Old Moats for New Models: Openness, Control, and Competition in Generative AI[1]

Pierre Azoulay
MIT Sloan and NBER
pazoulay@mit.edu

Joshua Krieger
Harvard Business School
jkrieger@hbs.edu

Abhishek Nagaraj
UC Berkeley-Haas and NBER
nagaraj@berkeley.edu

Abstract

Drawing insights from the field of innovation economics, we discuss the likely competitive environment in which generative AI advances will take place. Central to our analysis are the concepts of *appropriability*—whether firms in the industry are able to control the knowledge generated by their innovations—and *complementary assets*—whether effective entry requires access to specialized infrastructure and capabilities to which incumbent firms can ration access. While the rapid advancements in AI foundation models promises transformative impacts across broad sectors of the economy, we argue that tight control over complementary assets will likely result in a concentrated market structure, similar to past technological upheavals. We suggest the likely paths through which incumbent firms may restrict entry, confining newcomers to subordinate roles and stifling broad sectoral innovation. We conclude with speculations regarding how an oligopolistic future with blunted innovation could be averted. Policy interventions aimed at fractionalizing or forcing shared access to complementary assets might help preserve competition and incentives for extending the generative AI frontier. Ironically, the best hopes for a vibrant "open source" AI ecosystem might rest on the presence of a "rogue" technology giant, who might choose openness and engagement with smaller firms as a strategic weapon wielded against other incumbents.

[1] We thank Josh Lerner, Janet Freilich, Shikhar Ghosh, Pam Mishkin, Roland Szabo, Nikhil Naik, Rishi Bommasani, Juan Mateos-Garcia, Yoon Kim, and Roger Levy for useful discussions. Yanqi Cheng provided able research assistance. We acknowledge the help of GPT-4 and Claude 3 Sonnet as unparalleled brainstorming partners, expert paraphrasers, and whisperers of imaginary references.

1. Introduction

The history of technology is a steady march punctuated by great inflection points. In such moments, the steady accumulation of knowledge about how humans can manipulate our world coalesces into novel technologies with the potential for great jumps in productivity. Like steam engines, airplanes, antibiotics, semiconductors, and the internet in previous eras, generative artificial intelligence (AI) has the potential to be the next major transformative technology.

Interest in generative AI skyrocketed following the release of Open AI's large language models (LLMs) for text (ChatGPT) and image generation (DALL-E) in the second half of 2022. ChatGPT's uncanny ability to mimic human responses prompted individuals and firms to rethink what tasks might be aided or replaced by generative AI. Prominent business examples include novel protein design, writing software code, and weather forecasting. Meanwhile, individual consumers have found applications ranging from the mundane (automating email responses and shopping lists), to artistic (writing poetry, bedtime stories and screenplays), and skill building (e.g., personal tutors, coding copilots). As of early 2024, the GPT-4 model could pass the bar exam and a variety of Advanced Placement exams; score above the 93rd percentile on the SATs and above the 99th percentile on the GRE verbal exam; as well as perform a whole host of coding and programming challenges.

In sum, there is little doubt that many industries—including law, information technology, education, and entertainment—are poised to experience large productivity shocks and disruptions to their labor markets. This new wave of "foundational" generative AI models possess capabilities that are a clear step-improvements over previous generations of machine learning technologies, but we do not yet understand whether or to what extent any firm(s) will dominate the supply of generative AI technologies. While prognosticators face considerable uncertainty about the societal impacts of AI, regulators, investors, and managers must engage in sober assessments of innovation and competition in this fledgling sector in order to shape it. Understanding the industrial organization of the industry that will power these applications strikes us as a legitimate and necessary topic for entrepreneurship and innovation policy.

Rather than attempting a forecast regarding the depth and breadth of generative AI's applications, this paper harnesses the accumulated wisdom of research in the field of innovation economics to venture (with some trepidation and many caveats) a prediction for the likely *competitive environment* in which generative AI advances will take place. How might the technological features of generative AI shape the extent of product market competition in the industry? How likely are deviations from perfect competition to result in blockaded entry and delayed innovation? And what will determine the viability of a vibrant "open source" ecosystem, as opposed to a more secretive, incumbent-driven oligopoly? To guide our analysis, we draw on (Teece 1986), who gives primacy to the concepts of *appropriability* and *complementary assets* to identify the firms best able to profit from innovation: those who are first to market, follower firms, or firms that have related capabilities and assets that innovating firms require access to. *Our central claim is that while formal intellectual property and secrecy are unlikely to durably prevent entry by innovative firms, incumbent firms' tight control over key complementary assets will likely usher in a highly concentrated*

market structure. Thus, incumbents will have the ability to limit startup firm entry and confine new entrants to the "application layer" of the industry, in a pattern reminiscent of the smartphone industry.

Studies of previous general purpose technologies such as the steam engine, factory electrification, machine tools, and computer programming provide useful lessons for the competitive paths likely to emerge in generative AI (Bresnahan and Trajtenberg 1995; Shapiro and Varian 1999). As new and exciting as the early applications of AI may have been, they have mainly affected sectors that were either natively digital, or had already experienced sustained digital transformation, such as internet search, digital advertising, customer relationship management, and protein structure prediction. For generative AI to reach its general purpose technological potential, however, it will have to overcome bottlenecks that prevent invention in a broad set of application sectors, from agriculture to construction and financial services. These bottlenecks may be less technical in nature, and more rooted in the incentive problems and adjustment costs that organizations face as they digitize operations (Bresnahan 2024). An additional obstacle, and one central to the themes explored in this article, is that concentrated control in the upstream generative AI layer of the technology stack might well dim the incentives to innovate in the downstream sectors. As a result, a flowering of AI innovation throughout the economy faces substantial headwinds.

Industry predictions about the evolution of generative AI competition are quite varied, but ultimately hinge on whether early leaders can establish a competitive "moat." The flurry of new model releases has been matched only by feverish claims and counter-claims of model superiority. Evaluating these claims is challenging, especially given the gold rush of venture capital money flowing into the industry.[2] For example, in May 2023, a leaked internal Google memo titled *"We have no moat, and neither does OpenAI"* argued that the large and well-resourced leaders who pioneered the field using proprietary approaches would lose out to the nimbler open-source developers. Core to the argument was the idea that methods for improving and "fine tuning" open models were quickly dropping in cost.[3]

On the other hand, early adoption patterns suggested that early movers and well-resourced incumbents were skilled at turning new technologies into compelling products. Consumer adoption reflected the dominance of OpenAI's ChatGPT service, which reported to have more than 100 million monthly active users by January 2023, and 100 million weekly active users by November 2023 along with more than 2 million developers using its API.[4]

One software startup CEO noted that having a 'secret sauce' is rarely how defensible software businesses are made,[5] a claim backed by ample historical experience. Today's dominant technology platforms, including Google's search engine, Meta's social graph, and Amazon's e-commerce platforms were all built on technologies that were well-understood to competitors and potential

[2] Even after excluding the $14 billion in funding that went to LLM leaders Open AI and Anthropic, Pitchbook data showed that the first three quarters of 2023 saw more than $7.4 billion in venture capital deals supporting generative AI startups (https://pitchbook.com/news/articles/generative-ai-startups-vc-deals-decline).
[3] https://www.semianalysis.com/p/google-we-have-no-moat-and-neither
[4] https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/
[5] https://www.airplane.dev/blog/openais-moat-is-stronger-than-you-think

entrants during their entire rise. Rather these firms locked in customers through execution (e.g., in the form of compelling user interfaces), continuous innovation, and the harnessing of network effects. Further back, in industries from typesetting (Tripsas 1997) and aerospace, to personal computing and disk drives, a small number of firms have often dominated in fast-growing industries despite astonishing performance improvements, cost reductions and widespread technology diffusion. We draw on the insights gained in studying such breakthrough technologies to describe the forces likely to shape industry structure in generative AI, and how innovation incentives might play out under different competition scenarios.

We argue that tight control over specialized complementary assets is the most likely source of "moats" that would enable pioneering firms to durably entrench their dominant positions. We describe how producing and improving foundational generative AI models requires several interrelated complementary assets, such as a compute environment (to both train the models and serve their output), as well as access to massive quantities of non-public sources of training data, many of whom are protected by copyright. Early success begets further advantage through the endogenous development of other assets such as AI safety and governance expertise, compelling benchmarks and metrics, and data network effects—whereby users' engagement with a model generates information that dynamically improves its performance. In this context, hopes for a competitive market structure rest on the belief that vigorous public policy efforts will succeed in fractionalizing these assets or enforce their broad sharing across the industry. Alas, determined policy efforts of this type are likely to face strong political opposition.

Ironically, the oligopolistic future can perhaps best be averted through the actions of a single "rogue" technology giant. This rogue firm might pursue openness and engagement with smaller firms as a self-serving approach to improve its offerings in adjacent markets, win the "war" for scarce AI talent, or more generally to enhance its reputation as a benevolent leader in AI technology, influencing standards and practices across the industry. Hopes for a competitive generative AI marketplace resting on such a thin reed would ordinarily be construed as wishful thinking. We believe it deserves mention since one large tech firm—Meta, Inc.—seems intent on following through with their open source commitments, although whether this approach is sustainable in the face of competitive pressures remains in doubt.

Our paper begins with a technical primer on generative AI, describing the key technologies behind current day generative AI models. Next, we lay out a framework for assessing the prospects of the emergence of a vibrant competitive landscape in foundation models for generative AI, concluding that these prospects are rather dim. We end by proposing policy "low-hanging fruits" that have the potential to help usher in more competition in this domain.

2. Technical Primer

2.1 What do we mean by Generative AI?

During the 2010s, machine learning technology primarily focused on prediction tasks. Deep learning and neural networks transformed the field, enabling models to achieve high accuracy on tasks like financial fraud detection, image classification, and speech recognition. While groundbreaking, the main achievement was creating models that excelled at predicting outcomes from input data.

The transformer architecture has been used to develop so-called "foundation" large language models. These are large-scale AI models that serve as the basis for a wide range of downstream applications. Foundation models generate new content by learning from the patterns and structures present in the training data. During the training process, the model ingests vast amounts of text, images, or other forms of data and learns to predict the next token (e.g., word, pixel) based on the input sequence. By iteratively refining its predictions and adjusting the weights of the neural network, the model learns to generate outputs that closely resemble the training data. Once trained, foundation models can be fine-tuned for specific tasks or used for open-ended generation.

What makes foundation models different than the previous generation of machine learning models is that they are "pre-trained" (i.e. they do not need to be trained using a specific dataset) and they are generic and context-independent (i.e. they are not specifically trained for any particular task, but they are often powerful enough to perform a vast array of tasks without much additional training). For example, a language model like GPT-4 is extremely flexible and can perform varied tasks like language translation, text summarization, or question answering, with very little additional instruction. Alternatively, it can be used for creative writing tasks, such as generating stories, articles, or even code. Similarly, image-based foundation models like DALL-E can generate novel images based on textual descriptions, while audio-based models like AudioLM can generate music and speech. Foundation models are now being developed in a variety of other fields like in robotics and biology. It is this versatility and adaptability of foundation models that have opened up new possibilities for generative AI across various domains.

Recently, companies like OpenAI have exposed the functionality of foundation models like GPT-4 to third party players to build applications in a relatively unrestricted way. This logic is reminiscent of the smartphone application stores where third party developers leverage functional features of smartphones to develop targeted applications. Given their importance to the ecosystem of applications, foundational models can be a chokepoint for innovation and competition. Foundation models are therefore at the heart of policy debates about generative AI, especially as it relates to competition and entry. When we refer to generative AI in this article, *we are largely focused on competition for foundation models* rather than in the application layer.

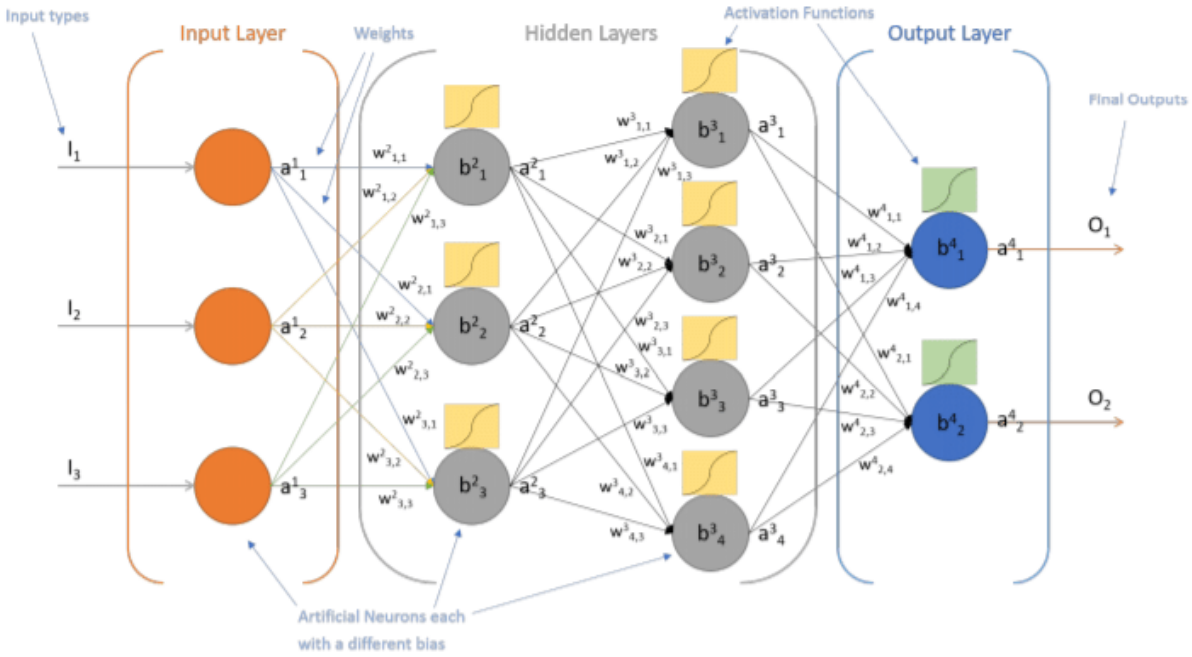
2.2 Large Language Models (LLMs)

Perhaps the most well-known of all foundation models is the Generative Pre-Trained Transformer (GPT) class of models from OpenAI, which belongs to a class of models called Large Language Models (LLMs) that are designed specifically to work with human language (and recently images and voice). Large Language Models (LLMs) are a subset of generative AI that analyze vast amounts of data

to understand context, recognize patterns, and generate new content. The term "language" in LLMs refers to various forms of symbolic representation, including human languages, computer languages, and any other type of encoded meaning. At their core, LLMs are algorithms that predict the next word or token when fed a string of words, progressively building up syntax that a human could plausibly create. LLMs model the ingested training data as a multi-dimensional space with numerous parameters, identifying artifacts (e.g., words, pixels) and their relationships to build new structures.

To understand how a model like GPT-4 works, it is useful to first take a deeper dive into the transformer architecture. The "M" in the LLM acronym stands for "Model," i.e., the empirical derivation of a mathematical formula that best captures the input text, manipulated as numerical data. At its core, each model runs on a neural network, a network of nodes and the connections between them that resemble the nerve cells (neurons) and their interconnections (synapses) in the brain.

Figure 1. Representation of a Neural Network in LLMs[6]



The neural network consists of nodes (the neurons) arranged in layers interconnected with nodes in the adjacent layer (See Fig 1). Each node performs some computation, with the result passed forward to the adjacent interconnected nodes. Models also incorporate activation functions which are nonlinear functions applied to the weighted sum of inputs coming into a neuron. It introduces non-linearity into the neural network, allowing it to learn and model complex relationships between inputs and outputs. Each interconnection is given a certain "weight," a multiplicative factor that captures the importance of the originating node's result. The weights are typically distributed in the form of model checkpoint files or serialized tensors. These files encode the architecture of the neural network, including the shape and dimensions of the weight tensors, as well as the actual numerical values of the weights themselves. The numerical values are derived iteratively by "training" the neural network with large amounts of known text samples to minimize a loss function that captures the difference between

[6] https://www.thedeload.com/p/understanding-ai-llm-basics-for-investors

the model's output and the expected outcome. These weights and some other adjustable inputs called "biases" are collectively called "parameters" and have become a proxy for representing the complexity of the neural network and the number of factors used to predict an output. Industry observers often compare generative AI models by the number of parameters. GPT-3, for example, has 175 billion parameters, while GPT-4 is estimated to have over a trillion.

Earlier generations of neural networks used for natural language processing suffered several drawbacks, including their inability to process whole sentences in one run and "long-term memory loss" – the inability to "remember" words occurring much earlier in the input text. These issues were overcome using the transformer architecture, created by Google in 2017 (Vaswani et al. 2017). Transformer architectures process input text – broken down into units called tokens, each representing a word, word fragment, strings of words, or a sentence – in parallel and can capture the context between words in sentences using a mechanism called "attention." When predicting the next word, the attention mechanism learns from the training data to "attend" or "focus" on specific words or phrases that are particularly relevant in the input context.

The training data used by prominent LLMs typically have trillions of sentences or images, taken from all publicly available sources such as the Web, books, social media content, audio, and video transcripts to recognize rules and patterns in the underlying data and make predictions for generating new content that uncannily resembles the training data. These days, some LLMs pre-train on specific categories, such as images (e.g., Open AI's DALL-E) or music (e.g., Google's AudioLM), to create new content specifically for these areas. Developers can also "fine-tune" models, a process which allows additional data to inform the output generated by a pre-trained model. While the most well-known LLMs distinguish themselves around content format (e.g., text, images, video, audio), other transformer-based generative AI models target specific knowledge domains and physical phenomena. One such example of a novel application is AlphaFold, an AI system pre-trained on data from the Protein Data Bank (PDB), a publicly funded effort started in the early 1970s to catalog the three dimensional structures of biological proteins. DeepMind trained their LLM on the PDB in order to predict a protein's 3D structure from its amino acid sequence. In 2018, Alphafold made a major leap in predictive accuracy, outperforming competitor protein structure predictions by a wide margin (Bertoline et al. 2023).

When considering alternate types of training data and extending the power of LLMs to other types of media, it is also worth noting that in addition to the transformer architecture, diffusion models (which are a type of convolutional neural network) have gained popularity. These models are based on reversing a diffusion process - they start with data (like an image or audio) and gradually add noise to it, diffusing and destroying the original data over many steps. They then learn to reverse this diffusion process to reconstruct the original data from the pure noise input. Popular diffusion models include OpenAI's DALL-E 2 and Stability AI's Stable Diffusion. While transformers model sequences via self-attention, diffusion models leverage the iterative diffusion-denoising process to generate new data from scratch in a very different way. The importance of training data and computing power remains as important for the diffusion models and other convolutional neural networks. Therefore, diffusion models and "chatbot"-style LLMs appear quite complementary in both their development and use, with many of the leading-edge generative AI developers offering both within the same user interfaces (e.g., "multi-modal" options within ChatGPT, Meta AI, or Google Gemeni).

The LLM landscape is dynamic with many new models and approaches and in practice, a mix of the transformer and diffusion architecture is used. Prominent examples of foundation models apart from GPT from OpenAI include Google's Language Model for Dialogue Applications (LAMDA), Large Language Model Meta AI (LLAMA) by Meta, and Claude from Anthropic. This wide class of foundational models serve as the basis for various applications, such as chatbots, language translation, and code generation. For instance, OpenAI's ChatGPT and Google's Gemini are chatbots built on top of their respective foundational models, GPT and LAMDA. LLMs are a crucial component of generative AI, enabling the creation of new content by analyzing vast amounts of data. By understanding how LLMs work, their training processes, and the applications they enable, we can better grasp the potential and implications of generative AI for innovation and competition.

2.3 Alternative foundation model architectures

LLMs have clearly brought a step-change increase to the immediate practical value of generative AI. The early success of *transformer-based* foundation models also strengthens the forecast that they represent a new "dominant design" for this technology that becomes the widely adopted paradigm that simultaneously constrains and enables progress in this domain (Utterback 1994).[7] If that prognosis is correct, then recent advances are only the beginning of a sequence during which incremental improvements in transformer-based models will continue for quite some time until their very design becomes a technological straitjacket, thus causing a marked slowing down as the current technological trajectory reaches a plateau.

A counter-argument might be that the field has not yet experienced the emergence of a dominant design, and that the enthusiasm for transformer-based models might recede if more promising alternatives emerge, or if different technology "architectures" prove more valuable for certain generative AI tasks. Aside from the convolutional neural networks (CNNs) used in many image generation models, liquid neural networks (LNNs) are another alternative, offering potential advantages such as greater interpretability, adaptability to new information and less computing needs– making them particularly attractive for robotics and autonomous vehicle applications (Hasani et al. 2021).[8]

Whether these alternative architectures surprise the industry or remain an academic research curiosity depends on both the uncertainty of scientific discovery and generative AI researchers' allocation of resources and attention across emerging technologies. While predicting the take-off of these alternatives (and their eventual commercial appeal) is beyond our remit and best left to computer scientists, the rest of the paper espouses the conventional view that foundation AI models currently situate themselves in the growth phase of an S-curve, a pattern followed by many new technologies, from tire materials to personal computing, artificial hearts, or X-Ray Lithography (Foster 1986; Henderson 1995).

---

[7] As noted in Section 2.2, the array of widely adopted early foundation models include transformer based LLM text-based models, as well as complementary image generation models that incorporate both transformer and convolutional neural network technologies to turn text prompts into novel images. Given these complementarities, we consider both as part of the same general class of foundation models that might represent a dominant design.

[8] see https://techcrunch.com/2023/08/17/what-is-a-liquid-neural-network-really/ for a discussion of advantages and disadvantages of LNNs, compared to other deep learning methods.

2.4 Computer scientists' openness fetish

As their name implies, foundation models form the substrate developers use to build a wide range of generative AI applications. Given this critical position in the generative AI ecosystem, foundation models can be a chokepoint for innovation and new entrants, offering incumbents a competitive edge that may be difficult to overcome.

Computer scientists have begun to tackle the question of competition in the area of foundation models by focusing on the concepts of "openness" and "transparency." The basic idea is that if the various elements of the "secret sauce" behind a model are made available through code, data, or detailed description, then it would be technically feasible to recreate it, and therefore compete against it. For example, one definition of openness claims that "open systems" are those that provide "transparency, reusability, and extensibility—they can be scrutinized, reused, and built on" (Widder et al. 2023). The idea is imported from the philosophy of the open source software movement, where openness means that the code behind a particular application is made available for public inspection and scrutiny.

How precisely openness and transparency should be measured, and what counts as genuinely "open" remain matters of active debate in the academic computer science community. For example, licensees of Open AI's GPT-4 have access to its APIs but have limited knowledge of its internals (such as the weights used in the model, a critical component for reuse). The company claims that it is open because it allows open access to the model's output, while critics argue that without transparency around how the model was built it remains relatively closed. They point to other large foundational models—Meta's Llama 2 and OPT-175B (although restricted for research only) and Google's BERT—that are identified as open source. But even in these cases, as Widder et al. note, the term "open source" does not yet have a precise meaning in AI, unlike traditional open-source software development, because many aspects of these purportedly open models remain black boxes. For example, some development efforts unaligned with major tech firms (such as Bloom from BigScience and GPT-J by EleutherAI) are considered open as the term is conventionally understood in the software industry, while Meta's Llama 2 has bespoke licensing terms that do not conform to more conventional open-source software.

Efforts are underway to standardize the definitions of openness and transparency. One notable effort in this spirit is the the Foundation Model Transparency Index (Bommasani et al. 2023) created by Stanford's Center For Research on Foundation Models, which aims to formally evaluate transparency and openness claims. These researchers score ten leading foundation models against a battery of one hundred "indicators" such as compute, data, model weights etc. The results of the 2023 index show that the ten major foundation models provide limited transparency, with a top score of 54% and a median of 37%, with the so-called "open" foundation models, Llama 2 and BLOOMZ, being the top scorers for transparency. Another such evaluation comes from Liesenfeld et al. (2023) which categorizes model openness across thirteen separate dimensions. Similarly, Schrepel and Pentland (2023) have proposed an openness taxonomy to guide policy recommendations that policy makers can use to maintain a competitive ecosystem.

These efforts are valuable in that they reveal the technical factors that might hinder the reverse-engineering of foundation models, which include access to training datasets, replicating their size and diversity of examples, the cost of training and serving them afterward, and their ability to grow their user base through positive feedback loops. The assumption is that if a particular technical component is not "transparent" or "open," then even well-resourced rivals would have trouble replicating these

frontier models. Openness advocates believe that unless leading foundation model developers lay bare the technical architecture, model weights, data sources etc., competition will necessarily be thwarted.

While these studies are valuable in assessing the technical landscape of foundation models, we disagree with the claim that competition and transparency go hand in hand. As the long history of technological innovation shows, there is usually only a weak relationship between openness in a technical sense and a regime where leading incumbents are unable to appropriate rents from their innovations. For example, in the pharmaceutical industry, firms protect their key discoveries using a mix of secrecy and patent protection, yet multiple incumbents can develop successful products using varied approaches within a broader "class" of therapies. Conversely, there are cases where generics have entered after a key patent has expired, but leading incumbents still enjoy considerable market power through their brand, distribution, and related assets, as seen with synthetic insulin and the continued dominance of Eli Lilly and Novo Nordisk despite the expiration of formulation patents over two decades ago.

We contend that philosophical discussions about the true meaning of openness constitute a distraction if one's goal is to elucidate the conditions that will enable ongoing competition in the emerging generative AI sector. Rather, as we will argue, the more important factors driving foundation model competition are incumbents' ability to endogenously raise the costs of reverse engineering over time, thus strengthening their grip on key complementary assets required to effectively compete.

3. Economic Moat-ivation: A framework for predicting likely competitive dynamics in foundation AI models

As practitioners of the dismal science, we are perhaps playing to type when we make light of the computer scientists' esthetic preferences for openness or transparency. But this stance stems from a dispassionate assessment of other technology markets. Historically, the relationship between openness of key technological ingredients and the extent of product market competition has been tenuous at best. To be sure, the circulation of technical talent across firms often means that access to key technical components eventually becomes democratized through feats of reverse engineering. This is especially the case when the national competitiveness stakes are high, as can currently be seen in the semiconductor industry. And yet, it is equally possible to point to numerous instances where technical know-how is widely dispersed across firms, and yet intense competition does not ensue.

Teece (1986) suggests that the ability of pioneer innovators to durably appropriate the returns from innovation is shaped by two fundamental forces: control over the knowledge at the core of the innovation—also called *appropriability*—and control over the *complementary assets* necessary to transform a firm's innovative know how into a value proposition customers might be willing to pay for. These assets might be tangible (like R&D tools, distribution networks, or manufacturing facilities), or intangible (like regulatory expertise, specialized skills, industry relationships, or brand reputation). Below, we analyze the emerging Generative AI industry through this lens. With a clear understanding of how the twin scissors of appropriability and complementary assets cut across the sector, we can derive implications for competition policy in this fascinating—but ultimately "normal"—domain.

3.1 Appropriability

Two broad approaches can be used by firms to protect the knowledge generated by their inventive and innovative activities: formal intellectual property rights (such as patents, copyrights, and

trademarks) and mechanisms to forestall reverse engineering (such as trade secrets, non-disclosure agreements, non-competes, and tacit knowledge).

Although there are plenty of AI-related patents (Giczy et al. 2022), patents do not appear to be especially valuable in the specific domain of foundation models. For example, the introduction of the transformer architecture was a significant milestone in the development of modern LLMs, but its broad contours were disclosed in a manuscript made available on arXiv, the most widely-used free distribution platform for pre-prints and working papers in computer science (Vaswani et al. 2017). That initial disclosure does not block all patents related to this architecture (e.g., pertaining to specific attention mechanisms), many of which might be quite valuable due to their wide applicability in models like BERT, GPT, and others. However, as is more broadly true in the software realm, much of AI/ML research may fall in the category of "abstract concept" and therefore fall short of the patentability standard.

Further, the fast pace of generative AI innovation may be mismatched with the pacing of patent examination in the modern patent system. This may lead inventors and may lead participants to turn to alternative mechanisms to control the knowledge associated with their ideas. It is worth noting that a variety of firms have at least applied for patents pertaining to technologies closely adjacent to foundation models.[9] Yet other firms hold patents on software features in their products which directly embed the output of foundation models.[10]

Whereas patents and other forms of IP have not proved to be central to firms' innovation strategies in these domains, much of the details necessary to implement foundation models in practice are shrouded in secrecy or deeply embedded in complex systems that may challenge the understanding of all but a handful of individuals working for the main firms active in these domains.

The development of foundation models is grounded in well-established scientific principles from fields like statistics, mathematics, and computer science, typically learned by data scientists over the course of their undergraduate or graduate studies. Understanding these principles is crucial for making informed decisions about model design and training. Training these models involve designing experiments to test hypotheses about model behavior, testing their performance on real-world tasks and datasets, and drawing conclusions in a methodical way. This scientific knowledge and associated practices are broadly available to all firms in industry, short-term skill shortage notwithstanding.

However, training generative AI models also require knowledge more akin to craft than the scientific method. Seasoned data scientists often develop an intuition for choosing the right model architectures, data preprocessing methods, and training parameters. Addressing challenges like overfitting, underfitting, or dealing with data irregularities often requires solutions that are seldom found in textbooks. Finally, the process of tuning hyperparameters and iterating over different model

[9] For instance, Google's "Processing of Deep Networks" (Szegedy and Vanhoucke 2017) covers methods for the compression of deep neural networks, i.e. reducing the computational complexity of deep neural networks while preserving, as much as possible, their performance. Microsoft's "Neural network categorization accuracy with categorical graph neural networks"(Du et al. 2024) pertains to the enhancement of neural network-based categorization using graph neural networks. Intel's "Optimizing Execution of Deep Learning Models on a Processor"(J. Ray et al. 2020) describes methods for optimizing the execution of deep learning models, including LLMs, on processors, specifically addressing challenges related to memory bandwidth and data movement.

[10] For instance, Adobe has patented technologies related to content generation and manipulation, such as the "Content-Aware Fill" feature in Photoshop, which intelligently fills in the space when a part of an image is deleted.

configurations involves much trial and error, rather than a systematic process prescriptively determined in advance.

Craft knowledge is more likely to remain tacit as it is deeply embedded and hard to unbundle from the broader architectural choices made by each firm in the design and training process. In this respect, generative AI models may be similar to other fields buffeted by technological breakthroughs. The spread of recombinant DNA research for instance, in spite of a seemingly broadly accessible scientific foundations, was for many years limited by pioneer researchers' ability to engage the next generation of innovators in long periods of craft-like apprenticeship (Zucker et al. 2002).

In addition to tacit knowledge, firms investing in the design and training of foundation models may be able to keep their models' weights proprietary. Without access to the weights, others cannot modify the pre-trained model without having access to the original training data or computational resources. Sponsors of generative AI models can forestall reverse engineering by maintaining tight control in other areas, including the specifics of the training data corpus—which encompasses the exact datasets used and any preprocessing steps that have been implemented, details about the infrastructure and scaling technology, the methods employed for fine-tuning (which are essential for adapting a pre-trained model to a particular task or dataset). The intricacies of the user interface layer, which facilitate smoother interaction with the model through mechanisms such as natural language prompts, are also another aspect that might be hard to replicate at scale. Lastly, foundation model developers might take inspiration from industries like stock photography and genetically modified seeds and develop their own technologies to detect IP leakage or unlicensed commercialization using their LLMs.[11]

Our interim conclusion is that pioneering firms in the industry benefit from a tight appropriability regime more owing to the many avenues to keep critical knowledge proprietary or tacit, rather than the assertion of formal intellectual property such as patents. This conclusion must remain guarded, however, since the field is still at an early stage of development. As AI talent becomes less scarce, the mobility of engineers and other technical staff will contribute to the circulation of knowledge that previously circulated as snippets within circumscribed communities of practice. "Closed" model developers may also develop new tactics to detect whether rival firms have illicitly learned and adopted their methods. In Section 3, the accidental leak of Meta's LLaMA large language model will provide an opportunity to revisit the question of whether a looser appropriability regime is likely to contribute to the democratization of innovation in this domain.

3.2 Complementary Assets

Complementary assets refer to the set of assets or capabilities required to effectively commercialize and extract value from that innovation (Teece 1986). These may include specialized manufacturing capabilities, access to distribution channels, service networks, or complementary technologies. Complementary assets may be generic and therefore not tightly held by incumbent firms. In contrast,

[11] Even before the current AI boom, Getty Images increasingly used technologies like reverse image searches and digital watermarks (https://www.stockphotosecrets.com/news/getty-images-is-leading-the- way-with-a-new-watermark.html) to detect and then enforce copyright protection (Luo and Mortimer 2017). Agricultural supplier Monsanto famously uses a range of methods to detect unauthorized use of their patented genetically modified seed. Monsanto's detection methods included collecting crop samples, genetic testing, anonymous "tip lines" for farmers to report on infringers, and investigative agents known as the "seed police" or "bean detectives" that visit hundreds of farmers a year. https://www.iatp.org/news/monsanto-the-seeds-of-dispute, https://investigatemidwest.org/2013/02/18/anonymous-tipsters-often-prompt-monsanto-farm-investigations/, https://www.centerforfoodsafety.org/files/cfsmonsantovsfarmerreport11305.pdf

other complementary assets may be narrowly tailored and dependent on a specific innovation, with few or no substitutes in the industry. In this case, complementary assets can be deployed by incumbents in such a way that they constitute potent barriers to entry and imitation. As we discuss below, leading firms in the generative AI industry have already begun developing complementary assets. At least some of these assets have the potential to become the exclusive fiefdoms of a narrow set of large, incumbent technology firms. The following six complementary assets are probably the most salient:

3.2.1 Compute environment

The compute environment refers to the hardware and software infrastructure used to train and fine-tune foundation models. A compute environment's principal components include (1) hardware for substantial computational power, often provided by GPUs, which are specialized hardware for handling the parallel computations that are common in machine learning; (2) a software stack (e.g., operating system, machine learning frameworks); (3) networking infrastructure to facilitate quick data transfer between different computers; (4) integration with cloud services to scale resources up or down based on the needs of the model training or inference; and (5) tools to monitor the performance of the hardware, the progress of model training, resource utilization, and to manage the deployment and operation of models.

A well-optimized compute environment is crucial for the efficient training and operation of foundation models, given their large scale and computational requirements. It involves a careful balance of hardware capabilities, software support, and infrastructure management to ensure that the models can be trained, fine-tuned, and served effectively. While an active computer science literature explores how model sponsors should trade-off model size, training dataset size, and compute budget at the frontier (Hoffmann et al. 2023), this research broadly points to "scaling laws" reminiscent of Moore's Law (Kaplan et al. 2020), and as a result the capital expenditures needed to set up, operate, and maintain the necessary infrastructure is increasing with model size and complexity. To fix ideas, Meta's CEO Mark Zuckerberg recently announced that in its effort to train the next generation of its Llama large language model (Llama 3), the company was building a massive compute infrastructure including 350,000 of NVIDIA's H100 GPUs by the end of 2024. Valued at the retail price of this crucial piece of equipment, the GPU investment alone would amount to roughly 10 billion dollars.[12]

3.2.2 Model serving and model inference capabilities and infrastructure

Once trained, foundational model sponsors must have the capability to deploy the model in a production environment where it can receive input data, process it, and provide outputs to end-users or other systems. The computational power needed for deploying large AI models in practical applications often surpasses the initial training requirements. While precise data on this subject is scarce, an informal analysis indicates that the operational costs of running the ChatGPT interface might exceed the training expenses of the GPT-4 model it is based on every week. Data scientists confusingly denote the task of applying a learned model to make predictions or decisions on new, unseen data as "inference," and high inference costs have prompted OpenAI's CEO, Sam Altman, to acknowledge in a congressional testimony that the company aims to design systems that don't prioritize user engagement, partly due to a scarcity of GPUs, which are crucial for AI training and operations (Oremus 2023).

[12] https://www.theverge.com/2024/1/18/24042354/mark-zuckerberg-meta-agi-reorg-interview

In addition to compute infrastructure, sponsors must make the trained model accessible for practical use, for example through APIs or other interfaces. Serving a model involves tracking its performance, ensuring its uptime, and updating it (or the underlying infrastructure) to handle additional load or to integrate improvements. When models also process user-generated data, sponsors must set up secure data transmission channels, ensuring data privacy, and complying with relevant regulations.

Lastly, the output of foundation models might be further enhanced by tight integration into the web browsing experience and wider ecosystem features, such as Google's chrome or Microsoft's edge web browsers. Not only might ecosystem owners elevate access to their preferred model by default (not unlike their current approaches to the selection of search engines), they might also enrich answers by leveraging location data or the past browsing history of individual users.

3.2.3 Safety and governance

Sponsors must engage in a range of approaches and practical measures to ensure that foundation models are developed and used responsibly. The specifics vary between organizations, but generally begin with a set of ethical guidelines that guide development and deployment, with the creation of boards or committees responsible for overseeing compliance with these principles. From a technical standpoint, sponsors may develop methods to make their systems more transparent and understandable to humans, and implementing measures to protect the data the models are trained on and ensure the models do not leak private information.

Sponsors also invest in research dedicated to understanding and mitigating potential risks associated with AI, such as studying the problem of alignment (ensuring AI systems reliably do what their operators intend) and robustness (ensuring AI systems behave safely under a wide range of conditions).[13] To some degree, these R&D investments are product- or brand-specific choices undertaken with the goal of differentiation and risk management. However, developing safety systems might prove to be necessary to placate regulators and earning public trust. When testing new LLMs and their derivatives, conducting audits of released generative AI products, and handling customer complaints, incumbents that invested and demonstrated such safety systems may have both efficiency and credibility advantages.

Investing in AI safety can be quite expensive, but the exact figures are often not publicly disclosed and can vary widely between organizations and depending on the scale of the AI systems being developed. The expenses include not only direct financial investments in research and development but also the opportunity costs of implementing rigorous safety measures, which can slow down the development process. Despite these costs, leading AI organizations tend to view these investments as essential, both for ethical reasons and for the long-term viability and acceptance of their AI products.

3.2.4 Benchmarks and metrics

Commercial sponsors invest in the development of benchmarks to evaluate the performance of foundation models, often in collaboration with academic researchers, developers, industry experts, and sometimes even policymakers.

[13] Industry observers caught a glimpse of the importance of these research investment with a series of events surrounding the departure of Dr. Timnit Gebru from Google in late 2020. This incident attracted significant attention and led to widespread discussion about academic freedom, the ethical implications of AI, and diversity and inclusion within the tech industry.

This might take the form of investing in the creation or curation of large, diverse, and representative datasets that can be used to rigorously test the performance of foundation models across various tasks (such as the case of ImageNet), or in the form of hosting contests where researchers worldwide work to achieve the best performance on a given benchmark, thereby pushing forward the state-of-the-art in model evaluation. For example, the GLUE and SuperGLUE benchmark are designed to evaluate the performance of models on a range of natural language understanding tasks such as sentence completion, question answering, and sentiment analysis. They were developed by researchers at NYU, the University of Washington, and DeepMind.

Another area in which benchmark development is becoming more salient is governance, where firms endeavor to publish and promote safety methodologies for AI systems. For instance, Open AI published Safety Gym, a suite of reinforcement learning environments and tools for developing AI systems that learn to perform tasks while respecting safety constraints (A. Ray et al. 2019).

3.2.5 Training data

The training data corpus for foundation models is typically extensive, diverse, and sourced from a wide range of public or semi-public sources. Foundation model sponsors must bear the costs of harvesting, ingesting, storing, and preprocessing these data for training purposes. In some cases, they must guard against contaminating the corpus with biased and sensitive content, and keep the data updated to prevent "data staleness."

Some of the data used for training might be in the public domain, such as protein structure contained in the Protein Data Bank (i.e., PDB, used to train protein structure prediction models such as AlphaFold or RosettaFold), ImageNet (a large visual dataset used for image classification), Common Crawl (a massive dataset of archived web content). But many other data sources should be characterized instead as "semi-public," i.e., data that is not explicitly private (like personal emails or messages) but exists in a gray area concerning usage and privacy expectations. Examples include social-media posts, user-generated content on forums and discussion boards, code repositories, product reviews and feedback, academic papers, and newspaper articles.

When using such semi-public data, model sponsors need to consider legal aspects (like copyright and data protection laws), ethical considerations (like the reasonable expectation of privacy and the intentions of the content creators), and the potential impact on their reputation and trust with users. Most sponsors do not typically disclose the exact datasets used in the training process, while also claiming they will respect opt-out requests and endeavor to use the data in ways that align with the expectations of the individuals and organizations that originally created the content.

The legal framework that will govern the use of copyrighted data in the training of foundation models is not settled (Gans 2024; Lemley and Casey 2020). While sponsors typically invoke the doctrine of fair use, this is contested by copyright holders, and a flurry of legal cases suggest that sponsors might mitigate legal risks by using data for which they have licenses or explicit permission to use for training purposes. Of course, a consequence might be that only sponsors with sufficient financial resources or legal expertise will be in the position of acquiring training datasets of sufficient breadth and depth, while taking care of purging personal information accidentally found in the training data and more generally deal with biased and sensitive content. How copyright policy in this area evolves, and whether and to what extent training on copyrighted material will be deemed fair use, will have important implications for competition in this area.

3.3.6 Data network effects

An additional consideration is whether a very particular type of network effect is likely to arise from sheer access to massive amounts of training data. For LLMs specifically, as more data is collected and used in training, the model generally becomes better at understanding and generating natural language. This is due to the model's increased exposure to various linguistic patterns, nuances, and contexts. In some implementations, user interactions with the model can be used to further refine its output. For example, when users correct or flag inappropriate responses, this feedback can be used to improve the model. As more users interact with the model, the quality of the model can improve, benefiting all users.

Reinforcement Learning from Human Feedback (RLHF) is a training strategy where the model is fine-tuned based on feedback that reflects human preferences or corrections. RLHF has been particularly influential in training models for complex tasks where traditional reward functions are insufficient to capture all aspects of desired behaviors, such as language models, robotics, and game playing. It typically involves an iterative process where the model is continuously updated based on ongoing feedback from human raters. In this iterative cycle, each new piece of data adds value to the model, and the improved model, in turn, provides better service to users, encouraging more engagement and feedback.

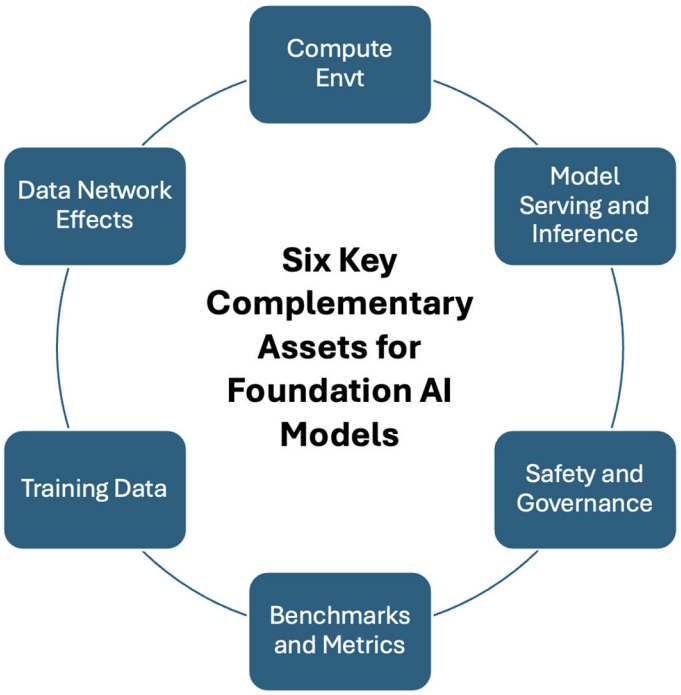
Since RLHF involves tradeoffs between generalization and diversity of output (Kirk et al. 2024), it is probably too early to assert it will give rise to data network effects that will favor firms able to create the infrastructure necessary to capture human feedback on a large scale. We choose to mention it as one of a broad class of complementary technologies that, if deployed in conjunction with massive training datasets, might endogenously lead initial performance advantages to cumulate and magnify over time.

3.3.6 The durability of complementary assets

While the list of complementary assets above (summarized in Figure 2) is not meant to be exhaustive, it suggests nonetheless that complementary assets have the potential to help turn foundation models from standalone technologies into integral components of products, services, or business processes. Their importance lies not only in enhancing the performance of these models but also in ensuring that they are usable, reliable, and aligned with broader organizational goals and societal norms.[14]

[14] At the same time, the gatekeeping effects of complementary assets need to be ascertained on a case by case basis. In structural biology, for example, the training datasets have so far mostly consisted of publicly available data, and the users of foundation models are technically sophisticated, lessening the importance of model serving capabilities. To date, the compute environment necessary to train these models has not put the frontier out of reach for academic labs, as attested by the arrival of the RoseTTAFold model (Baek et al. 2021), quick on the heels of Google DeepMind's AlphaFold (Jumper et al. 2021).

Figure 2. Key Complementary Assets Shaping Competition in Foundation Models



However, the inhibition of competition through tight control of complementary assets is subject to change as technological developments either amplify or depreciate their value (Tripsas 1997). For instance, much research effort has been recently directed to alleviating the burden of initial model training, thus lessening the need to invest in an extensive compute environment to make contributions at the frontier of the field. DeepMind's Chinchilla study highlighted that the size of large language models (measured in tokens) and the size of its training corpus are broad substitutes (Hoffmann et al. 2022). Simultaneously low-rank adaptation (LoRA) is a method designed to efficiently fine-tune large models without the need to retrain them fully, which can be computationally expensive and time-consuming (Hu et al. 2021). Combined with techniques in model distillation, these developments offer the tantalizing possibility of a level playing field between large tech companies, startups, and academic groups.

The research frontier in AI foundation models is perpetually shifting outward. On the one hand this could provide opportunity for laggards to leapfrog over the achievements of pioneer innovators. On the other hand, there is likely no model training "free lunch," with follow-on innovators requiring access to compute infrastructure–of their own to realize the promise of their innovative approaches.

3.4 A natural experiment in appropriability: The Llama leak and its consequences

Relative to Google and OpenAI, Meta's Facebook AI has distinguished itself by embracing an approach of relatively open sharing of its foundation models. In April 2022, the company released its OPT (Open Pre-trained Transformer) large language model, along with extensive documentation describing model architecture and design, training data and methodology, with detailed performance

and benchmark information. Even seemingly unabridged log books detailing the many steps needed to train the model were made available, which provided a rare glimpse into the craft-like techniques used by data scientists to generate the model weights.

However, the model weights for the 175 billion parameters OPT-175B were not released to the public. Meta's decision to withhold the model weights was likely due to concerns about potential misuse and the need for responsible AI development and deployment. One year later, Meta followed a similar approach with its more advanced Llama (Large Language Model Meta AI) model.[15] Merely a week after Meta began accepting inquiries for Llama access, the model found its way online. On March 3rd, 2023, a downloadable version of the system (including the model's weights) appeared as a torrent on 4chan, and it quickly disseminated through numerous AI forums.

The accidental leak ushered in a quick sequence of cascading events. Even though Llama's release was stripped to essentials, the availability of LLaMA's pre-trained weights enabled researchers who might not have the resources to train such large models from scratch to experiment with and improve upon a state-of-the-art LLM. Having a high-quality LLM available without constraints also allowed for comprehensive benchmarking and comparative studies.[16] Following the leak, numerous Llama-based developments emerged in quick succession, including Stanford's Alpaca, an instruction-following model; Vicuna from UC Berkeley, CMU, Stanford, and UC San Diego; BAIR's Koala, enhanced with internet dialogues; Nebuly's ChatLLama for custom conversational assistants; FreedomGPT, an Alpaca-based conversational agent; and UC Berkeley's Colossal-AI project's ColossalChat, a ChatGPT-style model using Llama. Within a month, these "spawns of Llama" included variants with instruction tuning (Zhang et al. 2023), quantization,[17] and RLHF (Köpf et al. 2024). By the end of 2023, our tabulation of the data gathered by the Center of Research on Foundation Models indicates that among the set of large language models introduced by startups and academic groups, close to 30% build on Llama or its direct descendants, Alpaca and Vicuna.

These open-source developments sent shockwaves through the major tech firms' engineering staff, as they seem to imply that these firms would struggle to maintain control of the knowledge at the core of the models they sponsored. With much trepidation, a leaked internal document from Google claimed that *"while our models still hold a slight edge in terms of quality, the gap is closing astonishingly quickly. Open-source models are faster, more customizable, more private, and pound-for-pound more capable. They are doing things with $100 and 13B params that we struggle with at $10M and 540B. And they are doing so in weeks, not months."*[18]
The impact of these investments has been amplified by the platform Hugging Face, whose vibrant developer community has catalyzed open-source innovation in foundation models by democratizing access to pre-trained models, offering access to libraries and tools, providing users with the ability to

[15] https://ai.meta.com/blog/large-language-model-llama-meta-ai/
[16] In text-to-image generation, Stable Diffusion's release in 2022 marked the first time that a foundation model was released as open source. Yet, until Llama's release, this approach had not been mirrored in the large language model (LLM) arena, where frontier advances such as GPT-4, Claude, and Cohere were API-bound, and therefore not enabling third party researchers to peek "under the hood."
[17] Quantization refers to a set of techniques to make it feasible to run powerful AI models directly on user devices, enabling real-time processing without the need for constant cloud connectivity. A cognate concept is that of model *distillation*, whereby knowledge from a large, complex model (or an ensemble of models) is transferred to a smaller, more efficient model.
[18] https://www.semianalysis.com/p/google-we-have-no-moat-and-neither

fine-tune models for specific tasks or datasets, and facilitating direct comparison of models on standardized benchmarks. In particular, Hugging Face hosts and maintains benchmark datasets and leaderboards, which are crucial for tracking the state-of-the-art and assessing the progress of foundation models over time.[19]

This natural experiment in openness demonstrates that leading firms' attempts to titrate access to their innovation by keeping some of the implementation details secret have not been successful. Less certain is whether the flourishing of the open source AI ecosystem also results in the development of technologies that contribute to the devaluation of some of the complementary assets mentioned in earlier. Even if the importance of the compute environment faded over time in shaping the performance of these models, other complementary assets, such as investments in AI safety, mounting regulatory burdens, and access to unique training data might still result in an industry with an oligopolistic structure, albeit one with a highly variegated fringe of smaller actors.

3.5 LLAMA in the ointment: Competitive scenarios with one "rogue" tech firm

Our pessimism regarding the likelihood of the emergence of a vibrant and competitive foundation models marketplace is grounded in an implicit assumption that large tech firms have incentives to close their models, multiple technical avenues to do so, and will likely exert a stranglehold on access to some essential complementary assets. But what if one "rogue" large tech firm chooses to not follow this predictable script? We mention this possibility because Meta's and founder/owner Mark Zuckerberg has publicly expressed his company's intent to pursue an open source strategy as it invests massively in the quest to achieve "General Artificial Intelligence."[20]

We can speculate on the reasons that might lead a powerful AI pioneer to choose an open source approach. That company could benefit indirectly from a vibrant ecosystem that uses its open models. For example, improvements by third parties could be integrated back into its own offerings, enhancing their value at minimal cost. By positioning itself as a leader in open AI, such a firm could significantly enhance its reputation and establish itself as a benevolent leader in technology, influencing standards and practices across the industry. Lastly, an open and collaborative approach could make that firm a magnet for top AI talent who are motivated by the prospect of working on high-impact, accessible technology.

In the scenario where Meta (or any other large tech firm) "goes rogue" and commits to a genuinely open-source model for foundational AI, the competitive landscape could be transformed in a manner akin to the early dynamics between Android and iOS in the smartphone industry. Meta's approach could democratize access to powerful AI tools, much as Android initially made smartphone technology widely accessible, enabling a broad range of manufacturers to produce smartphones that catered to various market segments. This could significantly lower the costs of exploration for startups and academic teams who might otherwise be unable to afford access to such advanced technologies, potentially leading to a surge in AI applications across various sectors.

---

[19] An important caveat is that, in the absence of well-validated benchmarks to compare the performance of different models, we should not treat claims of performance parity between open source models and large scale implementations such as OpenAI's GPT-4 or Google's BARD at face value. It may well be that recent developments democratizes access to language models useful for a broad range of niche applications, while human-like performance in general tasks remains the prerogative of firms endowed with the resources to train and retrain ever larger models using the latest, most expensive, and most energy-consuming hardware.

[20] https://www.dwarkeshpatel.com/p/mark-zuckerberg

In the smartphone market, Android's open system has served as a counterbalance to Apple's more closed and expensive iOS, fostering wider smartphone adoption globally and supporting a diverse array of application developers. This competition keeps iPhone prices somewhat in check and spurs continuous innovation relative to what might occur in a monopoly scenario. Similarly, an open-source AI model from a leading firm could encourage rapid advancements in AI applications, propelled by an open collaboration model that accelerates iteration and improvement.

However, just as Android's nominally open-source license has not resulted in a flurry of entry into the smartphone OS market, the existence of a leading open access foundation model would not by itself ensure access to the complementary assets that are necessary to build an viable product or service in the AI marketplace. Additionally, the long-term sustainability of Meta's openness "crusade" is uncertain. As in the Android ecosystem, where Google has occasionally shifted its strategies or tightened control over aspects of the platform, Meta could potentially alter the terms of access, especially if the economic or strategic benefits of the open model do not align with its broader corporate goals.

4. Implications for public policy

Given this understanding, how might policy makers think about competition policy in this setting? Below, we highlight four low hanging fruits that may prove socially beneficial without committing government authorities to wade in too deeply in the dynamics of a still nascent industry.

4.1. Credible performance benchmarks

First, it might be worthwhile for the government to coordinate the establishment of benchmarks for foundation models, both with respect to their performance as well as their safety. In spite of its crucial role in the field of Artificial Intelligence, the task of comparing foundation models is fraught with challenges, making definitive claims of superiority or advancements more complex than they appear.

One of the primary difficulties in comparing foundation models lies in the multifaceted nature of these models and the tasks they are designed to perform. LLMs and related image generation AI tools are not only evaluated based on their accuracy but also on factors like efficiency, scalability, adaptability to different languages or domains, and ethical considerations such as bias and fairness. Given this, a model that excels in one benchmark task or metric might underperform in another, complicating direct comparisons.[21] Moreover, the rapid pace of development in AI means that benchmarks themselves can quickly become outdated, as models evolve to address their limitations and new capabilities are developed. Further, given a lack of transparency around what data is used to train a model, whether a model truly "learned" to perform a particular task or whether it is simply regurgitating responses from its training data is hard to know.

[21] For example, the information retrieval and prediction needed to climb a leaderboard that ranks LLMs based on their performance on Jeopardy questions (https://github.com/aigoopy/llm-jeopardy) is likely quite different from the features that make LLMs perform well on software programming tasks (https://huggingface.co/spaces/mike-ravkine/can-ai-code-results). Furthermore, such performance metrics may not adjust for speed, cost, or model size.

Given these challenges, claims of one model being superior to another, or of open-source models catching up to their proprietary counterparts, should indeed be approached with caution. Without a comprehensive and nuanced understanding of both the benchmarks and the models being compared, such claims can be misleading. One recent example involves a careful comparison of several open-source model LLMs with the proprietary models GPT-4 and Claude 2 in the context of the Nephrology Self-Assessment Program, a multiple-choice questionnaire administered by the American Society of Nephrology to help clinicians assess their disciplinary knowledge (Wu et al. 2024). Large performance gaps emerged between the open-source models and the proprietary ones, casting doubt on the reliability of unvalidated leader boards such as those used by the leading open source repository Huggingface. Similarly, one open source project compared over 200 open source models to the six leading closed models on a variety of logic-based word questions. While the top-performing open source models produced impressive results and were competitive with the laggard closed source models, the top performing closed models (GPT-4 and Claude 3) outperformed all other models by a significant margin.[22]

This complexity and the need for nuanced comparison provide a potential role for government intervention, specifically in the development of widely accepted and well-validated benchmarks. An organization like the National Institute for Standards and Technology (NIST) could play a crucial role in this regard. NIST, with its history of developing standards across various domains, could bring a level of rigor, transparency, and neutrality to the process of benchmarking AI models. By convening a diverse set of stakeholders from academia, industry, and civil society, NIST could help ensure that benchmarks are comprehensive, up-to-date, and reflective of the broader societal impacts of AI technologies.[23]

To avoid the often-criticized incentive problems of third-party rating agencies in finance, a government-sponsored generative AI benchmarking organization might instead follow the structure of the National Renewable Energy Lab (NREL). As a government-owned, contractor operated (GOCO) organization, NREL plays an important role in the performance testing, measurement, and validation of new energy technologies like solar photovoltaics. Instead of sending prototypes of a solar cell for efficiency testing in a controlled environment, generative AI researchers could provide LLM access to these third party evaluators for a transparent set of performance tests and published results.

In the rapidly evolving domain of AI, the ability to accurately assess and compare the capabilities of different models is not just a technical necessity but also a regulatory imperative. Without clear standards, assessing the competitive landscape becomes fraught with uncertainty, making it challenging to identify actual market leaders or to evaluate claims of technological parity critically. This uncertainty may lead to either premature regulatory interventions or, conversely, a lack of action in the face of emerging monopolies.

Moreover, the fast emergence of such standards could help demystify the actual progress and capabilities of AI technologies for policymakers, businesses, and the public. It helps delineate hype

[22] https://docs.google.com/spreadsheets/d/1NgHDxbVWJFolq8bLvLkuPWKC7i_R6I6W

[23] One could argue that the US government has already heeded this advice. In February of 2024, the US Department of Commerce and NIST announced the creation of the AI Safety Institute Consortium (AISIC), which included a consortium of more than 200 large incumbent technology companies and financial institutions with the goal of developing AI safety tools and standards (https://www.reuters.com/technology/us-says-leading-ai-companies-join-safety-consortium-address-risks-2024-02-08/). The exclusive focus on safety appears misplaced to us, although a neutral arbiter like NIST offers prospects that emerging safety standards will not be tilted to favor entrenched incumbents.

from genuine advancement, ensuring that public and private resources are allocated efficiently and that innovations that truly push the boundaries of AI are recognized and fostered.

4.2. Hastening the pace of clarification of key legal issues

The development and adoption of AI foundation models, including LLMs and those capable of generating images or films, raise several legal issues that courts will likely need to assess. These issues create a degree of uncertainty regarding the returns on investments in this rapidly evolving domain. The legal challenges span the areas of intellectual property rights, privacy concerns, liability for harm, employment issues (such as the delicate interplay of non-compete agreements and trade secrecy law), and regulatory compliance, among others.

These legal challenges highlight the need for clear regulatory frameworks and legal precedents to guide the development and deployment of AI foundation models. Until these issues are more definitively addressed by courts and regulators, companies and investors in the AI domain will face ongoing uncertainty about the legal landscape and the risks associated with their AI endeavors.

Governments, including the U.S. government, can take several steps to hasten the pace of legal clarification surrounding AI foundation models without introducing comprehensive new legislation:

Wading into the use of copyrighted data for training purposes. A Supreme Court decision on fair use and AI would have a huge impact on investment incentives. The solicitor general could encourage the highest Court in the land to take on an AI fair use case. The Copyright Office could also issue guidance, as it has done on the question of whether AI output is copyrightable. Even if immediately challenged in the courts, this might force the legal system to adjudicate the underlying issues at a less leisurely pace.

Regulatory guidance and frameworks. Regulatory agencies can issue guidance, frameworks, and best practices for AI development and use. This approach can provide immediate clarity on how existing laws apply to AI technologies, including intellectual property rights, privacy, liability, and ethical considerations. For instance, the Federal Trade Commission (FTC) could offer guidance on applying consumer protection laws to AI products and services; or the FDA could develop disclosure standards that would be required when AI software is involved in patient care, or regulating the use of AI in medical devices.

Executive actions. The President and executive branch agencies can use executive orders and actions to direct federal agencies to prioritize AI governance, promote research into the impacts of AI, and encourage the development of ethical AI practices.[24] These actions can also include the creation of task forces or advisory committees to address specific AI legal and policy issues.

Judicial education and expertise. Investing in judicial education and the development of judicial expertise in technology and AI can prepare the legal system to handle disputes involving AI more

[24] In October 2023, the Biden Administration released an executive order directing agencies to focus on certain AI-related risks and encourage potential benefits (https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence).

effectively. This could include specialized training for judges and the establishment of expert advisory panels to assist courts in understanding AI technologies and their implications.

Interagency coordination. Enhancing coordination between agencies that regulate different aspects of AI (such as the Department of Commerce, the FTC, the Patent and Trademark Office, and the Department of Labor) can ensure a more coherent and comprehensive approach to AI regulation. This can help clarify legal responsibilities and reduce overlapping or conflicting regulations.

International cooperation. Engaging in international forums and bilateral agreements on AI governance can help harmonize legal standards and practices across borders. This is particularly important given the global nature of AI development and deployment, and it can reduce legal uncertainties for companies operating internationally.

These steps can help provide greater clarity and guidance for AI developers and users, addressing some of the most pressing legal uncertainties without the need for comprehensive new legislation. The options above would contribute to preparing the legal landscape for future advancements, perhaps introduced by Congress through new legislation.

4.3. Encouraging the fractionalization of infrastructure

Our discussion has leaned heavily on the importance of complementary assets in shaping the competitive landscape of the emerging foundation AI model industry. One way in which policy makers might mitigate the risks of market dominance by a few large players is to encourage the "fractionalization" of these assets so that smaller firms or academic researchers can build products and services on top of the leading technologies. In this context, fractionalization means the unbundling of the distinct activities and infrastructure required to supply and improve upon generative AI models.

The argument for policy intervention gains particular traction in areas where the market, left to its own devices, fails to ensure equitable access or distribution. While some complementary assets—such as access to Reinforcement Learning from Human Feedback (RLHF) data—are being democratized through the emergence of platforms akin to Mechanical Turk (e.g., Clickworker),25 others—such as access to training compute—remain expensive and tightly controlled by a handful of dominant players, creating significant barriers to entry and innovation. Focusing policy attention on the complementary assets most likely to elude market-led democratization efforts would amplify the effectiveness of such interventions.

Such efforts could take many forms, including investments to support innovators' access to cloud computing resources and open datasets; requirements for companies to share certain assets or make them available to competitors on the basis of fair, reasonable, and non-discriminatory terms; funding research programs in areas likely to result in the obsolescence complementary assets tightly controlled by incumbents (e.g., model distillation); promoting the development and adoption of open standards and platforms; or regulatory sandboxes that provide startup innovators the opportunity to experiment with new ideas and technologies without the burden of immediate full-scale regulatory compliance.

Beyond supplying infrastructure for generative AI research and testing, government policy might also support the fractionalization of AI assets through private-public partnerships. Modeled off of

[25] https://research.aimultiple.com/rlhf-llm/

government programs that have accelerated private industry development in frontier technologies like space launch (NASA's Commercial Orbital Transportation Service program) and fusion energy (the Department of Energy's Milestone-Based Fusion Development Program), the government might co-fund companies in the early development of cutting-edge generative AI "infrastructure" tools aimed at reducing barriers to entry, then provide subsequent larger awards to companies that achieve new milestones and publish their methods and findings. Milestones could include achievements in model compression and distillation (reducing needs for expensive compute), fine-tuning systems, performance benchmarking, detection of unauthorized model use, and reducing the cost and/or energy intensity of cloud computing.

Though governments may be the most obvious funder of fractionalized research infrastructure, leading foundation model firms may also want to fund efforts that encourage academic and startup teams to experiment using their AI infrastructure. Collaborating and opening up pieces of their LLM testing systems would enable small research teams to test ideas at much greater scale, but also allow incumbents to reinforce their competitive position by getting early insight into new technologies and potential acquisition targets. In this area, the emerging generative AI industry might take a page from the biopharmaceutical industry, in which partnerships between biotech startups and large pharmaceutical firms, combined with the sharing of pre-clinical research tools (e.g., shared lab spaces for startups) and unbundling of development activities (e.g., outsourcing of clinical trials) have encouraged startup investment by reducing barriers to entry and streamlining pathways to company acquisitions.

4.4 Platform Policy

Our discussion has been exclusively focused on competition between sponsors of AI foundation models. Distilled in a few words, our considered view is that the industry, left to its own devices, is likely to evolve an oligopolistic structure due to the stranglehold exercised by leading firms on at least some valuable complementary assets. Of course, limited competition in foundation models is compatible with the existence of one or more vibrant ecosystems of "fine-tuned" models that build on top of the infrastructure layer provided by foundation models. Such a platform evolution is portended by OpenAI's fledgling store of third party plugins.

Just as in the example of application stores for smartphone operating systems, competition policy challenges are sure to arise to limit the bargaining power of the caretakers of these ecosystems—the companies controlling the foundation models—vis à vis third party fine-tuned model/application providers. For example, how proprietary should the APIs that allow application developers to develop on top of foundation models be? Should foundation model sponsors be allowed to avail themselves of the data used by third party developers to fine-tune these models for their particular use case? And how unfettered should the power of model sponsors be in setting royalty rates and access fees to capture value generated by third-party developers?

5. Conclusion

Our analysis emphasizes how the technology and complementary assets of the nascent generative AI industry might favor a small set of leading firms as they build up a war chest of both physical assets and specialized know-how. Naturally, predicting another concentrated technology industry might suggest preemptive policy guardrails against market power or spark calls to subsidize entrants or offer

government-created "public option" models to check the private sector oligopolists.[26] However, if AI innovation policy is to foster progress and technological leadership through a creative, diverse and commercially viable ecosystem, that outcome does not strictly require many competitors, and maximizing innovation incentives might even be in-conflict with free entry. The economics of innovation literature highlights the precarious or "inverted U-shaped" relationship between the number of competitors and incentives for innovation: not enough competition and firms exploit their market power without the urgency to invest in disruptive R&D; too much competition and profits erode, lowering innovation returns for all (Aghion et al, 2005;Segal & Whinston, 2007).

Thus, any policy efforts face the high wire act of fostering intense competition, without dulling incentives for engaging in risky R&D. That daunting challenge is why our policy recommendations emphasize reducing barriers to experimentation–such as legal and regulatory uncertainty, milestone based private-public partnerships, public benchmarking, and access to AI research infrastructure. Instead of bluntly forcing incumbents to license/share their models for the sake of reducing consumer prices, generative AI innovation policies would ideally encourage a robust ecosystem of academic researchers and firms of different types competing to test and prove next-generation AI models and tools. Through policies that are agnostic about who ultimately brings the winning ideas to market (e.g., encourage collaboration through shared resources and enforce intellectual property rights), innovators can strive to complement or displace the leading firms' "dominant design" transformer-based models, rather than simply duplicating the production of similar foundation model efforts.

A related challenge more specific to the complex and rapidly evolving new frontier of AI foundation models is that U.S. policymakers find themselves navigating between the Scylla of fostering a competitive market and the Charybdis of ensuring technological safety. This dichotomy presents a nuanced challenge, reflecting a policy approach that can appear, at times, to be marked by a certain ambivalence, sometimes bordering on a dissociative identity disorder.

On one hand, there is a pronounced concern among policymakers regarding the potential for market domination by a select few large tech companies, such as Meta, Google, and Microsoft. The apprehension is that without intervention, these behemoths could monopolize the AI field, stifling innovation and controlling the direction of future technological developments. This perspective champions the idea of breaking down barriers to entry and encouraging a vibrant ecosystem where startups and academic teams innovate freely, thus preventing any single entity from wielding disproportionate influence over the AI landscape.

Conversely, there is a palpable sense of caution permeating policy discussions, primarily centered on the safety and ethical implications of AI technologies. The rapid pace of innovation, especially among smaller startups with fewer resources to dedicate to ethical considerations, raises concerns about the potential for harmful applications, such as the creation and proliferation of deep fakes. Here, the larger corporations, with their established reputations and experience with operating under regulatory scrutiny, are seen as entities that might be more accountable and responsible stewards of powerful AI technologies.

The reaction to the "Model Forum" convened by leading AI providers to develop safety standards represents a quintessential example of the ambivalence permeating the AI policy landscape. On one hand, this initiative could be lauded as a commendable example of self-regulation, reminiscent of the

[26] https://foreignpolicy.com/2023/06/12/ai-regulation-technology-us-china-eu-governance/

Asilomar conference, which played a pivotal role in addressing the ethical and safety concerns surrounding recombinant DNA technology (Berg et al. 1975; Frederickson 1991). This perspective views the forum as a proactive step towards ensuring AI technologies are developed and deployed responsibly, mitigating risks and fostering public trust. On the other hand, a more skeptical view might interpret the same forum as a strategic maneuver by incumbent firms to solidify their dominance. By setting the bar for compliance with safety standards they themselves crafted, these firms could inadvertently (or perhaps, tactically) raise the operational and financial hurdles for emerging and future competitors, potentially stifling innovation and entrenching their market position.

This ambivalence presents a conundrum for crafting policy recommendations aimed at fostering competition within the AI domain. In this article, we have argued that delineating a competition policy framework requires a nuanced understanding of the AI ecosystem, one that is grounded in the twin pillars of appropriability and complementary assets. The power of this conceptual framework, however, only becomes compelling if this sector is approached by policy makers like any other nascent industry. Rather than reinventing the competition policy wheel for the AI age, we have argued that there is nothing inherently special about this domain. While there is always a risk of constraining the interpretation of new innovation phenomena within the boundaries of existing paradigms, our analysis offers a counterpoint to the fevered claims and policy proposals advanced by industry insiders, as well as their detractors.

References

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Dijk, A. A. van, Ebrecht, A. C., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. https://doi.org/10.1126/science.abj8754

Berg, P., Baltimore, D., Brenner, S., Roblin III, R. O., & Singer, M. F. (1975). Asilomar conference on recombinant DNA molecules. *Science*, *188*(4192), 991–994.

Bertoline, L. M. F., Lima, A. N., Krieger, J. E., & Teixeira, S. K. (2023). Before and after AlphaFold2: An overview of protein structure prediction. *Frontiers in Bioinformatics*, *3*. https://doi.org/10.3389/fbinf.2023.1120370

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., & Liang, P. (2023). The foundation model transparency index. *arXiv Preprint arXiv:2310.12941*.

Bresnahan, T. (2024). What innovation paths for AI to become a GPT? *Journal of Economics & Management Strategy*, *33*(2), 305–316. https://doi.org/10.1111/jems.12524

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'Engines of growth'? *Journal of Econometrics*, *65*(1), 83–108.

Du, T., Chang, K.-H., Zhang, R., & Liu, P. (2024). *Neural network categorization accuracy with categorical graph neural networks* (U.S. Patent and Trademark Office Patent US 11,960,573 B1). https://patents.google.com/patent/US11960573B1/en?oq=11960573

Foster, R. (1986). *Innovation—The attacker's advantage*. Summit Books.

Frederickson, H. G. (1991). Toward a theory of the public for public administration. *Administration & Society*, *22*(4), 395–417.

Gans, J. S. (2024). *Copyright Policy Options for Generative Artificial Intelligence*. National Bureau of Economic Research.

Giczy, A. V., Pairolero, N. A., & Toole, A. A. (2022). Identifying artificial intelligence (AI) invention: A novel AI patent dataset. *The Journal of Technology Transfer*, *47*(2), 476–505.

Hasani, R., Lechner, M., Amini, A., Rus, D., & Grosu, R. (2021). Liquid Time-constant Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(9), 7657–7666. https://doi.org/10.1609/aaai.v35i9.16936

Henderson, R. (1995). Of life cycles real and imaginary: The unexpectedly long old age of optical lithography. *Research Policy*, *24*(4), 631–643. https://doi.org/10.1016/S0048-7333(94)00790-X

Hoffmann, J., Avocat, S. B. D., Sifre, L., & Mensch, A. (2023). *Allocating computing resources between model size and training data during training of a machine learning model* (Patent US 20,230,315,532 A1). https://patents.google.com/patent/US20230315532A1/en?oq=US+2023%2f0315532+A1

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L.,

Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). *Training Compute-Optimal Large Language Models* (arXiv:2203.15556). arXiv. http://arxiv.org/abs/2203.15556

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv Preprint arXiv:2106.09685*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., & others. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models* (arXiv:2001.08361). arXiv. http://arxiv.org/abs/2001.08361

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., & Raileanu, R. (2024). Understanding the effects of rlhf on llm generalisation and diversity. *arXiv Preprint arXiv:2310.06452*.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z. R., Stevens, K., Barhoum, A., Nguyen, D., Stanley, O., Nagyfi, R., & others. (2024). Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, *36*.

Lemley, M. A., & Casey, B. (2020). Fair learning. *Tex. L. Rev.*, *99*, 743.

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–6.

Luo, H., & Mortimer, J. H. (2017). Copyright Enforcement: Evidence from Two Field Experiments. *Journal of Economics & Management Strategy*, *26*(2), 499–528.

Ray, A., Achiam, J., & Amodei, D. (2019). *Benchmarking Safe Exploration in Deep Reinforcement Learning*. *7*.

Ray, J., Ashbaugh, B., Surti, P., Ramani, P., Harihara, R., Justin, J. C., Huang, J., Cui, X., Costa, T. B., Gong, T., Ould-Ahmed-Vall, E., Balasubramanian, K., Thomas, A., Elibol, O. H., Bobba, J., Zhuang, G., Subramanian, B., Keskin, G., Sakthivel, C., & Poornachandran, R. (2020). *Compression in machine learning and deep learning processing* (U.S. Patent and Trademark Office Patent US 10,546,393 B2). https://patents.google.com/patent/US10546393B2/en?oq=US10546393B2

Schrepel, T., & Pentland, A. (2023). *Competition Between AI Foundation Models: Dynamics and Policy Recommendations*.

Shapiro, C., & Varian, H. R. (1999). *Information rules: A strategic guide to the network economy* (Nachdr.). Harvard Business School Press.

Szegedy, C., & Vanhoucke, V. O. (2017). *Processing images using deep neural networks* (U.S. Patent and Trademark Office Patent US 9,715,642 B2).

https://patents.google.com/patent/US9715642B2/en

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, *15*(6), 285–305.

Tripsas, M. (1997). Unraveling the Process of Creative Destruction: Complementary Assets and Incumbent Survival in the Typesetter Industry. *Strategic Management Journal*, *18*, 119–142. JSTOR.

Utterback, J. M. (1994). *Mastering the Dynamics of Innovation*. Harvard Business School Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Widder, D. G., West, S., & Whittaker, M. (2023). Open (for business): Big tech, concentrated power, and the political economy of open AI. *Concentrated Power, and the Political Economy of Open AI (August 17, 2023)*.

Wu, S., Koo, M., Blum, L., Black, A., Kao, L., Fei, Z., Scalzo, F., & Kurtz, I. (2024). Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology. *NEJM AI*, *1*. https://doi.org/10.1056/AIdbp2300092

Zhang, Y., Cui, L., Cai, D., Huang, X., Fang, T., & Bi, W. (2023). Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *arXiv Preprint arXiv:2305.13225*.

Zucker, L. G., Darby, M. R., & Armstrong, J. S. (2002). Commercializing Knowledge: University Science, Knowledge Capture, and Firm Performance in Biotechnology. *Management Science*, *48*(1), 138–153. JSTOR.