patterns in the training data, without explicit labels or human supervision, ultimately building a robust internal representation of the data. During this process, foundation models can

[36] Ian Sample, "Race to AI: The Origins of Artificial Intelligence, from Turing to ChatGPT," *Guardian*, October 28, 2023.

[37] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al., *On the Opportunities and Risks of Foundation Models*, Center for Research on Foundation Models, August 2021.

[38] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan,, Johannes Gehrke, Eric Horvitz, Eve Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang, "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," arXiv, arXiv:2303.12712, March 22, 2023; Benj Edwards, "ChatGPT Sets Record for Fastest-Growing User Base in History, Report Says," *arsTechnica*, February 1, 2023.

gain a wide array of new and sometimes unforeseen capabilities, which allow them to perform a variety of tasks.[39]

To achieve optimal task performance, however, developers often need to tailor the model for the specific downstream task using fine-tuning or prompt engineering.[40] In *fine-tuning*, a pre-trained model is further trained using supervised learning and smaller, task-specific, labeled datasets, enhancing the model's performance on that task. In *prompt engineering*, a prompt is provided to the same context window in which user input is provided to the model at the time of inference (i.e., without modifying the underlying model weights). This prompt is chosen to condition the model's performance on a particular downstream task, promoting the use of some aspects of pre-trained knowledge while suppressing others.

Modalities

In the evolving AI landscape, foundation models are characterized not only by their scale and versatility but also by the variety of data modalities (e.g., text, image, audio, video) they can process. The box on this page summarizes some of the primary modalities of foundation models.[41]

Common Foundation Model Modalities

Natural language: LLMs are foundation models that can interpret and generate text. Their versatility lies in their capacity to be fine-tuned for a variety of language-based tasks—including summarization, content creation, and coding.

Visual models: Visual models are capable of general-purpose image or video processing. Their significance lies in their ability to be fine-tuned explicitly for tasks such as facial recognition or medical imaging.

Tactile models: General-purpose robotics have already demonstrated their versatility in performing various physical tasks across diverse environments.

Multimodal models: Multimodal models represent a significant shift toward more integrative AI systems. Their capability to process multiple modalities—such as text and images—marks an important step in the evolution of AI, moving toward models that can ingest a variety of data types seamlessly.

The Market for Foundation Models

The market for foundation models is complicated, evolving, and multifaceted. Currently, there are many organizations that produce foundation models. For example, data from Epoch,[42] which tracks the existence of foundation models across

[39] Bommasani et al., 2021.

[40] Bommasani et al., 2021.

[41] The selected modalities in the box on this page are where foundation models have the greatest market impact; ancillary modalities (e.g. audio, olfactory, gustatory) are not yet sufficiently economically significant to warrant inclusion in this market analysis.

[42] Epoch, "Notable AI Models," webpage, undated.

modalities, suggest that there have been 530 foundation models released by 167 organizations since 2012.[43] Of the 530 models released since 2012 in the Epoch data, the largest share, 43 percent, were language models, meaning that their primary purpose is to generate text.

Foundation models differ in terms of the level of access that developers provide to the model. Some developers provide full access to the model's weights, making them available for distribution and reuse. Model weights are the parameter values that have been tuned during training to capture patterns in the data. By sharing these weights, developers enable others to build on top of the foundation model for their own applications without needing to train a model from scratch. In contrast, other developers restrict access to the weights, instead providing access to model inputs and outputs over an API. Currently, frontier models (GPT-4, Gemini, Claude 3) are all closed-source.[44]

Foundation Model Development

Foundation models require several critical inputs in development and operation: algorithms, data for training and fine-tuning, and computing resources (compute). Some inputs are more critical in the model developing and training stages (e.g., training data), while others play important roles in both training and inference (e.g., compute). Of course, human capital and other operational costs (e.g., office space) are important inputs to model production and operation as well.

Figure 3.1 displays a stylized version of the collection of technologies and tools used to build AI-enabled applications. At the bottom of the stack are compute resources, human capital, and training data. Compute resources can either be purchased directly or rented from cloud providers. All of these factors feed into the development of pre-trained foundation models.

Foundation models are shown in the middle of Figure 3.1, highlighting their centrality to the overall ecosystem. The foundation model is the result of the training process, where the model learns patterns in the training data. However, foundation models generally require fine-tuning before they can be used in downstream apps. For instance, OpenAI's GPT-4 is tuned for a chat context and used in the in-house ChatGPT app.

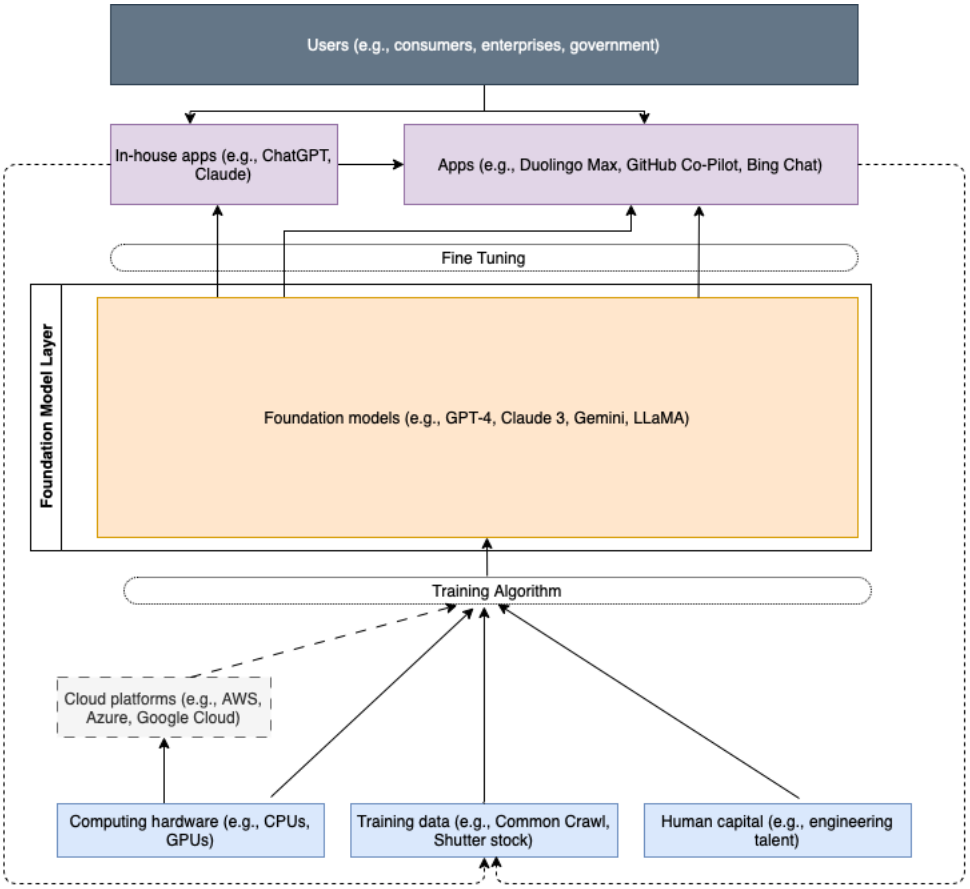
The apps layer is divided into in-house apps (e.g., ChatGPT or Claude) and third-party apps. Third-party apps are those that rely on an underlying foundation model but are not produced by

[43] The number of foundation models and organizations is calculated as the number of unique models and organizations (including groups of organizations). To avoid double counting, a fuzzy deduplication method was used to remove different generations of the same model.

[44] Open and closed models may cater to different users. For instance, developers using a pre-trained foundation model to create an application might prioritize the adaptability and customization of open source, while other users might prioritize user-friendliness and the ease of use of closed models. Some foundation model producers also produce in-house end use applications (e.g., ChatGPT), allowing for vertical expansion within the broader market for AI capabilities. Alternatively, foundation model developers may license their models to application developers (e.g., OpenAI licenses GPT-4 to Microsoft), creating a market for foundation models as a service.

the same entity that produces the foundation model. For instance, language education application Duolingo's Duolingo Max app augments the company's existing software by integrating OpenAI's GPT-4 as the underlying foundation model.

Figure 3.1. Notional Depiction of AI Ecosystem Stack



NOTE: AWS = Amazon Web Services, CPU = central processing unit, GPU = graphics processing unit.

In the following subsections, we discuss the key inputs that flow into the pre-trained foundation model layer in greater detail. These key inputs are those below the foundation model layer in Figure 3.1 and represent factors that drive costs of producing a pre-trained model. We also discuss other costs associated with developing a foundation model, such as the cost of labor.

Compute

Computing resources—which encompass processing hardware, servers, and networking equipment—are critical to foundation model development and operation. The compute needed

for frontier foundation model training is substantial.[45] Key computing resources include graphics processing units (GPUs) or tensor processing units (TPUs), which facilitate parallel operations and are crucial for AI workloads; central processing units (CPUs), which are tasked with interpreting and executing instructions; and random access memory (RAM), responsible for storing intermediate computations. Finally, networking equipment enables models to leverage distributed computing and train across multiple machines.

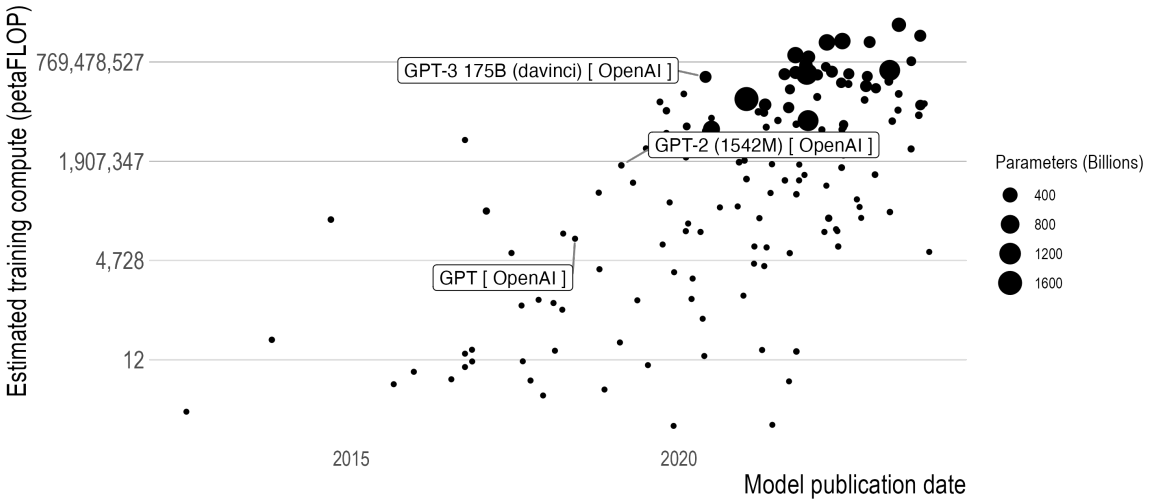
Figure 3.2 displays the growth in training compute in floating-point operations (FLOP) for approximately 150 large language foundation models over time. The data in the figure come from Epoch's database on foundation model attributes.[46] There is a clear upward trend over time in the compute used for model training. The figure's y axis is in log scale, meaning that a growth trend over time that appears linear in the plot would actually represent exponential growth. For instance, there was an 84-fold increase in the amount of training compute used between GPT and GPT-2 and a 201-fold increase in training compute between GPT-2 and GPT-3. On average, across the models in the figure, the amount of training compute used has increased by 485 percent per year.[47]

[45] Jordan Hoffman, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al., "Training Compute-Optimal Large Language Models," *DeepMind*, March 29, 2022.

[46] Epoch, undated.

[47] This is estimated from the coefficient that results from estimating log compute on a yearly time trend. The raw coefficient estimate is 1.77 and is translated into a percentage as $100 * (e^{1.77} - 1)$. The results are quantitatively similar when controlling for the number of parameters.

Figure 3.2. Training Compute over Time



SOURCE: Authors' calculation based on data from Epoch (Epoch, undated).
NOTE: This figure shows the estimated training compute (in petaFLOP) for models contained in the Epoch dataset (Epoch, undated). The x axis shows the publication date of the model. Points are sized by the number of parameters in the model (in billions). Three OpenAI models are highlighted as examples.

Compute Acquisition and Training

Compute resources can either be purchased directly or rented from a cloud provider. Renting resources from a provider offers more flexibility and less upfront cost. However, for organizations operating at a large scale and developing very large models, acquiring compute resources directly may be more cost-effective.[48]

Direct costs include the cost of acquisition, which involves purchasing GPUs directly from a supplier (e.g., Nvidia) or designing GPUs in house and paying for fabrication (e.g., from Taiwan Semiconductor Manufacturing Company Limited [TSMC]), as well as the cost of operation (i.e., storing and powering GPUs in a datacenter).

Direct acquisition of compute can be expensive. For instance, the Nvidia A100 GPU was reported in February 2023 to cost $10,000 per chip.[49] The successor to the A100, Nvidia's H100, which was released globally in 2023, is reported to be priced at $30,000 per chip, on average.[50] GPU demand has outpaced supply, and GPU shortages reportedly stalled OpenAI's deployment of its multimodal model.[51] Meta has announced plans to purchase more than 340,000 of Nvidia's H100 GPU. Assuming that each H100 costs $30,000, Meta's purchase could come to

The Compute Market Influences the Foundation Model Market

The market structure of the foundation model market is closely tied to that of the high-performance compute market, on which these models rely. The compute market is highly concentrated, with a small number of providers producing the high-performance logic chips required to train large models. For instance, in the second quarter of 2023, Nvidia had more than 70 percent of the market share in advanced AI chips (Don Clark, "Nvidia Revenue Doubles on Demand for A.I. Chips, and Could Go Higher," *New York Times*, August 23, 2023).

Several foundation model developers (e.g., Google, Meta) and organizations with ties to developers (e.g., Microsoft) also own substantial compute resources. This vertical integration allows these developers to avoid cloud compute rental markups and amortize the fixed cost of their compute infrastructure across multiple models. Google and (as of April 2024) Amazon and Microsoft plan to do the same (design their own compute), which allows them to reduce their reliance on vendors like Nvidia.

Vertical integration between large foundation model developers and compute owners will likely raise barriers to entry and could result in a less competitive foundation model market.

[48] Guido Appenzeller, Matt Bornstein, and Martin Casado, "Navigating the High Cost of AI Compute," *Andreessen Horowitz*, April 27, 2023.

[49] Kif Leswing, "Meet the $10,000 Nvidia Chip Powering the Race for A.I.," *CNBC*, February 23, 2023.

[50] Doug Eadline, "Nvidia H100: Are 550,000 GPUs Enough for This Year?" *HPCwire*, August 17, 2023.

[51] Dylan Patel, Myron Xie, and Gerald Wong, "AI Capacity Constraints—CoWoS and HBM Supply Chain," *Semianalysis*, July 5, 2023.

over $10 billion.[52] Epoch estimates that training costs for Google's Gemini Ultra model were approximately $630 million.[53]

On top of purchasing chips, operating chips in a datacenter has associated costs as well. For instance, estimates suggest that creating GPT-3, a model released by OpenAI in 2020, required 1,287 megawatt hours of electricity.[54] According to data from the Bureau of Labor Statistics, the average energy price in the United States was $0.14 per kilowatt hour in 2019. Assuming that GPT-3 was trained in 2019, the energy cost comes to approximately $176,000.[55] Relative to the cost of compute acquisition, energy costs associated with training GPT-3 seem to be relatively insignificant.[56] However, the energy costs associated with model inference could eventually grow to be large.

Most computing resources can be rented from cloud providers, allowing users to scale their usage up or down over time, based on their needs. However, most larger users negotiate cloud costs with providers and typically commit to a minimum time requirement, locking them into a longer-term contract.[57] While renting compute requires less upfront cost than purchasing it directly, rental costs can still be relatively high. For instance, Stable Diffusion, an image-generating foundation model produced by Stability AI, was trained on 256 A100s rented from Amazon Web Services (AWS).[58] Stability AI's chief executive officer suggested in a tweet that the market price of training compute for Stable Diffusion was $600,000.[59] Unlike direct purchase of compute, renting compute does not allow foundation model developers to amortize this fixed cost across multiple models or rent it out to other users. Due, in part, to the high costs and lack of flexibility of cloud rental agreements, some large foundation model producers have partnered

---

[52] Michael Kan, "Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs," *PC Magazine*, January 18, 2024.

[53] Epoch, "Machine Learning Trends," webpage, 2023.

[54] David Patterson, Joseph Conzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean, "Carbon Emissions and Large Neural Network Training," arXiv, arXiv:2104.10350, April 21, 2021.

[55] U.S. Bureau of Labor Statistics, "Databases, Tables & Calculators by Subject," webpage, undated.

[56] While the electricity costs of model training are relatively small compared with other costs, the carbon emissions resulting from multiple organizations all training their individual foundation modes could sum to a meaningful amount. When considering whether the market for foundation models has characteristics of a natural monopoly, the social cost of the inefficiencies that stem from multiple organizations producing a similar product (a pre-trained model) may manifest as increased carbon emissions.

[57] Appenzeller, Bornstein, and Casado, 2023.

[58] Robin Rombach, Patrick Esser, and David Ha, "High Resolution Image Synthesis with Latent Diffusion Models," Hugging Face, June 2022.

[59] Emad [@EMostaque], "We actually used 256 A100s for this per the model card, 150k hours in total so at market price $600k," post on the X platform, August 28, 2022.

with datacenter and compute providers to limit the costs of compute acquisition.[60] For instance, OpenAI is designing its own chips with Microsoft and also has access to Microsoft's Azure Cloud. Similarly, Anthropic is designing its own chips with Amazon and has access to AWS. Google develops its foundation models using its own datacenters and chips.

Compute for Inference

Compute is also required when using foundation models. Each call of the model, referred to as an inference, requires computing resources and has a monetary cost. Data on inference compute is not readily available, though several studies have suggested rules of thumb in calculating inference compute costs. For instance, Desislavov, Martinez-Plumed, and Hernandez-Orallo (2023) found that the compute FLOP required to make a single inference of a transformer-based model was roughly the same as the number of model parameters.[61] Others suggest that a good rule of thumb for estimating inference FLOP is 2*n*p, where n is number of tokens in the input (e.g., the user's question) and output (e.g., the model's answer) and p is the number of parameters.[62] Villalobos and Atkinson (2023) suggest that a good rule of thumb is that the compute of an inference is approximately equal to the square root of the training compute of the model, due to the quadratic cost of transformers.[63] However, the expenditure on inference compute is significantly lower than the expenditure on training compute. On a per token basis, training a transformer-based model takes approximately three times as long as performing inference. Because training datasets are significantly larger than the average inference prompt, training takes significantly longer than inference, resulting in greater costs.[64]

Data

Foundation model developers use data from a variety of sources, including web-scraped public websites, public and private databases, and synthetic data.[65] Unlike compute and labor inputs, data are a non-rival input, meaning that the same data can be used by multiple foundation model producers at the same time.

---

[60] Microsoft, Amazon, Nvidia, and Google (MANG) provided an estimated $25 billion in investment funding to emerging AI companies in 2023 (Apoorv Agrawal, "New VC in Town: 'MANG,'" *Apoorv's Notes* blog, January 18, 2024).

[61] Radosvet Desislavov, Fernando Martinez-Plumed, and Jose Hernandez-Orallo, "Compute and Energy Consumption Trends in Deep Learning Inference," arXiv, arXiv:2109.05472v2, March 29, 2023.

[62] Appenzeller, Bornstein, and Casado, 2023.

[63] Pablo Villalobos and David Atkinson, *Trading Off Compute in Training and Inference*, Epoch, July 28, 2023.

[64] Appenzeller, Bornstein, and Casado, 2023.

[65] Samir Sampat, "Where Do Generative AI Models Source Their Data & Information?" *Smith.ai*, September 20, 2023; Adam Zewe, "In Machine Learning, Synthetic Data Can Offer Real Performance Improvements," *MIT News*, November 3, 2022.

The largest training dataset used, as of 2023, consists of 1.87 trillion words, which is approximately 20 percent of the text data on the internet.[66] In general, the greater the volume, variety, and quality of the data, the greater the capabilities of the model.[67] A larger volume of data with greater variety can contribute to better model performance by presenting the model with greater sources of variation from which to learn patterns. At the same time, data quality is also critical, as high-quality data enable improved overall model performance and more accurate fine-tuning.

Some research suggests that current models may have already used most of the high-quality data available on the internet. For instance, Villalobos et al. (2022) project that the stock of high-quality text data used to train and tune large language foundation models will be exhausted by 2026 or earlier.[68] Other research suggests that, given a compute budget, optimal model training requires that training tokens (approximately 0.75 words, in English) should grow in proportion with model size (parameters).[69] Thus, using common training algorithms, more compute-intensive and potentially more-capable models could require more data. Algorithmic improvements and incorporating synthetic data in model training could alleviate this constraint.

Algorithms and the Transformer

Foundation models are developed by performing algorithms on data using computing resources. Algorithms are procedures for manipulating data and include procedures used to train and run inference using a model. Furthermore, the model itself is a kind of algorithm. Most foundation models, as of January 2024, are based on the transformer, which is a particular algorithm for learning relationships in sequential data (i.e., words in a sentence, pixels in an image, etc.).[70] The transformer was discovered in 2017 and has been responsible for an incredible series of advancements. These advancements have been made possible by three features of the algorithm: (1) It can be easily adapted to different domains; (2) it can learn from large, unlabeled datasets; and (3) it can be trained using large amounts of parallel computing, letting it take advantage of huge clusters of AI accelerators.[71]

---

[66] Epoch, 2023.

[67] Christophe Carugati, *Competition in Generative AI Foundation Models*, SSRN, Working Paper 14/2023, September 18, 2023.

[68] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho, "Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning," arXiv, arXiv:2211.04325v1, October 26, 2022.

[69] Hoffman et al., 2022.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," arXiv, arXiv:1706.03762, June 12, 2017.

[71] Jonathan Gillham, "What Are Transformer Models—How Do They Relate to AI Content Creation?" *Originality.ai* blog, undated.

Training

During model training, algorithms process data to teach the model the underlying patterns and relationships within it, and this process requires substantial computational resources. Researchers have discovered predictable relationships—termed *scaling laws*—relating compute expenditure and model performance.[72] Scaling laws suggest that model performance scales as a power law with model size, dataset size, and the amount of compute used for training. Model performance is primarily influenced by its size, while the impact of its architecture, or the hyperparameters' shape, is relatively less significant.[73] This relationship has spurred enormous investments in training transformer-based models using ever more computing resources as competitors in the foundation model market compete over model performance.

Improvements in algorithmic efficiency have contributed to the declining cost of training compute required to achieve a given level of model performance.[74] For instance, research finds that from 2012 to 2019, training a model to achieve AlexNet-level (a benchmark in image recognition) image classification performance required 97.7 percent less computing power due to algorithm improvements.[75] This trend has continued, with more-recent studies finding that compute requirements are now halving every eight to nine months.[76]

Inference

Model size is particularly relevant in the context of inference because it partially determines the cost of inference,[77] where the model can be deployed, and the latency of inference.[78] The cost of inference depends on model size because the amount of compute used is proportional to both model size and cost. Model size is also proportional to the amount of memory required to run the model, which, in turn, restricts the settings in which the model can be used. Additionally, the latency of inference depends on model size. The forward pass, which must partially be

---

[72] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling Laws for Neural Language Models," arXiv, arXiv:2001.08361, January 23, 2020.

[73] Fabio Chiusano, "Two Minutes NLP—Scaling Laws for Neural Language Models," *Medium*, March 18, 2022.

[74] Konstantin Pilz, Lennart Heim, and Nicholas Brown, "Increased Compute Efficiency and the Diffusion of AI Capabilities," arXiv, arXiv:2311.15377v2, February 13, 2024.

[75] Danny Hernandez and Tom B. Brown, "Measuring the Algorithmic Efficiency of Neural Networks," arXiv, arXiv: 2005.04305, May 8, 2020.

[76] Ege Erdil and Tamay Besiroglu, "Algorithmic Progress in Computer Vision," arXiv, arXiv:2212.05153, December 10, 2022; Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla, "Algorithmic Process in Language Models," arXiv, arXiv:2403.05812, March 9, 2024.

[77] Nikhil Sardana and Jonathan Frankle, "Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws," arXiv, arXiv:2401.00448v1, December 31, 2023.

[78] DeepSpeed Team, Rangan Majumder, and Andrey Proskurin, "DeepSpeed: Accelerating Large-Scale Model Inference and Training via System Optimizations and Compression," *Microsoft Research Blog*, May 24, 2021.

performed sequentially, takes longer to perform when the model has more layers.[79] Latency has significant economic implications because some applications may require very low latency.

New methods and processes are emerging that could further reduce the positive relationship between compute and model performance. For instance, the Alternating Updates (AltUp) method has been shown to increase transformer model capacity without increasing the computational burden of the model.[80] Other studies have proposed successors to the transformer, like the retentive network[81] and Mamba[82], which have lower inference compute costs than the transformer.

Labor

Building foundation models requires highly specialized expertise in domains like machine learning (ML), data engineering, and compute optimization. Skilled labor is essential for tasks ranging from developing initial model architecture to ongoing maintenance and improvements. Beyond technical roles, human capital in such areas as ethics, AI oversight, communications, and legal policy analysis may be required to ensure that models comply with corporate ethics codes or government regulations.

Attracting and retaining workers for these roles can be expensive. There are reports that organizations producing cutting-edge foundation models often attempt to poach workers from their competitors.[83] Additionally, given the close physical proximity of many organizations producing foundation models, local labor market churn results in workers moving between employers frequently. For instance, according to data from Lightcast, which collects information on worker job histories from public worker profiles like LinkedIn, 104 of the roughly 800 past and current OpenAI employees worked at Google, Meta, Amazon, or DeepMind in the past five years.[84] This is likely an underestimate, as not all employees at these companies have public professional profiles.

Public information on employment within individual foundation model producers is not available. However, data on job postings from these companies can provide some insight into the

---

[79] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean, "Efficiently Scaling Transformer Inference," arXiv, arXiv:2211.05102v1, November 9, 2022.

[80] Xin Wang and Nishanth Dikkala, "Alternating Updates for Efficient Transformers," *Google Research* blog, November 7, 2023.

[81] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei, "Retentive Network: A Successor to Transformer for Large Language Models," arXiv, arXiv:2307.08621v4, August 9, 2023.

[82] Albert Gu and Tri Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," arXiv, arXiv2312.00752, December 1, 2023.

[83] Christine Hall, "Chaos at OpenAI Adds Fuel to the AI Talent Poaching War," *TechCrunch*, November 20, 2023.

[84] Lightcast, homepage, undated. We arrived at these numbers by analyzing Lightcast data on OpenAI job transitions. Note that this is an incomplete list of transitions between foundation model producers because Lightcast only collects data from public worker profiles.

relevant occupations these organizations employ. Of course, an employer may hire multiple workers from the same job posting, so job posting data do not directly translate to number of workers. Additionally, employers may hire some workers outside of the typical labor market, through word of mouth or internal promotions, which might not be reflected in the available job postings data.

We collected data on job postings between 2020 and 2022 from Lightcast.io on the following foundation model producers (number of unique job postings over the time period): Hugging Face (238), OpenAI (348), Stability AI (29), Anthropic (237), Cohere (45), Deepmind (11), and Inflection (5).[85] These foundation model producers have created some of the most popular foundation models, according to analysis by CRFM.[86] Job postings from other organizations, such as Meta, were excluded from this analysis due to the sheer size of their workforce, which makes it difficult to determine whether a job posting is specific to foundation model development or other operations. Lightcast translates the specific job title in the posting to a harmonized occupation code used by the Department of Labor's Occupational Information Network (O*NET) database, which facilitates comparison of labor demand over time across employers.

Overall, there were 913 unique job postings from these seven foundation model organizations collected in the Lightcast database. The number of unique postings grew from 65 in 2020 to 535 by 2022. Table 3.1 shows the top ten occupations in terms of share of total 2022 job postings. The table also shows the relevant shares for these occupations in 2020 and 2021, as well as the percentage point change in the share between 2022 and 2020. Software developers (a category that includes software engineers) make up the largest share each year and also had the second largest increase in job postings between 2020 and 2022.[87] The largest increase, five percentage points, was in job postings for architectural and engineering managers.[88] Data scientists, human resources specialists, computer system engineers and architects, sales representatives, lawyers, web developers, other engineers, and other managers make up the top ten occupation postings by these firms in 2022.[89]

---

[85] Lightcast, undated.

[86] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Masleg, Betty Xiong, Daniel Zhang, and Percy Liang, "The Foundation Model Transparency Index," *GitHub*, October 2023.

[87] O*NET lists a sample of job titles that fall under the software developer occupation: application developer, application integration engineer, developer, infrastructure engineer, network engineer, software architect, software developer, software development engineer, software engineer, and systems engineer.

[88] O*NET lists a sample of job titles that fall under the architectural and engineering managers occupation: civil engineering manager, electrical engineering manager, engineering director, engineering group manager, engineering program manager, mechanical engineering manager, process engineering manager, project engineering manager, and project manager.

[89] O*NET describes the data scientist occupation as workers who "develop and implement a set of techniques or analytics applications to transform raw data into meaningful information using data-oriented programming

Table 3.1. Job Postings by Occupation for Select Foundation-Model-Producing Organizations

Occupation	Percentage in 2020	Percentage in 2021	Percentage in 2022	Change Between 2022 and 2020 (percentage point)
Software developers	12.3	16.6	16.6	4.3
Architectural and engineering managers	3.1	1.5	8.4	5.3
Data scientists	4.6	16.0	6.9	2.3
Human resources specialists	9.2	8.3	4.7	–4.6
Computer systems engineers/architects	3.1	2.5	4.5	1.4
Sales representatives	3.1	2.1	4.1	1.0
Lawyers	1.5	0.9	4.1	2.6
Marketing managers	3.1	3.4	3.9	0.8
Web developers	0.0	8.9	3.4	3.4
Engineers, all other	4.6	11.4	3.2	–1.4

SOURCE: Authors' calculations from Lightcast.io data (Lightcast, undated).
NOTE: This table shows the share of total unique job postings from Hugging Face, OpenAI, Deepmind, Inflection AI, Stability AI, Anthropic, and Cohere between 2020 and 2022. The table shows the top ten occupations based on the 2022 share. The last column shows the change in the share for the occupation between 2022 and 2020 in percentage points.

As noted above, while job posting data do not reflect total employment by occupation, they do suggest the types of roles that are relevant to foundation model producers. However, job postings may be thought of as a signal of labor demand, and the job posting data for the seven foundation model organizations suggest that software developers, engineering managers, and data scientists are in high demand.

Apart from job posting data, research report authorship also reveals the scale and scope of employment at foundation model organizations. Many organizations publish research papers that discuss their models, often with lists of authors who made significant contributions to the development of the model. For instance, OpenAI released a technical report discussing GPT-4 with 280 authors, of which 32 contributed to pre-training.[90] Google's technical report for the

languages and visualization software. Apply data mining, data modeling, natural language processing, and machine learning to extract and analyze information from large structured and unstructured datasets. Visualize, interpret, and report data findings. May create dynamic data reports."

[90] OpenAI, "GPT-4 Technical Report," arXiv, arXiv:2303.08774v4, December 19, 2023.

Pathways Language Model (PaLM) 2 model lists 38 people who contributed to model pre-training tasks.[91]

Data on salaries at foundation model organizations are not available from typical sources. However, Levels.fyi, a database of technology salary compensation, provides estimates of the average annual compensation for software engineers at OpenAI. According to Levels.fyi data, the average level 4 software engineer earns $245,000 (excluding stock options) per year, and a level 5 engineer earns $303,000 (excluding stock options).[92] Assuming that all 32 authors of the GPT-4 technical report are either Level 3 or Level 4 software engineers, the annual labor costs (excluding fringe) come to $7.8 and $9.7 million, respectively. Of course, this is a rough approximation, as not all 32 authors are software engineers, not all are Level 4, and some are members of OpenAI leadership (e.g., Greg Brockman) who likely make salaries significantly higher than those listed in Levels.fyi data.

Foundation Model Costs Classification

Given the emphasis on cost functions as a primary determinant of whether a natural monopoly exists, we categorize the input factors discussed in the prior section into two types of costs associated with producing AI foundation models: fixed and variable costs. Fixed costs are those that do not change with the level of output from the model. They are typically incurred regardless the level of output, and the quantity of resources required is determined primarily before the model is developed. Variable costs change with the level of model output and are usually incurred when the model is used.

Defining output in the context of pre-trained foundation models is complicated because these models are highly versatile and adaptable. Unlike other products that have a fixed function or purpose, foundation models can be used for various tasks and in multiple domains, depending on how they are trained and applied. These models can also be fine-tuned or adapted to specific domains or datasets. This implies that the output of a foundation model is not a single product, but a variety of potential products.

The typical approach to modeling the costs of foundation model development involve treating inference as the variable output of the model and the compute, energy, and resources used for inference as the variable cost.[93] We adopt this framework here as well, with the caveat that pre-trained models are generally not used for inference at a large scale. Most users who will

---

[91] Task categories include large model training, pre-training data and mixture workstream, and code pre-training workstream (Andrew M. Dai, David R. So, Dmitry Lepikhin, Jonathan H. Clark, Maxim Krikun, Melvin Johnson, Nan Du, Rohan Anil, Siamak Shakeri, Xavier Garcia, et al., "PaLM 2 Technical Report," Google, May 17, 2023).

[92] *Levels* refer to the hierarchy of workers within an occupation at an employer. In the case of OpenAI, a Level 3 software engineer is the entry level (Levels.fyi, "OpenAI Software Engineer Salaries," webpage, undated).

[93] Anton Korinek and Jai Vipra, "Concentrating Intelligence: Scaling Laws and Market Structure in Generative AI," working paper, February 28, 2024.

eventually interact with the foundation model will do so with a fine-tuned model, which we consider a separate product from a pre-trained model.[94] While the pre-trained model forms the core of the foundation model, it is typically not the version of the model that end users interact with directly. Instead, users engage with models that have undergone additional fine-tuning. This fine-tuning can include domain-specific training, as well as alignment processes to improve the model's truthfulness and reduce harmful outputs. As a result, the final consumer-facing model is often a specialized version of the foundation model, distinct from the initial pre-trained models.

The discussion in this chapter highlights the outsized role that fixed costs, particularly the acquisition of training compute, play in the firm's cost function. As a result, the firm's total cost is dominated by the fixed costs. As the number of inferences run on the model grow, the variable cost can become substantial.[95] However, in the context of pre-trained models, inference calls are likely to be relatively few. As a result, the variable costs associated with model development are likely substantially smaller than the fixed costs of development.

Table 3.2 shows the various cost factors, the cost type (fixed and variable), and a short description of the cost factor. In general, the majority of the costs associated with model training are fixed because many of these costs must be incurred before the model is trained. For instance, training a model requires sufficient training compute, networking, and storage capacity. Compute acquisition requires either building a datacenter with advanced chips or reserving cloud computing resources for an extended period of time, both of which lock developers into a certain level of compute.

Model training also requires curating training data that are sufficiently large and representative for the task. In theory, data can be scaled up or down depending on the level of output; however, acquiring some datasets may involve fixed licensing fees. While the quantity of training data can be increased, accessing proprietary datasets requires upfront payments regardless of usage levels. Additionally, given the importance of a large datasets in model training, a minimum viable dataset matching the model's intended use cases is likely needed before the model can be trained.

---

[94] With this framework in mind, assuming Cobb-Douglas production, the firm's total cost function can be represented as follows:

$$C(q) = r\left(\frac{q}{\overline{K}^{\beta}\overline{D}^{\gamma}\overline{R\&D}^{\xi}}\right)^{\frac{1}{\alpha}} + a_K\overline{K} + a_D\overline{D} + a_{R\&D}\overline{R\&D}$$

where q is an inference call of the model and r is the variable cost of the inference (primarily compute and energy costs). Inputs with an overbar are assumed to be fixed: $\overline{K}$ is the fixed compute used to train the model, which was acquired at cost a_K; $\overline{D}$ is the data inputs used to train the model, which have acquisition cost a_D; and $\overline{R\&D}$ is the fixed amount of labor R&D required to train the model, which is acquired at wage $a_{R\&D}$. The parameters α, β, γ, and ξ are coefficients of the Cobb-Douglas production function assumed to describe production. The Cobb-Douglas production function that underlies the total cost function implicitly assumes that there is some degree of substitution between variable inputs (e.g., inference compute) and fixed inputs (e.g., R&D).

[95] Patterson et al., 2021.

R&D activities, such as developing specialized techniques and model architectures, are likely primarily conducted prior to model training. This work requires a workforce that includes researchers who work on algorithmic improvements and engineers who help implement the model training process. While some incremental innovation may be required during the training process, the majority of this work is completed upfront and must be done regardless of the model's eventual use. R&D costs, which mainly consist of labor expenses, are classified as fixed in Table 3.2 to account for the fact that the workforce needed to train the model likely does not scale with each inference of the model.

Lastly, the variable costs of producing a pre-trained model involve the costs associated with storing the model weights, distributing the model to other developers, electricity, and, primarily, the compute used to run inference on the model. Inference costs scale with model use as well as model size, though inference compute costs are significantly smaller than the acquisition of compute for training.

Table 3.2. Cost Factor Categorization for Pre-Trained Models

Cost Factor	Cost Type	Description
Compute acquisition	Fixed	The cost of purchasing or reserving the computing resources needed to train and run the model
Training data acquisition	Fixed	The cost of acquiring the data needed to train the model
R&D	Fixed	The labor cost of developing algorithms or model architecture
Inference compute	Variable	The costs of compute used per model inference

NOTE: This table shows the relevant input factors required for building a pre-trained foundation model. Each input factor is classified as a fixed, semi-fixed, or variable cost based on its relationship to the production of a pre-trained model.

Chapter 4. Adapting the Natural Monopoly Criterion to Foundation Models

Chapter 2 set forth a generic set of criteria for defining a natural monopoly. Chapter 3 described the market and production conditions for foundation models. This chapter adapts the natural monopoly criterion to the particular economic features of foundation models, describing how indicators of natural monopoly would manifest in practice, so that we can apply the natural monopoly criteria to real and hypothetical markets in Chapter 5. Table 4.1 presents the set of criteria defined in Chapter 2, adding indicators specific for the foundation model market. The more of the indicators that are observed in the focal market, the stronger the case for a natural monopoly.

Table 4.1. Criteria for Natural Monopoly in the Context of Pre-Trained Foundation Models

Criteria	Definition	Indicators
Product homogeneity	A homogeneous product is one that is perfectly substitutable: A consumer will be indifferent to the exchange of one version for another one and will base purchasing decisions solely on price.	Limited product variation (at pre-trained foundation model level)
Economies of scale	Economies of scale exist when the average cost of production decrease when its production increases.	High cost of training relative to variable costs
Sunk cost	Sunk costs exist when the value of an investment in an asset is worth less than an alternative use of the investment.	Low resale value of initial compute, data, labor, and foundation model
Network effects	Network effects exist when the utility of a good or service increases as the number of users grows (i.e., with scale).	Community-led development
Economies of scope	Economies of scope exist when the total cost of producing two or more distinct products within a single firm is less than the total costs of producing those products within more than one firm.	Firms have multiple products using common foundation model

Product Homogeneity in Foundation Models

In this analysis, the market under scrutiny is defined as the market for pre-trained foundation models. The processes associated with fine-tuning—while critical to downstream applications of foundation models—is thus outside of the scope of analysis.

Homogeneity in this context refers to a scenario in which developers, acting as the primary users of these models, experience negligible differences in performance when switching between various foundation models of the same generation. This characteristic is pivotal because it underscores limited differentiation in the market. Measuring the change in satisfaction from switching between models is infeasible, but evidence of product homogeneity can be found by considering features of the models themselves. Evidence of homogeneity might include similar technical characteristics (e.g., model architecture), similar performance, similar cost of downstream differentiation (e.g., fine-tuning), or competition among foundation model developers based on cost.

Economies of Scale and Foundation Models

Recall that economies of scale depend on the distribution of a firm's costs that are fixed versus those that vary as a function of output, and that high fixed costs and low variable costs yield economies of scale. In the context of a foundation model, the fixed costs are training compute acquisition, training data acquisition, algorithm development, and some aspects of labor costs. There are very little variable costs associated with pre-trained model production, as an individual inference of the model is inexpensive.

To gauge whether economies of scale are present in the market for foundation models, we produce estimates of one of the largest fixed costs: the compute used to train a new model.[96] Because variable costs are significantly smaller than fixed costs, and we do not expect this to change over the forecast window, we assume that they will remain relatively constant, and therefore we do not produce cost estimates.

Sunk Costs and Foundation Models

Recall that fixed costs are sunk to the extent that they cannot be easily recovered. To the extent that the fixed costs of model development cannot be recovered in the event of bankruptcy or a failed training run, they will be considered, in the analysis to follow, to be sunk. Furthermore, if the probability of attempting and failing to develop a foundation model is high and costs are sunk, market contestability will suffer because would-be entrants may avoid entry due to the expectation of sunk costs.

---

[96] Using the cost of compute as a stand-in for total fixed costs of foundation model development is a simplification due to limited public data on the internal costs of model producers. While a complete breakdown of all fixed-cost components is challenging due to these limitations, compute acquisition costs are substantial and likely to represent a significant portion of total fixed costs. Other fixed costs, such as labor R&D costs and data acquisition, also contribute but are harder to quantify and project into the future. The rapid growth in training compute requirements over time underscores the significance of this cost component. As more research on the cost structures of foundation model producers becomes available, future research could provide a more granular breakdown of the fixed-cost components.

Network Effects and Foundation Models

Recall that network effects occur when the value of a product increases as more people use the product. In the case of a pre-trained foundation model, network effects are most likely to come by way of model improvements to open-source models via a user and development community.[97] That is, as they gain users, open-source models broaden the developer base beyond that assigned by the original firm. As the user base (at the level of the foundation model, users will largely be developers) grows, so too do opportunities for improvement to model performance and model safety. In explaining the decision to publicly release Llama 2, Mark Zuckerberg echoed this rationale, stating, "Open-source drives innovation because it enables many more developers to build with new technology."[98]

These effects are unlikely to be present in closed-source models because within such models, the development community does not scale with model usage.

Economies of Scope and Foundation Models

Economies of scope occur when a firm realizes a cost advantage from producing multiple products using a common input. If a firm's foundation model can be used in more than one of that firm's final products, economies of scope will be present. The size of economies of scope will depend on how widely diffused the foundation model is across a firm's final products and the proportion of each of a firm's final products' total production cost that is occupied by the foundation model. Economies of scope require that the multiple products into which a common input are incorporated be produced by the same firm that produces the common input. Cases in which a foundation model developer licenses its model to another firm which then uses the model as the basis for its own products do not constitute economies of scope. Economies of scope will be measured by examining the extent to which firms produce distinct products that use a common foundation model that was developed by that firm.

---

[97] While not relevant to pre-trained models, models that interact with end users (e.g., chatbots such as ChatGPT) have the potential to realize data network effects, which occur when the data collected from a product's users improve the product in a way that increases the product's utility to the user (Robert Wayne Gregory, Ola Henfridsson, Evgeny A. Kaganer, and Harris Kyriakou, "The Role of Artificial Intelligence and Data Network Effects for Creating User Value," *Academy of Management Review*, March 2021). While the mechanism is different, the basic relationship between scale and demand in network effects and network effects is identical: user growth for a product begets increased utility. In the case of language models, data network effects will be high when the returns to model performance from user data (e.g., dialog data and preference data) are high. The additional user value associated with data network effects can manifest in either improved overall model performance or personalized improvement to the user providing data (Gregory et al., 2021).

[98] Meta, "Let's Get Building!" Facebook post, July 18, 2023.

Chapter 5. Is the Market for Foundation Models a Natural Monopoly?

Having put forth a natural monopoly criterion and having described how this criterion can be applied to the market for foundation models, in this chapter we apply the criterion to five scenarios. We first apply the criterion to the status quo market, answering the following question: Is the current foundation model market a natural monopoly? We then apply the criterion to four hypothetical scenarios, in which each scenario assumes a distinct technological future set three years hence (2027).[99] Specifically, in imagining the future scenarios, we vary two technology variables critical for determining the foundation model cost structure: the scaling hypothesis and the computing performance/cost.

Is the Status Quo Market a Natural Monopoly?

Product Homogeneity in the Status Quo Market

Large, pre-trained, state-of-the-art foundation models exhibit pronounced homogeneity, primarily due to the widespread adoption of similar model architectures and training data across frontier foundation models.[100] Models such as LLaMA,[101] GPT-4,[102] and Gemini[103] share the transformer architecture as their backbone. Widespread adoption stems from the transformer's proven efficiency in processing sequential data, making it the architecture of choice for natural language processing (NLP) tasks.[104] This convergence on a single model architecture significantly limits opportunities for meaningful differentiation at the architectural level.

---

[99] We chose a three-year forecasting window because it seemed to strike a reasonable balance between the increased confidence in short-term forecasts and the growth in uncertainty associated with long-term forecasting. Extending the forecasts further into the future would decrease our confidence in our assumptions regarding the persistence of current trends, which we use to calculate the forecasts. Selecting too short of a forecast window would limit our ability to discuss how the temporal change in compute would affect the market for foundation models.

[100] Interestingly, the theoretical potential for variability introduced by the random initialization of model weights, which could enable divergent learning pathways even among models with identical architectures and data, does not translate into significant performance disparities in practice.

[101] Touvron et al., 2023.

[102] OpenAI, 2023.

[103] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, et al., "Gemini: A Family of Highly Capable Multimodal Models," *Google DeepMind*, December 19, 2023.

[104] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled, "Overview of the Transformer-Based Models for NLP Tasks," *IEEE*, 2020.

Compounding this is the reliance on overlapping, internet-based datasets for model training. Common sources, including the Common Crawl and the Pile, are widely used—leading to a convergence in foundational knowledge. Although there is a growing effort among developers to enrich training processes with a broader array of data sources, including synthetic data, the performance gap between leading models remains surprisingly narrow.[105] This suggests that the variation in data has yet to confer a significant competitive advantage to any single model developer. Thus, the current performance similarities among models highlight the limited impact of data diversity on model differentiation at the frontier.

Nevertheless, it is important to recognize certain distinctions that exist among models—such as differences in supported context lengths.[106] Variability in the length of context supported, which can affect a model's utility in specific applications and its performance in tasks requiring different data sequence lengths or computational efficiency, suggest a level of heterogeneity. However, these differences are relatively minor and do not suffice to classify foundation models as heterogeneous goods.

Economies of Scale in the Status Quo Market

Economies of scale in the status quo market appear to be high given the high fixed cost and low variable cost associated with pre-training a state-of-the-art foundation model. While definitive pre-training cost information is not available, Epoch estimates that the cost to train Google's Gemini Ultra was roughly $630 million.[107] Apart from fixed costs, the variable costs of a pre-trained model primarily stem from the cost of compute associated with inference. However, the focus on pre-trained models limits the relevance of inference costs for this analysis. Some components of labor costs are also variable, as organizations may need to scale up effort as more downstream developers and organizations use the pre-trained model for downstream uses. Overall, the variable costs of developing a pre-trained model are relatively low.

It is worth noting that while the cost of pre-training a large state-of-the-art foundation model may be prohibitive to a small start-up company, it represents less than 3 percent of the individual R&D budgets of Alphabet, Meta, Amazon, Microsoft, and Apple, based on September 2023 R&D reporting.[108] This suggests that these companies, hypothetically, have the funding to produce foundation models at the same level as Gemini, at least in terms of model training. Table 5.1 displays the annual R&D funding of these firms over the last four years.

---

[105] LMSYS Chatbot Arena Leaderboard, webpage, undated.

[106] Gemini Team and Google, "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context," *Google DeepMind*, 2024.

[107] Epoch, 2023.

[108] FactSet, homepage, undated. Other state-of-the-art foundation models cost significantly less to train. It has been reported that training Google's PaLM cost just $20 million (Lennart Heim, "Estimating PaLM's Training Cost," *xyz* blog, April 5, 2022).

Table 5.1. R&D Spending ($ billions)

FY	Alphabet	Amazon	Apple	Meta	Microsoft	Average
2020	27.57	42.74	18.75	18.45	19.27	21.28
2021	31.56	56.05	21.91	24.66	20.72	26.02
2022	39.50	73.21	26.25	33.62	24.51	33.10
2023	42.73	84.40	29.92	34.37	27.21	40.97
Average	35.34	64.10	24.21	27.77	22.93	30.34

SOURCE: Authors' calculation from FactSet data (FactSet, undated).
NOTE: FY = fiscal year.

With any fixed cost, the low variable costs suggest that there likely are economies of scale in foundation model development. However, currently the fixed costs are not so high that they significantly limit competition in the market, as there are a number of companies that do, or likely could, participate in the market. Of course, simply because a number of companies *can* participate in the market does not mean that all of these companies *should* participate in the market. Depending on the homogeneity of the product, the investments these companies are making through their fixed cost may lead to duplicated efforts without significant value creation.

Sunk Cost in the Status Quo Market

Recall that the *risk* associated with not recovering sunk costs can affect market entry. If the risk of a project failing is high and costs are sunk, would-be market entrants may forgo entry. Training a foundation model is often risky due to the possibility of poor performance, training inefficiencies, and technical challenges.[109] Additionally, such constraints as limited budgets and tight timelines sometimes necessitate concluding training despite unresolved issues—contributing significantly to non-recoverable expenses. For this reason, when developing a foundation model, fixed costs (e.g., compute acquisition, R&D) are largely sunk—encompassing non-recoverable investments in compute, labor, and pre-training data. While the resale value of a completed foundation model is not entirely zero, the high degree of risk and investment in their development underlines the sunk-cost nature of foundation model development. Below, we discuss the potential risk of sunk costs across the various fixed costs required in model development.

Compute acquisition: *Highly sunk*. While compute has resale value, reselling compute may not allow for the complete recovery of the full cost, particularly considering depreciation and the speed at which compute becomes outdated. Cloud-based compute rental may offer some degree

[109] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, et al., "Opt: Open Pre-Trained Transformer Language Models," arXiv, arXiv:2205.01068, May 2, 2022.

of flexibility from scaling up and down based on needs, but, as discussed above, most large models relying on cloud-based compute enter into longer-term contracts with compute providers, which suggests that cloud-based compute costs are also highly sunk.

Training data acquisition: *Partially sunk.* Training data come from a variety of sources, including publicly available data. If a model is trained primarily on publicly available data, acquiring the data may involve limited sunk costs apart from processing and data curation costs. However, if training data must be purchased or rented, the acquisition costs can be irreversible and the data may not have significant use cases outside of model training, creating the risk of sunk cost.

R&D: *Highly sunk*. The specialized knowledge, algorithms, and training techniques developed for a specific model or class of models may not be transferable to other projects, making R&D costs highly sunk. Of course, some research findings may have broader applicability, but extracting value from research can be challenging and may require additional fixed investments.

Labor: *Partially sunk*. The accumulated expertise gained by model development teams could be applied to future projects, even if those projects are not related to the specific model. However, given the specialized nature of labor required to develop foundation models, specific skills may not have broad applications and may not be fully transferable to other areas.

Network Effects in the Status Quo Market

In the status quo market, open-source foundation models appear to yield significant network effects via community development.[110] Perhaps the clearest evidence of foundation model network effects is observed in the robust community-led model improvement associated with Meta's Llama 2 suite of models. Llama 2—an open-source transformer-based foundation suite of models ranging from 7 billion to 70 billion parameters—employs a permissive licensing approach: a licensing approach with very few limits on how the model and code are used, changed, or distributed.[111] Meta's stated objective in open-sourcing Llama 2 and selecting a permissive licensing approach is to allow model improvement via extra-Meta community development.[112] Evidence of improved user experience includes the official Llama-recipes repository containing example implementations and explanations of fine-tuned models; a robust Llama 2 Chinese-language community dedicated to improvement of the model's performance

[110] It is possible that in a market with open-source and closed-source foundation models, economies of scale and network effects will be negatively correlated. Given that much of their utility comes from the convenience of hosting the model on a personal computer, open-source models are likely to be small and less costly to train, thus exhibiting relatively low economies of scale. Yet, open-source models are also more likely to exhibit network effects via community development. We thank Edward Parker for this astute observation.

[111] Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos, "Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model," *Preprints*, 2023.

[112] Meta, 2023.

and application in Chinese; and myriad user communities providing access to code, example applications, and tools to use Llama 2 on various platforms and for various end uses. [113]

The use of Llama 2 has been particularly common in academic research. The article introducing Llama 2, submitted to arXiv in July 2023, has already been cited more than 3,700 times, as of April 2024.[114] Researchers have used Llama 2 to develop models to detect ophthalmic disease,[115] detect potentially sexual predatory chat behavior,[116] and write code.[117] Furthermore, because the model is open source, researchers have been able to begin to probe its security and safety characteristics; this scrutiny may lead to security improvement.[118]

Economies of Scope in the Status Quo Market

In the status quo market, there is significant evidence of economies of scope. OpenAI has used its foundation model, GPT, in at least three products: ChatGPT, Dall-E, and GitHub Copilot. Gemini, a foundation model developed by Google, is being integrated into existing Google products, such as Search, Gmail, Google Drive, Google Docs, Google Maps, and YouTube.[119] The effect of economies of scope is to spread fixed cost over more production, decreasing average cost and increasing the strength of incumbent market participants.

Is the Status Quo Market a Natural Monopoly?

The market for pre-trained foundation models exhibits several attributes, such as high barriers to entry and negligible marginal costs, limited product differentiation, and economies of scale, that align with those of a natural monopolistic industry. Additionally, competition in the market currently may be a "war of attrition," characterized by (Bertrand) competition among developers on model scale and performance while incurring financial losses due to the high fixed

---

[113] Roumeliotis, Tselikas, and Nasiopoulos, 2023.

[114] Touvron et al., 2023.

[115] Huan Zhao, Qian Ling, Yi Pan, Tianyang Zhong, Jun-Yu Hu, Junjie Yao, Fengqian Xiao, Zhenxiang Xiao, Yutong Zhang, San-Hua Xu, Shi-Nan Wu, Min Kang, Zihao Wu, Zhengliang Liu, Xi Jiang, Tianming Liu, and Yi Shao, "Ophtha-LLaMA2: A Large Language Model for Ophthalmology," arXiv, arXiv2312.04906, December 8, 2023.

[116] Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins, "Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts," arXiv, arXiv:2308.14683, August 28, 2023.

[117] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al., "Code Llama: Open Foundation Models for Code," arXiv, arXiv:2308.12950, August 24, 2023.

[118] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou, "Safety-Tuned LLaMAs: Lessons from Improving the Safety of Large Language Models That Follow Instructions," arXiv, arXiv:2309.07875, September 14, 2023; Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish, "LoRA Fine-Tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B," arXiv, arXiv:2310.20624, October 31, 2023.

[119] Sundar Pichai and Demis Hassabis, "Introducing Gemini: Our Largest and Most Capable AI Model," Google, December 6, 2023.

costs of model development. In a war of attrition, market participants expect that over time the high costs of development will force market exit by weaker firms, consolidating the market to a small number of providers or to a single provider. Once weaker players are forced out of the market, the remaining firms price their output to deter re-entry, resulting in a fully captured market. The war of attrition dynamic is common in the early stages of industrial development,[120] and it is particularly common in new technology sectors.[121]

In sum, application of the natural monopoly criterion to the status quo foundation language model market (as of January 2024) indicates that the current case for a natural monopoly is relatively strong (Table 5.2). This conclusion rests on the observations that the current generation of foundation models is reasonably homogeneous, economies of scale are high, costs are largely sunk, and network effects and economies of scope are present.

Importantly, this conclusion does not mean that the foundation model market should be regulated. That decision depends on the existence of social costs. We discuss the potential social cost associated with a natural monopoly in the foundation model market in the next chapter.

Table 5.2. Application of Natural Monopoly Criteria to the Status Quo Market

Natural Monopoly Criteria	Foundation Model Variable	Status Quo Market
Homogeneous good	Limited product variation (at pre-trained foundation model level)	Yes
Economies of scale	Cost of training relative to variable costs	High
Sunk cost	Resale value of initial compute, data, labor, and foundation model	High
Network effects	Community-led development	Moderate
Economies of scope	Firms have multiple products using common foundation model	High
Case for a natural monopoly		**Strong**

Future Scenarios

Recall that technological change has the potential to alter the case for natural monopoly and, more generally, the competitive climate of the market. Here we consider two technology

[120] Joskow, 2007.

[121] Jeremy Bulow and Paul Klemperer, "The Generalized War of Attrition," *American Economic Review*, Vol. 89, No. 1, March 1999.

variables critical to determining the future competitive structure of the market for foundation models: the scaling hypothesis and the cost of compute technology. For each of these variables, we consider two conditions. Table 5.3 presents the conditions of each variable in the four scenarios considered here. In scenarios 1 and 2, the scaling hypothesis is assumed not to hold—i.e., the relationship between performance and model size/compute expenditure breaks down. In scenarios 3 and 4, the scaling hypothesis holds. In scenarios 1 and 3, compute technology is assumed to realize no significant technological discontinuity: Costs proceed to decrease at roughly the current rate. In scenarios 2 and 4, we assume, relative to current trends, a negative effect on computing performance relative to cost—i.e., the cost of compute more or less stagnates. For each scenario, we discuss the potential changes in fixed costs (focusing on compute costs), sunk cost risks, product homogeneity, network effects, and economies of scope.

Table 5.3. Technology Variables in Four Hypothetical Scenarios

		Cost of Compute	
		Current Trends Persist	**Stagnant Cost Improvement**
Scaling Hypothesis	**Does not hold**	Scenario 1. Proliferation of differentiated foundation models	Scenario 2. Toward a foundation model oligopoly
	Holds	Scenario 3. Toward a foundation model monopoly	Scenario 4. Optimal conditions for a foundation model monopoly

Technology Variable 1: The Scaling Hypothesis

Recent language model performance has been shown to depend primarily on the size of the pre-training dataset, the amount of compute used in pre-training, and the number of model parameters (i.e., performance depends on scale).[122] For example, Liang et al. (2023) tested large language foundation models across several dimensions using a set of scenarios (e.g., information retrieval, question answering, summarization) and metrics (e.g., accuracy, bias, fairness),[123] finding that larger models tended to perform better than smaller models.[124] Scaling also improves

[122] Kaplan et al., 2020.

[123] Liang et al. (2023) notes that to benchmark general language models, an adaptation procedure is required. The authors adapt all models through few-shot prompting, which involves providing the model with a few examples and instructions relevant to the specific task (Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Xhang, Deepak Narayanan, Yuhai Wu, Ananya Kumar, et al., "Holistic Evaluation of Language Models," *Transactions of Machine Learning Research*, August 2023).

[124] Liang et al., 2023.

the performance of models in other modalities—as demonstrated across vision,[125] robotics,[126] and multimodal settings.[127]

The impact of scale on performance can be seen in the substantial improvement between GPT-2 and GPT-3. This enhancement followed an increase of over an order of magnitude in model size and training compute resources, without any other significant alterations to the model's architecture.[128] In this case, performance growth manifested not in mere improvements on existing capabilities, but a qualitative shift in performance capacity; GPT-3 could perform wholly new tasks, such as meta-learning and direction-following.[129]

The scaling hypothesis simply contends that the observed relationship between scale and performance will continue. Put another way, even given a relatively fixed model architecture, the model performance relationship associated with data, compute, and model size will not significantly diminish.

Expert disagree on how long the relationship between model performance and model size will last.[130] On one hand, models like Koala-13B, a model small enough to run on a personal computer that leveraged high-quality dialog data scraped from the internet, have performed well in a head-to-head evaluation against ChatGPT.[131] Schick and Schütze (2020) show comparable performance to GPT-3 (174 billion parameters) with a much smaller model (223 million parameters) when the smaller model was trained using an alternative approach called pattern-exploiting training.[132]

On the other hand, there is also recent evidence that the scaling hypothesis continues to hold across various machine-learning architectures. Neumann and Gros (2023) found the scaling hypothesis to hold in the domain of reinforcement learning, observing that model (AlphaZero) performance scales as a power law based on parameter count.[133] Constantin (2023) observed that scaling law exponents in recent publications tended to be larger than those in older papers,

---

[125] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer, "Scaling Vision Transformers," Google Research, June 20, 2022.

[126] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," *Google DeepMind*, August 1, 2023.

[127] Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen, "An Empirical Study of Scaling Instruction-Tuned Large Multimodal Models," arXiv, arXiv:2309.09958v1, September 18, 2023.

[128] Gwern, "The Scaling Hypothesis," webpage, January 2, 2022.

[129] Gwern, 2022.

[130] Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," *Wired*, April 17, 2023.

[131] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song, "Koala: A Dialogue Model for Academic Research," *BAIR*, April 3, 2023.

[132] Timo Schick and Hinrich Schütze, "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners," arXiv, arXiv:2009.07118, September 15, 2020.

[133] Oren Neumann and Claudius Gros, "Scaling Laws for a Multi-Agent Reinforcement Learning Model," arXiv, arXiv:2210.00849, September 29, 2022.

suggesting the continued operation of scaling relationships.[134] Others have recently documented the relationship in transfer learning,[135] LLMs,[136] supervised neural machine translation models,[137] and neural language models.[138]

If the scaling hypothesis holds, the case for foundation models becoming a natural monopoly strengthens, assuming that the inputs to scale—data and compute—remain costly and largely fixed.[139] As these costs increase—in pursuit of significant model performance growth—the potential for economies of scale increases, raising the likelihood of reaching cost subadditivity. In contrast, if the scaling hypothesis fails to hold, the potential for natural monopoly decreases.[140] This assertion rests on the decreased relative importance of training costs in a world in which performance does not necessarily increase with model size.

Technology Variable 2: Compute Performance

To train modern foundation models is computationally expensive. This is due to the extreme scale and complexity of solving the algorithmic problems associated with such large models. Already the computational requirements of large AI models have driven innovation on the part of firms and developers in chips, partitioning, weight streaming, and distributed model training techniques, such as model parallelism and pipeline parallelism. As an example of innovation in chips, Hobbhahn and Besiroglu (2022) found that the FLOP per second per dollar for the GPUs most often used in ML have doubled at a rate of 2.07 years.[141] Such improvement in hardware performance translates into significant reduction in training costs; Nvidia announced at an industry conference in 2023 that when compared with CPU-based servers, using its GPUs would reduce the cost of an LLM training run by a factor of 25.[142]

---

[134] Sarah Constantin, "'Scaling Laws' for AI And Some Implications," *Rough Diamonds*, April 12, 2023.

[135] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish, "Scaling Laws for Transfer," arXiv, arXiv:2102.01293, February 2, 2021.

[136] Hoffmann et al., 2022.

[137] Mitchell A. Gordon, Kevin Duh, and Jared Kaplan, "Data and Parameter Scaling Laws for Neural Machine Translation," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[138] Kaplan et al., 2020.

[139] In the scenarios to follow, we assume that persistence of the scaling hypothesis (a technical phenomenon) will coincide with firms continuing to scale compute (a firm behavior phenomenon). We believe that this assumption is realistic in the near term, given that the primary dimension on which firms appear to be competing is model performance.

[140] This perspective is informed by the assumption, highlighted in our analysis, that in a scenario in which the scaling hypothesis does not hold, we fix the computational benchmark at the level required by the latest Gemini Ultra model, as estimated by Epoch at 9e25 FLOP. This provides a conservative estimate for future modeling, representing the upper limits of current technology. It allows for a simplified analysis framework, though it may not fully capture the nuances of incremental technological progress, external influences, or potential underestimations of future innovations.

[141] Marius Hobbhahn and Tamay Besiroglu, *Trends in GPU Price-Performance*, Epoch, 2022.

[142] Usman Pirzada, "Nvidia: Reduce the Cost of C CPU-Training an LLM from $10 Million to Just $400,000 USD by Buying Our GPUs," *WCCFTech*, May 28, 2023.

The scenarios presented below consider two conditions for the cost of future compute. The first assumes continuity of current per-unit compute cost trends. Given the rapid decrease in compute cost during the deep learning (DL) era, this constitutes an optimistic scenario. The second condition considers stagnation in the per-unit cost of compute (i.e., cessation in the trend of regularly decreasing compute cost). This condition could come from various sources. Perhaps the simplest would be a decrease in production capacity of high-end GPUs, CPUs, TPUs, or memory chips. Supply is already limited; Nvidia has not been able to keep up with demand for its high-end chips, leading to shortages, especially among smaller firms and university research groups.[143] A compute supply shock could occur for myriad reasons, including conflict in the Asia-Pacific or a natural or human-caused disaster that affects the highly concentrated semiconductor manufacturing sector. A fall in production capacity could limit industry-wide compute performance simply by limiting access to state-of-the-art chips.[144] Alternatively, because recent compute performance has depended on a series of technological and process innovations, the pessimistic cost scenario could arise simply due to a failure to innovate at the current rate, rather than an exogenous event.

The cost of compute has already likely affected the composition of the AI market. Ahmed and Wahed (2020) show that growth in the cost of compute necessary to conduct AI research has led to an AI research system that is increasingly dominated by large firms and elite universities, while mid- and lower-tier universities have been crowded out.[145]

Scenario 1. Proliferation of Differentiated Foundation Models

In Scenario 1, we assume that the scaling hypothesis breaks down and that current trends in the cost of compute continue. If the marginal benefit of increasing model scale diminishes, developers may seek performance or market advantages through other means, or they may seek to differentiate their products on other dimensions. For example, developers may pivot to experimentation with novel architectures or algorithms. Alternatively, developers might seek to gain competitive advantage by implementing alternative training strategies or by incorporating structured data or expert knowledge.[146] The breakdown of the scaling hypothesis could result in a shift away from general-purpose models to domain-specific pre-trained models. Furthermore, because of the decoupling between compute acquisition and model performance in this scenario,

---

[143] Erin Griffith, "The Desperate Hunt for the A.I. Boom's Most Indispensable Prize," *New York Times*, August 16, 2023.

[144] Appenzeller, Bornstein, and Casado, analysts at Andreessen Horowitz, are skeptical about the prospect of near-term production catching up to demand, noting the expected continued high near term growth in GPU demand and stating, "There is no sign that the GPU shortage we have today will abate in the near future"(Appenzeller, Bornstein, and Casado, 2023).

[145] Nur Ahmed and Muntasir Wahed, "The De-Democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research," arXiv, arXiv:2010.15581, October 22, 2020.

[146] Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso, "The Computational Limits of Deep Learning," arXiv, arXiv:2007.05558v2, July 27, 2022.

the entry cost is relatively low, and new market entrants may contribute to further differentiation in the foundation model market.

In Scenario 1, economies of scale can be expected to be present, but of a significantly lower magnitude than in other scenarios. We estimate that the cost of compute required to train the largest ML model in 2027 in this scenario would be $9.2 million (see the appendix for details about how costs were estimated). While such training costs are not trivial, they are well within the reach for a large portion of the corporate market. To help put this value in context, the R&D expenditures of the large technology companies listed in Table 5.1 have grown by an average annual rate of roughly 18 percent since 2020. Assuming that this growth rate holds, the average R&D expenditures in 2027 will be approximately $79 billion, making the cost to train the largest ML model in this scenario roughly 0.1 percent of annual R&D expenditures.[147]

Assuming that smaller, more specialized models result from the lower compute requirements in the scenario, the size of the user base that interacts with any given model will likely shrink. As a result, there may be limited scope for network effects, and this could limit the benefits of scale and communal development seen in very general foundation models. However, within niche areas, network effects could still emerge around domain-specific models.

Sunk costs may also decline relative to the status quo, largely due to lower fixed costs of model training. However, even with smaller fixed training costs, sunk costs will still be present given the inherent risk of developing a foundation model, particularly if the model's domain turns out not to generate significant downstream commercial applications. While the risk of sunk costs will likely remain, the absolute dollar value of sunk costs associated with model training and development would likely fall as compute requirements decline. The reduced financial exposure to sunk costs may encourage new firm entry into the market.

The potential for economies of scope may be more limited in this scenario than in the status quo market. With the potential emergence of smaller, more specialized models, firms may be less likely to reuse a common foundation model across multiple settings. However, within niche areas, firms may still be able to realize economies of scope by reusing domain-specific models for multiple purposes. For instance, a model specialized for medical imaging may have additional applications in diagnosis and treatment domains.

Overall, the market for pre-trained models in this scenario is likely to be competitive while exhibiting characteristics of monopolistic competition. While material fixed costs will likely persist in model development, these costs are unlikely to be so large that they deter entry, and, to gain market share, developers are likely to attempt to differentiate their model from their competitors' models. This aligns with the broad theory of monopolistic competition, which notes

[147] This future value calculation assumes a 17.97 percent growth rate and annual compounding until 2027, and it began with a 2023 R&D expenditure value of $40.97 billion. That is, $40.97 billion $(1 + 0.1779)^4 \approx$ $78.9 billion.

that product differentiation can displace prices as a competitive focal point when differentiation can be achieved at a reasonable cost.[148]

Scenario 2. Toward a Foundation Model Oligopoly

In Scenario 2, we assume a breakdown of scaling laws and stagnation in the cost of compute. Product homogeneity in this scenario can be expected to decrease relative to the status quo market but not to the extent predicted in Scenario 1. That is, in this scenario, model developers can still be expected to seek differentiation in the ways described above; however, the increased cost of entry associated with compute cost growth will limit market participation relative to that of Scenario 1. For instance, it is possible that the increased cost of compute could result in developers being less willing to take on the risks of novel model architectures in a high-cost environment, thus limiting entry and innovation.

In Scenario 2, economies of scale can be expected to be higher than in Scenario 1 but still lower than if the scaling hypothesis were to continue to hold. We estimate that the fixed cost of compute required to train the largest ML model in 2027, assuming sluggish improvement in price-performance for compute and no model scaling, would be $23.1 million. Assuming that trends in R&D expenditures of the companies listed in Table 5.1 hold, the cost to train the largest ML model in this scenario is approximately 0.03 percent of annual R&D expenditures.

Network effects would likely be more muted in this scenario compared with the status quo, though likely not as muted as they would be in Scenario 1. The failure of scaling laws would imply that smaller, more specialized models may emerge in this scenario, but the increased cost of compute could limit the number of new specialized models developed. However, like in Scenario 1, the specialized models that are developed in Scenario 2 could result in network effects in domain-specific areas.

The fixed and sunk costs required for model development would be higher in Scenario 2 than in Scenario 1 due to increasing compute costs. The increased magnitude and risk of sunk costs could deter market entry. Whether the fixed and sunk costs associated with model development in Scenario 2 are smaller than those in the status quo market depends on the relative reduction in compute requirements and increase in compute costs. If developing cutting-edge models in Scenario 2 requires significantly lower compute, the overall fixed costs of compute acquisition may fall below those of the status quo market.

The potential for economies of scope in Scenario 2 is similar to that in Scenario 1. The emergence of smaller, domain-specific models may reduce the opportunities to reuse foundation models across domains. However, in Scenario 2, the higher compute costs may discourage firms

[148] Corwin D. Edwards, "Reviewed Works: *The Theory of Monopolistic Competition* by Edward Chamberlain; *The Economics of Imperfect Competition* by Joan Robinson," *American Economic Review*, Vol. 23, No. 4, December 1933.

from creating multiple models and result in firms reusing existing specialized models across domains, even if doing so is not ideal.

Overall, the market for pre-trained models in this scenario is likely to trend toward oligopoly, in which a small number of firms dominate market share. The persistence of some economies of scale alongside rising compute barriers and increased financial risks involved with R&D could enable market power concentration among incumbents. However, the failure of scaling laws suggests that product differentiation and specialization may be required for performance gains, which creates niche markets, curbs the potential for monopolization, and creates opportunities for more than a single provider to operate profitably.

Scenario 3. Toward a Foundation Model Monopoly

Scenario 3 represents a continuation of current trends: The scaling hypothesis continues to hold, and the cost of compute decreases at prevailing rates. In this scenario, we contend that the relatively high observed homogeneity in the current market for pre-trained foundation models will continue. In this condition, developers can be expected to continue to realize significant model performance improvement by increasing model size and pre-training compute, leaving model architectures (e.g., the ubiquity of transformer architectures for pre-trained NLP models) as a secondary concern. Furthermore, the high cost of training a new model in this scenario may limit market participation, limiting opportunities for product differentiation by new entrants. In this scenario, we assume other potential differentiating features (e.g., novel architectures, algorithms, training regimes, or domain focus) to be relatively uncommon at the level of the pre-trained model.

Economies of scale in this scenario are high. In this condition, model size growth will significantly increase the fixed cost of compute. We estimate that if the scaling hypothesis holds, the cost to train the largest ML model in 2027 will be $6.2 billion where improvements in price-performance of compute maintain observed trends. Assuming that the trends in R&D expenditure growth hold, the cost to train the largest ML model in this scenario is approximately 7.4 percent of annual R&D expenditures among companies listed in Table 5.1.

If the scaling hypothesis holds and compute costs continue falling at their current rate, fixed training costs for leading models are likely to remain high. Under these conditions, fewer organizations will have resources to develop new models. As a result, most R&D activity may center on incremental improvements to dominant models. The resulting concentration in the market could drive network effects for key models.

As training costs grow, the financial risk associated with the sunk costs of model development will likely also grow. Increased compute demand could result in the continuation of current trends in limited compute availability. As a result, significant model development expenses will likely accrue before the model is trained and evaluated for potential end uses. However, there may be less uncertainty surrounding which model architectures are likely to succeed in producing a useful model. Generally, the dollar value increase in training costs will

likely limit entry into the market because few firms may be able to absorb a potential of a model failure. With an assumed continued focus on large general-purpose models, the potential for economies of scope would likely be as large as it is in the status quo market.

Overall, in this scenario, competition in the market for pre-trained models is likely to trend toward monopoly. The persistence of scaling laws suggests that high levels of model homogeneity will continue, with most models relying on similar architectures. With model performance and AI progress tied to model size and compute intensity, the marginal value of new model architectures and methods will likely decline. This dynamic could inhibit horizontal differentiation. Additionally, the increased fixed costs of model training are likely to deter entry of resource-constrained firms, and the irreversibility of these investments likely creates risk that may discourage innovation and market participation. In this scenario, the developer that is able to spend the most to train the largest model will likely be able to gain market share over competitors.

Scenario 4. Optimal Conditions for a Foundation Model Monopoly

In Scenario 4, the scaling hypothesis holds and the per-unit cost of compute fails to fall. As described in Scenario 3, if the scaling hypothesis holds, the observed homogeneity in the current market for foundation models can be expected to persist, as developers realize model performance gains using existing model architectures by increasing model size and training compute. As in Scenario 3, market participation, and thus the associated innovation, in this condition can be expected to be limited given the cost of entry.

Of the four scenarios considered, economies of scale are likely the highest in Scenario 4. In the pessimistic training cost condition, we estimate that if the scaling hypothesis holds and improvements in price-performance of compute slows, the cost to train the largest ML model in 2027 will be $15.7 billion. Assuming that R&D expenditure trends hold, this amounts to approximately 19 percent of average R&D expenditures in 2027 among the large technology companies shown in Table 5.1.

This scenario would likely result in the strongest network effects of the four scenarios, driven largely by the likely lack of alternatives provided in the market. With scaling laws holding despite rising compute costs, barriers to entry will probably remain high. With relatively few model providers and a continued concentration on general (as opposed to domain-specific) models, the communal benefits may grow. However, even in this scenario, the network effects may remain relatively minor compared with supply-side economies of scale.[149]

Scenario 4 represents the costliest training environment, given the adverse shift in compute costs and continuation of scaling laws for model performance. As costs increase, the financial risks associated with market entry would likely limit entry and experimentation. Multi-billion-

[149] Vipra and Korinek, 2023.

dollar sunk investments per model attempt without longer time horizons to reclaim expenses could discourage innovation and prohibit new entrants from developing competitive models.

Like Scenario 3, the assumed persistence of large general-purpose models could provide opportunities for economies of scope through model reuse across domains and modalities. However, the high compute costs in Scenario 4 may result in fewer firms developing foundation models, and that vertical integration with compute providers may allow some firms to benefit from scope more than others. For example, a compute vendor-developer pair may be able to train a large model with multiple applications, while developers without a compute partner may struggle to develop a sufficiently capable model with multiple alternative applications.

Like Scenario 3, the market in Scenario 4 is likely to trend toward a monopoly. However, the increased compute costs in Scenario 4 further limit competition in the market. The persistence of scaling laws is likely to lead to greater model homogeneity and, with the heightened barriers to entry from increased compute costs, less room for entrants to differentiate themselves from incumbents. Again, the best-resourced firms are likely to be able to capture the market by training larger models on more compute, though, unlike Scenario 3, the increasing compute costs further limit the number of potential developers with such resources. As a result, this scenario may result in greater vertical integration between compute providers and model developers. Table 5.4 summarizes all four scenarios.

Table 5.4. Summary of Scenario Outcomes (relative to status quo)

Natural Monopoly Criteria Relative to Status Quo	Scenario 1 (scaling hypothesis fails, cost low)	Scenario 2 (scaling hypothesis fails, cost high)	Scenario 3 (scaling hypothesis holds, cost low)	Scenario 4 (scaling hypothesis holds, cost high)
Homogeneous good	Lower	Slightly lower	Same	Same
Economies of scale	Lower	Slightly lower	Higher	Much higher
Sunk cost	Lower	Slightly higher	Higher	Much higher
Network effects	Lower	Lower	Same	Higher
Economies of scope	Lower	Slightly lower	Same	Uncertain
Likely market structure	Monopolistic Competition	Oligopoly	(weak) Monopoly	(strong) Monopoly
Case for a natural monopoly relative to status quo	Weaker	Weaker	Stronger	Stronger

Chapter 6. Does It Matter If the Market for Foundation Models Is a Natural Monopoly?

Our analysis indicates that the current market for large state-of-the-art foundation models and the hypothetical 2027 markets (if scaling laws persist until then—i.e., Scenarios 3 and 4) exhibit characteristics of natural monopoly. The traditional economic case for regulation of a natural monopoly depends on meeting two conditions: (1) The candidate market exhibits economic inefficiency or social costs, and (2) there exist plausibly implemented policy instruments that could reduce these costs. In this chapter, we briefly discuss the first condition; we leave discussion of the potential policy options (the second condition) to future researchers.

While we find that the status quo market has the characteristics of a natural monopoly, we think that the case for natural monopoly regulation is weak. This is due to the low observed social cost associated with the current market structure. The absence of social costs that would be typical of monopoly (e.g., high prices or low-quality products) likely is due to the robust competition in the current market for foundation models. For example, large tech companies—X, Meta, Microsoft, and Google—have made sizable investments in in-house and external foundation model development. Other AI labs, such as OpenAI, Cohere, and Anthropic, have developed highly capable models with growing user bases. Recently, firms in France, Abu Dhabi, China, and India have announced large investments in LLM development.[150] Finally, myriad open-source models, such as Bloom and OPT, are increasingly powerful and diffused.

However, if evidence of significant social costs emerges, the question of regulation should be reconsidered. Here, we recommend several areas to be monitored to identify potential social costs stemming from inefficient competition in the market for foundation models. The first two areas—the potential for pricing above marginal cost and low product quality—are social costs associated with all natural monopolies. The last three areas—competitive dynamics in the market for compute, the environmental impact of large training runs, and systemic risk—are particular to the foundation model market.

Prices

An unregulated monopoly can yield a social cost by enabling a monopolist to price its product or service higher than would be allowed given more competition. In a competitive or contestable market, competition for market share will force prices down until they reach marginal cost or, in the case of monopolistic competition, a markup over marginal cost that depends on the level of product differentiation. In an unregulated monopoly, the monopolist has

[150] "Welcome to the Era of AI Nationalism," *The Economist*, January 1, 2024.

the ability to set prices above marginal cost, resulting in economic inefficacy, or deadweight loss, that can be borne by the consumer.

However, prices may not always be good indicators of emerging market concentration. For example, it is possible that as the costs of training a foundation model grow, the current period of fierce competition will result in a single model provider. This market dynamic is commonly known as a war of attrition. In this dynamic, the near-term profitability of a model matters less than a firm's ability to withstand the increasing costs of innovation and its potential for realizing monopoly rents from being the last developer standing. In war of attrition competition, prices may be kept below marginal costs to gain market share and undercut competitors. Given how well-capitalized many of the players in the market for pre-trained models appear to be, monitoring prices for access to pre-trained models and their downstream uses may not be a sufficient tracking metric to understand how competition in the market is evolving.

Nonetheless, when paired with additional metrics, the continual monitoring of foundation model prices may act as an indicator of consumer harm due to market concentration. In the case of pre-trained foundation models, *prices* may refer to the licensing fees paid by application developers to use a foundation model or to user data that end users may be required to provide to access the model. In the case in which a single firm is offering consumer-facing products built on top of its own foundation model, foundation model prices will be embedded in the price of the final output and will thus require a step of price decomposition. Such price monitoring may require more transparency on the part of model developers.

Product Quality

There are at least two ways in which a monopoly might lead to consumer harm by way of inferior product quality. When a monopolist can set prices, in a static setting, it may choose to produce less than what would satisfy demand in order to set prices that maximize its profit. In practice, this may lead to a rate of innovation in the focal product below that expected given a competitive market. Second, an unregulated monopolist may select socially suboptimal R&D investment, upgrading, or service coverage strategies, resulting in a product quality that is lower than would be expected given perfect competition. Just as the current competitive market appears to have kept prices reasonably low, competition appears to be assuring a high standard of product quality and innovation. This is evident in the frequent release of upgraded state-of-the-art foundation models that demonstrate performance improvements relative to their predecessors. We would, nevertheless, suggest monitoring product quality going forward to ensure that concentration does not depress the incentive to innovate.

Competitive Dynamics in the Market for Compute

Given the important role of compute in model development and overall model costs, monitoring trends in compute costs can provide insights into competition in the market for pre-trained models. As such, tracking the competitive environment in the market for compute may

offer clues into competition in the market for pre-trained models and the potential emergence of a natural monopoly. If the market for compute becomes increasingly concentrated, with a few dominant players, it could result in higher prices and further increases in barriers to entry in the foundation model market. This could exacerbate the advantages for incumbent firms, making it more difficult for competitors to enter and challenge the incumbents' market power.

Additionally, growing concentration in the market for compute and higher compute costs could incentivize mergers and partnerships between model developers and compute providers. A movement toward greater vertical integration between developers and compute providers could lead to greater concentration in the market for foundation models, as some developers would have a competitive advantage over non-integrated firms.

Environmental Impact

Currently, the social costs stemming from inefficient competition in the pre-trained foundation model market are relatively limited. For example, there is an opportunity cost involved in duplicated investments in pre-trained models, as these resources may be better used in other domains. However, these opportunity costs are present in all investment decisions and likely only become a significant social cost if alternative causes become so critical and underfunded that spending on pre-trained models is limiting human progress. Additionally, any government-led means of avoiding duplication of investment is likely to be impractical, given that the direct costs are borne by firms themselves and directing private investment is not typically considered within the policy mix used to redress market concentration.

However, some emerging research suggests that there are externalities associated with model training. Specifically, the carbon cost of large training runs, duplicated across model developers, could become large and socially costly.[151] To the extent that pre-trained models remain relatively homogenous, excessive competition in the market could result in emissions exceeding the socially optimal level. More research is needed to determine how these potential social costs from carbon emissions relate to the potential benefits of AI capabilities. Monitoring energy use in model training could provide insights into the potential for excess competition in the pre-trained model market to cause meaningful social costs. The traditional approach to dealing with such an externality, a tax on carbon emissions resulting from model training runs, could add to the fixed costs associated with model training and further concentrate market power among large developers. Research is needed to determine the optimal way of dealing with potential carbon emissions resulting from model training efficiently and in a way that does not create additional barriers to entry in the market.

[151] Patterson et al., 2021.

Systemic Risk

Lastly, there are potential social costs associated with market concentration in pre-trained foundation models from systemic risks. As the market becomes more concentrated, more end users rely on a single model. While this creates network effects that can have positive spillovers, it also creates systemic risk in downstream markets. If a single private company provides the pre-trained foundation model on which the majority of economic uses rely, a service disruption could have wide-ranging economic effects. Illustrative of the potential for systemic risk that might be associated with market concentration, the chairman of the U.S. Securities and Exchange Commission (SEC) warned that increasing reliance on a small number of foundation models could lead to future financial crises.[152]

Model Safety

Recent and emerging concerns about the perceived safety risks posed by advanced AI foundation models have highlighted an additional potential social cost that could be exacerbated by the natural monopoly characteristics of the market.[153] As models become more capable and are deployed in increasingly high-stakes domains, their robustness, transparency, predictability, and alignment with human values may become more important. Ensuring that foundation models possess these qualities likely requires significant investment in safety R&D. In turn, the degree to which AI developers invest in safety likely depends on the *elasticity of demand for safety*—the sensitivity of customer demand to perceived and actual model safety risks.

In a market with elastic demand for safety, even a monopolistic provider would face pressure to invest in model safety, as customers would abandon an unsafe model. On the other hand, inelastic demand for safety could lead to underinvestment in safety even in a highly competitive market, as customers would continue to use unsafe models and developers may invest less in safety measures to remain competitive on price and performance. Factors that may influence the elasticity for safety include the stakes of the domain of model application (e.g., military uses), the visibility of safety auditing, overall awareness and attitudes about the risks of AI, and the availability of safe alternatives.

To illustrate the potential relationship between market concentration and investment in safety, consider two extreme scenarios. In a scenario in which demand is highly elastic and either perfect competition or pure monopoly exists, significant investments in model safety would be expected, as developers either invest to meet the public's high bar for safety or to protect their

---

[152] Lauren Sforza, "SEC Chairman Warns of Risk to Financial Systems from AI," *The Hill*, August 7, 2023.

[153] For instance, former and current employees at leading AI development organizations have made public their perceived social harms from AI. This includes comments by Jan Leike, the former head of OpenAI's superalignment team, an internal group focused on model safety (Jan Leike [@janleike], "Building smarter-than-human machines is an inherently dangerous endeavor. OpenAI is shouldering an enormous responsibility on behalf of all of humanity." post on the X platform, May 17, 2024). Similar concerns have been raised by current employees and published on the website righttowarn.ai (righttowarn.ai, undated).

market position. However, in a scenario of highly inelastic demand, even fierce competition among developers may not be sufficient to drive safety investments if customers are willing to accept models that are deemed unsafe, while a monopolist would face little pressure to prioritize safety. Monitoring the evolution of safety practices and metrics, as well as public response to these metrics, can act as a leading indicator of the elasticity of demand for safety.

Concluding Thoughts

We find that the status quo market for foundation models has characteristics of a natural monopoly. We also find that in future scenarios in which the scaling hypothesis holds, the case for a natural monopoly is likely to strengthen. However, we believe that the rationale for natural monopoly regulation is currently weak. This is due to the low observed social cost associated with the current market structure. We recommend that several factors be monitored as potential sources of social costs stemming from inefficient competition in the market for foundation models. Specifically, we recommend monitoring price relative to marginal cost, product quality, competition within the market for compute, environmental impact, systemic risk, and model safety. If these factors begin to show evidence of social cost imposition, the question of regulation should be revisited.

Appendix. Method for Estimating Future Model Training Costs

This appendix outlines the methodology employed in forecasting future training costs of foundation models. Our analysis hinges on two pivotal elements: (1) calculating the requisite FLOP for the training of future models and (2) forecasting the progression of the cost-effectiveness of computing hardware, quantified as FLOP per dollar (FLOP/$). Estimating these elements allows us to calculate the training costs of foundation models under a variety of hypothetical future scenarios.

Projecting Training FLOP of Future Foundation Models

This section outlines the methodology that projects the required FLOP for training future foundation models, updating the analysis originally performed by Tamay Besiroglu, Lennart Heim, and Jaime Sevilla in their 2022 report *Projecting Compute Trends in Machine Learning*.[154] Our approach builds on their work, extrapolating historical trends in foundation model training compute requirements under various assumptions—and utilizes Epoch's database of notable ML systems.[155]

Rates of Model Scaling

Our projections are grounded in two distinct scenarios based on the scaling hypothesis's applicability:

> **Scaling hypothesis continues:** This scenario assumes ongoing rapid growth in the compute required for developing foundation models, a trend that has been evident since the DL era began.[156] Historically, the usage of compute for training has doubled approximately every six months, a notable acceleration from the pre-DL era, in which compute doubled roughly every 20 months, aligning with Moore's Law.[157] This scenario envisions an eventual moderation of growth rates to align with Moore's Law, influenced

[154] Tamay Besiroglu, Lennart Heim, and Jaime Sevilla, *Projecting Compute Trends in Machine Learning*, Epoch, March 7, 2022.

[155] Epoch, "Notable AI Models," webpage, undated.

[156] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos, "Compute Trends Across Three Eras of Machine Learning," arXiv, arXiv:2202.05924v2, March 9, 2022.

[157] Our scenarios make separate assumptions about the future of compute demand growth and changes in compute costs over time. In reality, these two factors are interconnected. The trajectory of compute cost is influenced by shifts in compute demand and the resulting incentives to innovate. Our analysis does not account for this linkage and, as a result, may not fully capture the economic mechanisms that determine cost trajectories.

by economic and technological constraints as suggested by Lohn and Musser (2022).[158] To model this transition, we employ three scenarios from Besiroglu, Heim, and Sevilla, 2022—bearish, middle of the road, and bullish—reflecting the potential duration that DL-era growth rates persist before reverting to Moore's Law. These durations are informed by the time required for the cost of computation to decrease by an order of magnitude.

Scaling hypothesis does not continue: In the case in which the scaling hypothesis ceases to apply, we fix the compute level at that required for the last training run of Gemini Ultra, as estimated by Epoch at 9e25 FLOP as of January 30, 2024.[159] This approach serves as a conservative benchmark, representing the upper limits of current technology. It simplifies modeling by providing a stable metric, albeit with the limitations of potentially neglecting incremental advances, not accounting for external factors, and possibly underestimating future innovation.

Model

A Monte Carlo simulation is utilized to address the inherent uncertainty in long-term forecasting. Monte Carlo simulation generates a broad spectrum of potential future outcomes. By sampling from 40 different growth rates per model and conducting 10,000 model runs, mirroring the approach used in Besiroglu, Heim, and Sevilla (2022), we achieve a statistically reliable spread of outcomes. The simulation is executed with functional form: For each model **j** and year **i**, the logarithm of compute (FLOP) required for training is given by

$$\log C_{j,i} = \log C_{j,i-1} + \left(\frac{end - 2022}{timeline\_length - 1}\right) \cdot growth_i$$

where $\log C_{j,i-1}$ represents the logarithm of compute (FLOP) required for training in the previous year, $growth_i$ is a weighted growth rate based on the sampled reversion date from the distribution of reversion dates to Moore's Law, *coef_pre_dl* represents the growth rate consistent with Moore's Law prior to the advent of DL, and *coef_post_dl* denotes the growth rate observed after the introduction of DL.

$$growth_j = (coef\_post\_dl^{\,weights[i]}) \cdot (coef\_pre\_dl^{\,1-weights[i]})$$

The weight for each year, $weights[i]$, is determined by a logistic-like function that smoothly transitions based on the reversion date relative to 2022. The term scaling $growth_i$ adjusts the growth rate to match the number of years covered in each step of the simulation.

---

[158] Andrew Lohn and Micah Musser, "AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?" CSET, January 2022.

[159] Epoch, undated.

Results

We assess the median predicted annual FLOP for future training cost projection under the continuation of the scaling hypothesis. These estimates represent the expected FLOP required for training the largest foundation model each year.

While our projections focus on raw compute trends, it is important to acknowledge the potential influence of algorithmic progress on required training compute. Advances in algorithms could significantly increase the efficiency of compute usage—potentially requiring less FLOP for equivalent or superior outcomes compared with current models. However, quantifying the precise effect of these advancements is challenging and introduces additional uncertainties. Therefore, our current projections do not explicitly account for these algorithmic improvements. By applying the growth rates from the DL era and pre-DL era to our Monte Carlo simulations and prior over reversion dates to Moore's Law, we provide updated projections that reflect potential future compute requirements for ML systems.

Projecting Price-Performance (FLOP/$) of Compute Hardware

This section outlines the methodology used for projecting the future price-performance of compute hardware, specifically focusing on FLOP/$ for ML applications. Our approach utilizes a combination of regression analysis and bootstrapping techniques to estimate the annual change in FLOP/$ for different numeric precisions, particularly FP32 and FP16, due to greater data availability and relevance in the ML field. This method reproduces the work detailed in Epoch's *Trends in Machine Learning Hardware* report, which emphasizes the importance of precision-specific performance analysis in contemporary ML hardware evaluation.[160]

Assumptions

Our estimation is predicated on several key assumptions:

Continued trend of numeric precision: The analysis primarily focuses on FP32 and FP16 precision in ML hardware, reflecting their relevance for contemporary ML applications.[161] This focus aligns with recent trends that favor lower precision, such as FP16, for efficiency gains in training ML models.

Data reliability: The data used for this analysis, sourced from the *Trends in Machine Learning Hardware* report, provide a comprehensive view of ML hardware performance

[160] Marius Hobbhahn, Lennart Heim, and Gökçe Aydos, *Trends in Machine Learning Hardware*, Epoch, November 9, 2023.

[161] Nvidia, *Train with Mixed Precision: User's Guide*, DA-08617-001_v001, February 2023.

from 2010 to 2023.[162] Price-performance is calculated using release prices or cloud service rates, adjusted for inflation and with assumed profit margins. Including FLOP/$ data for 40 ML accelerators, methods used for determining FLOP/$ are extensively described in Hobbhahn, Heim, and Aydos (2023).[163]

Model

The methodology for projecting the future price-performance of ML hardware involves several stages. Initially, the data is prepared by converting all dates into numerical values that represent years since the dataset's start date, a step essential for facilitating time-based regression analysis. Next, regression analysis is performed for each numeric precision. This involves performing a linear regression of the logarithm of FLOP per dollar (log10(FLOP/$)) against the numeric date representation, aiming to estimate the rate of change in FLOP/$ over time. To enhance the robustness of these estimates and account for potential variability, a bootstrapping method is employed. This method entails resampling the dataset with replacement and recalculating the regression coefficients over 1,000 iterations, from which the 5th and 95th percentile estimates for FLOP/$ growth rates are derived to provide a confidence interval for the projections. Finally, using the obtained regression coefficients and the confidence intervals from bootstrapping, projections are made for the FLOP/$ for 2023 for both FP32 and FP16 precisions to serve as the base case on which varying growth rates are extrapolated.

Estimating Training Costs of Future Foundation Models

Our methodology for estimating the training costs of future foundation models uses as inputs the two projections described above: the annual median required FLOP for training these models and the price-performance (FLOP/$) of compute hardware. It is important to note that while our model separately estimates the required FLOP for training and the FLOP/$ of computing hardware, these variables are potentially correlated—i.e., they are not strictly independent. This acknowledged limitation should be taken into account when evaluating the projections made by our model.

Formula

The cost of training the largest foundation model in a given year for a specified precision level is calculated for two scenarios—with the scaling hypothesis (SH) and without the scaling hypothesis (~SH).

[162] Hobbhahn, Heim, and Aydos, 2023.

[163] Hobbhahn, Heim, and Aydos, 2023.

Scaling Hypothesis (SH)

The formula for estimating training cost when the scaling hypothesis is assumed is as follows:

$$Cost_{\,Year,\,Precision,\,Rate\,|\,SH} = \frac{FLOP_{\,Year}}{FLOP\_\$_{\,Year,\,Precision,\,Rate}}$$

where:

- $Cost_{\,Year,\,Precision,\,Rate\,|\,SH}$ is the estimated cost for the specified year, precision level, and assumed growth rate of FLOP/$ under the scaling hypothesis.
- $FLOP_{\,Year}$ represents the median predicted FLOP required for training in the specified year.
- $FLOP\_\$_{\,Year,\,Precision,\,Rate}$ is the projected FLOP/$ for the specified year, precision level, and growth rate.

Without Scaling Hypothesis (~SH)

In the scenario in which the scaling hypothesis does not hold, the formula is modified as follows:

$$Cost_{\,Year,\,Precision,\,Rate\,|\,\sim SH} = \frac{9 \times 10^{25}}{FLOP\_\$_{\,Year,\,Precision,\,Rate}}$$

Here, 9×10^{25} is a constant representing the fixed FLOP estimate, based on the assumption that required training FLOP remains stable at the level of Epoch's estimate of the most compute-intensive training run, Gemini Ultra, as of January 30, 2024.[164]

Results

Table A.1 details the resulting projections, providing a breakdown of estimated training costs for future foundation models between 2024 and 2027.

[164] Epoch, undated.

Table A.1. Training Costs of Future Foundation Models

	Projected Growth Rate in FLOP/$			
	Median (0.14 Order of Magnitude/Year) *Compute trends persist*		5th Percentile (0.04 Order of Magnitude/Year) *Stagnant cost improvement*	
Year	**SH**	**~SH**	**SH**	**~SH**
2024	$337,222,427	$24,237,634	$424,537,883	$30,513,374
2025	$911,220,979	$17,558,614	$1,444,187,927	$27,828,528
2026	$1,916,747,917	$12,720,091	$3,824,414,886	$25,379,919
2027	$6,233,530,230	$9,214,892	$15,657,920,006	$23,146,761

NOTE: Estimates assume use of FP16 in training and provide estimates under the scaling hypothesis (SH) and the scaling hypothesis not holding (~SH). Numbers cited in this report are highlighted. Growth rates are provided in orders of magnitude.

Abbreviations

AI	artificial intelligence
AltUp	Alternating Updates
API	application programming interface
AWS	Amazon Web Services
CPU	central processing unit
CRFM	Center for Research on Foundation Models
DL	deep learning
FLOP	floating-point operations
FLOP/$	floating-point operations per dollar
FY	fiscal year
GPT	generative pre-trained transformer
GPU	graphics processing unit
LLM	large language model
MANG	Microsoft, Amazon, Nvidia, and Google
ML	machine learning
NLP	natural language processing
O*NET	Occupational Information Network
PaLM	Pathways Language Model
R&D	research and development
RAM	random access memory
SEC	U.S. Securities and Exchange Commission
SH	scaling hypothesis
TPU	tensor processing unit
TSMC	Taiwan Semiconductor Manufacturing Company Limited
VoIP	voice over internet protocol

References

Abdusalomov, Akmalbek Bobomirzaevich, Mukhriddin Mukhiddinov, and Taeg Keun Whangbo, "Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging," *Cancers*, Vol. 15, No. 16, August 2023.

Agrawal, Apoorv, "New VC in Town: 'MANG,'" *Apoorv's Notes* blog, January 18, 2024.

Ahmed, Nur, and Muntasir Wahed, "The De-Democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research," arXiv, arXiv:2010.15581, October 22, 2020.

Ai, Chunrong, and David E. M. Sappington, "The Impact of State Incentive Regulation on the US Telecommunications Industry," *Journal of Regulatory Economics*, Vol. 22, No. 2, 2002.

Anil, Rohan, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, et al., "Gemini: A Family of Highly Capable Multimodal Models," *Google DeepMind*, December 19, 2023.

Appenzeller, Guido, Matt Bornstein, and Martin Casado, "Navigating the High Cost of AI Compute," *Andreessen Horowitz*, April 27, 2023.

Arrow, Kenneth, "Economic Welfare and the Allocation of Resources," in National Bureau of Economic Research, *The Rate and Direction of Inventive Activity: Economic and Social Factors*, Princeton University Press, 1962.

Baumol, William J., Janusz A. Ordover, and Robert D. Willig, "Parity Pricing and Its Critics: A Necessary Condition for Efficiency in the Provision of Bottleneck Services to Competitors," *Yale Journal on Regulation*, Vol. 14, No. 145, 1997.

Besiroglu, Tamay, Lennart Heim, and Jaime Sevilla, *Projecting Compute Trends in Machine Learning*, Epoch, March 7, 2022.

Bianchi, Federico, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou, "Safety-Tuned LLaMAs: Lessons from Improving the Safety of Large Language Models That Follow Instructions," arXiv, arXiv:2309.07875, September 14, 2023.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al., *On the Opportunities and Risks of Foundation Models*, Center for Research on Foundation Models, August 2021.

Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Masleg, Betty Xiong, Daniel Zhang, and Percy Liang, "The Foundation Model Transparency Index," *GitHub*, October 2023.

Bresnahan, Timothy F., and M. Trajtenberg, "General Purpose Technologies 'Engines of Growth'?" *Journal of Econometrics*, Vol. 65, No. 1, January 1995.

Brohan, Anthony, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," *Google DeepMind*, August 1, 2023.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan,, Johannes Gehrke, Eric Horvitz, Eve Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang, "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," arXiv, arXiv:2303.12712, March 22, 2023.

Bulow, Jeremy, and Paul Klemperer, "The Generalized War of Attrition," *American Economic Review*, Vol. 89, No. 1, March 1999.

Carugati, Christophe, *Competition in Generative AI Foundation Models*, SSRN, Working Paper 14/2023, September 18, 2023.

Chiusano, Fabio, "Two Minutes NLP—Scaling Laws for Neural Language Models," *Medium*, March 18, 2022.

Clark, Don, "Nvidia Revenue Doubles on Demand for A.I. Chips, and Could Go Higher," *New York Times*, August 23, 2023.

Constantin, Sarah, "'Scaling Laws' for AI And Some Implications," *Rough Diamonds*, April 12, 2023.

Dai, Andrew M., David R. So, Dmitry Lepikhin, Jonathan H. Clark, Maxim Krikun, Melvin Johnson, Nan Du, Rohan Anil, Siamak Shakeri, Xavier Garcia, et al., "PaLM 2 Technical Report," Google, May 17, 2023.

Debaere, Peter, and Andrew Kapral, "The Potential of the Private Sector in Combating Water Scarcity: The Economics," *Water Security*, Vol. 13, No. 1, August 2021.

DeepSpeed Team, Rangan Majumder, and Andrey Proskurin, "DeepSpeed: Accelerating Large-Scale Model Inference and Training via System Optimizations and Compression," *Microsoft Research Blog*, May 24, 2021.

Desislavov, Radosvet, Fernando Martinez-Plumed, and Jose Hernandez-Orallo, "Compute and Energy Consumption Trends in Deep Learning Inference," arXiv, arXiv:2109.05472v2, March 29, 2023.

Eadline, Doug, "Nvidia H100: Are 550,000 GPUs Enough for This Year?" *HPCwire*, August 17, 2023.

Edwards, Benj, "ChatGPT Sets Record for Fastest-Growing User Base in History, Report Says," *arsTechnica*, February 1, 2023.

Edwards, Corwin D., "Reviewed Works: *The Theory of Monopolistic Competition* by Edward Chamberlain; *The Economics of Imperfect Competition* by Joan Robinson," *American Economic Review*, Vol. 23, No. 4, December 1933.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," arXiv, arXiv2303.10130, August 21, 2023.

Emad [@EMostaque], "We actually used 256 A100s for this per the model card, 150k hours in total so at market price $600k," post on the X platform, August 28, 2022.

Epoch, "Notable AI Models," webpage, undated. As of March 1, 2024: https://epochai.org/data/epochdb/table

Epoch, "Machine Learning Trends," webpage, last updated February 7, 2023. As of March 28, 2024: https://epochai.org/trends

Erdil, Ege, and Tamay Besiroglu, "Algorithmic Progress in Computer Vision," arXiv, arXiv:2212.05153, December 10, 2022.

FactSet, homepage, undated. As of August 7, 2024: https://www.factset.com

Fisher, Irving, *Elementary Principles of Economics*, The Macmillan Company, 1912.

Gemini Team and Google, "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context," *Google DeepMind*, 2024.

Geng, Xinyang, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song, "Koala: A Dialogue Model for Academic Research," *BAIR*, April 3, 2023.

Gillham, Jonathan, "What Are Transformer Models—How Do They Relate to AI Content Creation?" *Originality.ai* blog, undated.

Gillioz, Anthony, Jacky Casas, Elena Mugellini, and Omar Abou Khaled, "Overview of the Transformer-Based Models for NLP Tasks," *IEEE*, 2020.

Gordon, Mitchell A., Kevin Duh, and Jared Kaplan, "Data and Parameter Scaling Laws for Neural Machine Translation," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Gregory, Robert Wayne, Ola Henfridsson, Evgeny A. Kaganer, and Harris Kyriakou, "The Role of Artificial Intelligence and Data Network Effects for Creating User Value," *Academy of Management Review*, March 2021.

Griffith, Erin, "The Desperate Hunt for the A.I. Boom's Most Indispensable Prize," *New York Times*, August 16, 2023.

Gu, Albert, and Tri Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," arXiv, arXiv2312.00752, December 1, 2023.

Gwern, "The Scaling Hypothesis," webpage, January 2, 2022. As of February 29, 2024: https://gwern.net/scaling-hypothesis#scaling-hypothesis

Hall, Christine, "Chaos at OpenAI Adds Fuel to the AI Talent Poaching War," *TechCrunch*, November 20, 2023.

Heim, Lennart, "Estimating PaLM's Training Cost," *xyz* blog, April 5, 2022.

Hernandez, Danny, and Tom B. Brown, "Measuring the Algorithmic Efficiency of Neural Networks," arXiv, arXiv: 2005.04305, May 8, 2020.

Hernandez, Danny, Jared Kaplan, Tom Henighan, and Sam McCandlish, "Scaling Laws for Transfer," arXiv, arXiv:2102.01293, February 2, 2021.

Ho, Anson, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla, "Algorithmic Process in Language Models," arXiv, arXiv:2403.05812, March 9, 2024.

Hobbhahn, Marius, and Tamay Besiroglu, *Trends in GPU Price-Performance*, Epoch, 2022.

Hobbhahn, Marius, Lennart Heim, and Gökçe Aydos, *Trends in Machine Learning Hardware*, Epoch, November 9, 2023.

Hoffman, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al., "Training Compute-Optimal Large Language Models," *DeepMind*, March 29, 2022.

Hur, Johnson, "The History of VoIP," *BeBusinessed*, webpage, undated. As of February 29, 2024: https://bebusinessed.com/history/voip-history/

Joskow, Paul L., "Regulation of Natural Monopoly," in A. Mitchell Polinsky and Steven Shavell, eds., *Handbook of Law and Economics*, North-Holland, 2007.

Kan, Michael, "Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs," *PC Magazine*, January 18, 2024.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling Laws for Neural Language Models," arXiv, arXiv:2001.08361, January 23, 2020.

Kiesling, Lynne, "Economic Foundations: Natural Monopoly Theory II," *Knowledge Problem*, February 9, 2023.

Knight, Will, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," *Wired*, April 17, 2023.

Korinek, Anton, and Jai Vipra, "Concentrating Intelligence: Scaling Laws and Market Structure in Generative AI," working paper, February 28, 2024. As of April 3, 2024: https://www.dropbox.com/scl/fi/3vmu5q8js8bcgvujbc9bg/Economic-Policy-Draft-R2-Concentrating-Intelligence.pdf?rlkey=imw3sdkhu45em92j9kg4pl5vm&dl=0

Leike, Jan [@janleike], "Building smarter-than-human machines is an inherently dangerous endeavor. OpenAI is shouldering an enormous responsibility on behalf of all of humanity." post on the X platform, May 17, 2024. As of August 16, 2024: https://x.com/janleike/status/1791498183543251017

Lermen, Simon, Charlie Rogers-Smith, and Jeffrey Ladish, "LoRA Fine-Tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B," arXiv, arXiv:2310.20624, October 31, 2023.

Leswing, Kif, "Meet the $10,000 Nvidia Chip Powering the Race for A.I.," *CNBC*, February 23, 2023.

Levels.fyi, "OpenAI Software Engineer Salaries," webpage, undated. As of February 29, 2024: https://www.levels.fyi/companies/openai/salaries/software-engineer?country=254

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Xhang, Deepak Narayanan, Yuhai Wu, Ananya Kumar, et al., "Holistic Evaluation of Language Models," *Transactions of Machine Learning Research*, August 2023.

Lightcast, homepage, undated. As of August 7, 2024: https://lightcast.io

LMSYS Chatbot Arena Leaderboard, webpage, undated. As of February 29, 2024: https://chat.lmsys.org/?leaderboard

Lohn, Andrew, and Micah Musser, "AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?" *CSET*, January 2022.

Lu, Yadong, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen, "An Empirical Study of Scaling Instruction-Tuned Large Multimodal Models," arXiv, arXiv:2309.09958v1, September 18, 2023.

Marathe, Anoli, Deva Ramanan, Rahee Walambe, and Ketan Kotecha, "WEDGE: A Multi-Weather Autonomous Driving Dataset Built from Generative Vision-Language Models," *Computer Vision Foundation*, 2023.

Meta, "Let's Get Building!" Facebook post, July 18, 2023. As of February 29, 2024: https://www.facebook.com/Meta/posts/292330510122833/?paipv=0&eav=AfbQGViEIQliSCecvVWyO_nYeThMNxe4DGXf3d8pMVVYn8Qb6--KZ8EFVwH6xWQ8ztY&_rdr

Mohr, Tanga McDaniel, "Natural Monopoly," in Timothy C. Haab and John C. Whitehead, eds., *Environmental and Natural Resource Economics: An Encyclopedia*, Greenwood, 2014.

Narechania, Tejas N., "Machine Learning as Natural Monopoly," *107 Iowa Law Review*, Vol. 1543, 2021.

Neumann, Oren, and Claudius Gros, "Scaling Laws for a Multi-Agent Reinforcement Learning Model," arXiv, arXiv:2210.00849, September 29, 2022.

Nguyen, Thanh Thi, Campbell Wilson, and Janis Dalins, "Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts," arXiv, arXiv:2308.14683, August 28, 2023.

Nvidia, *Train with Mixed Precision: User's Guide*, DA-08617-001_v001, February 2023.

OpenAI, "Aligning Language Models to Follow Instructions," webpage, January 27, 2022. As of June 5, 2024: https://openai.com/index/instruction-following/

OpenAI, "GPT-4 Technical Report," arXiv, arXiv:2303.08774v4, December 19, 2023.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, et al., "Training Language Models to Follow Instructions with Human Feedback," arXiv, arXiv:2203.02155, March 4, 2022.

Patel, Dylan, Myron Xie, and Gerald Wong, "AI Capacity Constraints—CoWoS and HBM Supply Chain," *Semianalysis*, July 5, 2023.

Patterson, David, Joseph Conzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean, "Carbon Emissions and Large Neural Network Training," arXiv, arXiv:2104.10350, April 21, 2021.

Pichai, Sundar, and Demis Hassabis, "Introducing Gemini: Our Largest and Most Capable AI Model," Google, December 6, 2023.

Pilz, Konstantin, Lennart Heim, and Nicholas Brown, "Increased Compute Efficiency and the Diffusion of AI Capabilities," arXiv, arXiv:2311.15377v2, February 13, 2024.

Pirzada, Usman, "Nvidia: Reduce the Cost of C CPU-Training an LLM from $10 Million to Just $400,000 USD by Buying Our GPUs," *WCCFTech*, May 28, 2023.

Pope, Reiner, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean, "Efficiently Scaling Transformer Inference," arXiv, arXiv:2211.05102v1, November 9, 2022.

righttowarn.ai, homepage, undated. As of August 1, 2024: https://righttowarn.ai/

Rombach, Robin, Patrick Esser, and David Ha, "High Resolution Image Synthesis with Latent Diffusion Models," Hugging Face, June 2022.

Roumeliotis, Konstantinos I., Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos, "Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model," *Preprints*, 2023.

Roziere, Baptiste, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al., "Code Llama: Open Foundation Models for Code," arXiv, arXiv:2308.12950, August 24, 2023.

Sampat, Samir, "Where Do Generative AI Models Source Their Data & Information?" *Smith.ai*, September 20, 2023.

Sample, Ian, "Race to AI: The Origins of Artificial Intelligence, from Turing to ChatGPT," *Guardian*, October 28, 2023.

Sardana, Nikhil, and Jonathan Frankle, "Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws," arXiv, arXiv:2401.00448v1, December 31, 2023.

Schick, Timo, and Hinrich Schütze, "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners," arXiv, arXiv:2009.07118, September 15, 2020.

Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos, "Compute Trends Across Three Eras of Machine Learning," arXiv, arXiv:2202.05924v2, March 9, 2022.

Sforza, Lauren, "SEC Chairman Warns of Risk to Financial Systems from AI," *The Hill*, August 7, 2023.

Sharkey, William W., *The Theory of Natural Monopoly*, Cambridge University Press, 1982.

Shmitz Jr., James A., *New and Larger Costs of Monopoly and Tariffs*, Federal Reserve Bank of Minneapolis, Research Department Staff Report 468, September 11, 2012.

Sun, Yutao, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei, "Retentive Network: A Successor to Transformer for Large Language Models," arXiv, arXiv:2307.08621v4, August 9, 2023.

Sytsma, Tobias, and Éder M. Sousa, *Artificial Intelligence and the Labor Force: A Data-Driven Approach to Identifying Exposed Occupations*, RAND Corporation, RR-A2655-1, 2023. As of February 28, 2024: https://www.rand.org/pubs/research_reports/RRA2655-1.html

Thompson, Neil C., Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso, "The Computational Limits of Deep Learning," arXiv, arXiv:2007.05558v2, July 27, 2022.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Tasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv, arXiv:2307.09288v2, July 19, 2023.

U.S. Bureau of Labor Statistics, "Databases, Tables & Calculators by Subject," webpage, undated. As of February 29, 2024: https://data.bls.gov/timeseries/APU000072610?amp%253bdata_tool=XGtable&output_view=data&include_graphs=true

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," arXiv, arXiv:1706.03762, June 12, 2017.

Villalobos, Pablo, and David Atkinson, *Trading Off Compute in Training and Inference*, Epoch, July 28, 2023.

Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho, "Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning," arXiv, arXiv:2211.04325v1, October 26, 2022.

Vipra, Jai, and Anton Korinek, "Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT," *Brookings*, Center on Regulation and Markets Working Paper #9, September 2023.

Wang, Xin, and Nishanth Dikkala, "Alternating Updates for Efficient Transformers," *Google Research* blog, November 7, 2023.

"Welcome to the Era of AI Nationalism," *The Economist*, January 1, 2024.

Zaretsky, Adam M., "I Want My MTV . . . and My CNN and My ESPN and My TBS and . . . ," *Regional Economist*, July 1995.

Zewe, Adam, "In Machine Learning, Synthetic Data Can Offer Real Performance Improvements," *MIT News*, November 3, 2022.

Zhai, Xiaohua, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer, "Scaling Vision Transformers," Google Research, June 20, 2022.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, et al., “Opt: Open Pre-Trained Transformer Language Models,” arXiv, arXiv:2205.01068, May 2, 2022.

Zhao, Huan, Qian Ling, Yi Pan, Tianyang Zhong, Jun-Yu Hu, Junjie Yao, Fengqian Xiao, Zhenxiang Xiao, Yutong Zhang, San-Hua Xu, Shi-Nan Wu, Min Kang, Zihao Wu, Zhengliang Liu, Xi Jiang, Tianming Liu, and Yi Shao, “Ophtha-LLaMA2: A Large Language Model for Ophthalmology,” arXiv, arXiv2312.04906, December 8, 2023.

About the Authors

Jon Schmid is a political scientist at RAND. He specializes in the measurement and assessment of technological innovation and scientific progress, with a particular focus on emerging and military technologies. Schmid holds a Ph.D. in international affairs, science, and technology.

Tobias (Toby) Sytsma is an economist at RAND and a professor of policy analysis at the Pardee RAND Graduate School. His research focuses on international trade and supply chain frictions, technology, natural disasters and climate change, economic development, and workforce development. Sytsma holds a Ph.D. in economics.

Anton Shenk is a quantitative research assistant at RAND. His research examines the economics of emerging technologies. Shenk holds a B.S. in economics and mathematics.

Because of the wide variety of tasks they can be used to perform, *foundation models*—a class of artificial intelligence (AI) models trained on large and diverse datasets and capable of performing many tasks—have the potential to have a large effect in shaping the economic and social effects of AI. The authors of this report examined the economic and production attributes of pre-trained foundation models to answer the following questions: Does the market for foundation models have the characteristics of a natural monopoly, and, if so, is regulation of that market needed?

A *natural monopoly* refers to a market in which the total cost of serving the full range of demand is lower for a single firm than for multiple firms. Unlike a conventional monopoly, in a natural monopoly, competition and traditional antitrust policy cannot be assumed to alleviate the problems associated with concentrated market power. The authors established empirical criteria for classifying a market as a natural monopoly and applied them to the status quo foundation model market and to four hypothetical scenarios set in 2027 to understand possible future market dynamics.

Application of the natural monopoly criteria to the status quo AI foundation language model market (as of January 2024) indicates that the current case for a natural monopoly is relatively strong. This conclusion is based on the observations that the current generation of foundation models is reasonably homogeneous, economies of scale are high, costs are largely sunk, and network effects and economies of scope are present.

www.rand.org

RR-A3415-1