Integration of Generative AI in the Digital Markets Act: Contestability and Fairness from a Cross-Disciplinary Perspective

Ayse Gizem Yasar
London School of Economics

Andrew Chong
University of California

Evan Dong
Cornell University

Thomas Krendl Gilbert
The New York Academy of Sciences

Sarah Hladikova
Tufts University

Carlos Mougan
University of Southampton
The Alan Turing Institute

Xudong Shen
National University of Singapore

Shubham Singh
University of Illinois Chicago

Ana-Andreea Stoica
Max Planck Institute for Intelligent Systems, Tübingen

Savannah Thais
Columbia University

Abstract

The EU's Digital Markets Act (DMA) aims to address the lack of contestability and unfair practices in digital markets. But the current framework of the DMA does not adequately cover the rapid advance of generative AI. As the EU adopts AI-specific rules and considers possible amendments to the DMA, this paper suggests that generative AI should be added to the DMA's list of core platform services. This amendment is the first necessary step to address the emergence of entrenched and durable positions in the generative AI industry.

This is a working paper and will be subject to changes before publication. Please do not cite without the authors' permission.

Last update: March 2024

Corresponding author: Ayse Gizem Yasar (a.g.yasar@lse.ac.uk)

Contents

1 Introduction

The European Union's response to platform size and power in digital markets goes beyond traditional competition law enforcement. Under EU competition law, historically, large companies and market power have not been viewed as inherently problematic. It is only when market power is abused that EU competition law intervenes. However, there is an increasing concern that 'a few large platforms act as gatekeepers between businesses and users and enjoy an entrenched and durable position, often as a result of the creation of conglomerate ecosystems around the core platform services, which reinforce existing entry barriers.' (European Commission, 2020c) The underlying logic is that these entrenched positions lead to unfair behaviour vis-à-vis business users of these platforms, as well as reduced innovation and contestability in core platform services.

These concerns led to the creation of a new European regulation that goes beyond traditional competition law rules: the Digital Markets Act (DMA). The DMA is set up to counteract platform size rather than abuse of dominance or monopoly power that competition laws target. It aims to address the shortcomings of competition law in keeping the entry barriers low and ensuring fair game between 'gatekeepers' and their smaller rivals that depend on gatekeepers' services. The regulation covers some well-established platform services like operating systems, messaging platforms, and online advertising. The DMA is silent on AI, but there is an increasing concern that gatekeepers are likely to emerge in generative AI applications.

In this paper, we argue that generative AI should be included in the DMA as a core platform service. Our argument relies on an analysis of the gatekeeping potential of companies that develop and deploy generative AI models. We argue that generative AI services cannot be categorised as one of the current examples of core platform services under the DMA and must therefore obtain a category of their own. We illustrate ways in which certain generative AI companies may undermine fairness and contestability as defined in the DMA. We draw on the landscape of existing generative AI research (discussing both existing and incipient characteristics of generative AI technology), as well as characteristics of digital markets and past business practices by core platform service providers that led to the adoption of the DMA.

The paper is structured as follows: In Section 2, we provide the legal and technological foundations of our proposal. A brief overview of the goals of the DMA is followed by the qualitative and quantitative criteria for designating gatekeepers. In the absence of a clear definition of 'core platform service' in the DMA, we identify and detail three cumulative conditions under which a service can be added to the DMA's list of core platform services, namely: (1) constituting a gateway between business users and end users, (ii) exhibiting certain characteristics that facilitate the emergence of gatekeepers, and (iii) being sufficiently distinguishable from existing core platform services. In light of the EU's recently-passed AI Act, we also discuss differences between the DMA and AI Act, highlighting that they cover fundamentally different aspects of technology regulation. More specifically, we discuss why the AI Act does not address the same kind of fairness and contestability issues sought by the DMA.

In Section 3, we describe the landscape of generative AI in the context of online platforms. We survey current generative AI models and describe the pathways through which they become *gateways* between businesses and users, as defined by the DMA. More specifically, we identify the ways in which certain types of generative AI models emerge as *novel types of digital platforms*, fundamentally different from existing platforms.

We note that being a gateway is a necessary but not sufficient condition to become a gatekeeper. However, we argue in Section 4 that companies that develop large generative AI models strongly exhibit four main characteristics of digital services which facilitate the emergence of gatekeepers: first-mover advantage, capital-intensive resources, data resources, and system integration. First, we illustrate how a first-mover advantage functions in the generative AI sector through setting industry benchmarks and rapid scaling strategies. The operative distinction here is that a first developer of a generative AI model may not *necessarily* benefit from a first-mover advantage, unless it can follow up with rapid expansion. Second, we employ an analysis of the computing requirements for the development of current generative AI models. We show that the concentration of computing power in the hands of a few is directly linked to concerns of contestability and fairness as defined in the DMA. Third, data resources represent a powerful input that is not available for all. Indeed, large and integrated companies developing generative AI models benefit from troves of data unavailable to smaller actors. Finally, we argue that a variety of companies are opting for integrated services, combining search, email, personal assistants, and many others, with generative AI models through a user interface. We link this phenomenon with contestability issues as defined by the DMA.

In light of these concerns, we outline a regulatory proposal in Section 5, arguing that the DMA has the potential to address contestability and fairness in the generative AI industry if it is amended to include generative AI as a core platform service. Even if gatekeeper designations are possible in cases where the qualitative criteria are met while the quantitative criteria are not, we draw on both sets of criteria to support our proposal. Finally, we provide examples of existing gatekeeping obligations that would apply to generative AI if it were directly included in the list of core platform services.

We conclude in Section 6 with an overview of our argument and final suggestions, noting open directions for research at the intersection of law, policy, and computational fields of study.

2 Foundations

2.1 Principles of the Digital Markets Act

Regulation 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector, also known as the 'Digital Markets Act', came into force on 1 November 2022. It is the result of a recognition by the EU legislature that existing European competition law rules, established under Articles 101 and 102 of the Treaty on the Functioning of the European Union (TFEU), were insufficient to address forms of economic power associated with certain features of digital markets. In particular, many digital services were thought to have characteristics that had a tendency to result in unfair practices and a lack of contestability. In the run up to its DMA proposal in December 2020, the European Commission ('Commission') announced that it would explore '**ex ante rules** to ensure that markets characterised by large platforms with significant network effects acting as gatekeepers, remain fair and contestable for innovators, businesses, and new market entrants', which culminated in the DMA (European Commission, 2020b).

The DMA is built on two new regulatory categories: 'core platform services' and 'gatekeepers'. 'Core platform service' is not defined in the DMA, but a list of core platform services is provided in Article 2(2), comprised of services with a 'capacity to affect a large number of end users and businesses' (Recital 14). The DMA outlines numerous characteristics that providers of 'core platform services' in the digital sector exploit. These include scale economies, strong network effects, the ability to intermediate between multiple groups of users, high dependence of users on platforms, lock-in effects, and other data-driven advantages (Recitals 2, 13, 25). As a result, some providers of core platform services reach gatekeeping positions. Importantly, such characteristics and their likelihood in resulting in unfair practices and a lack of contestability are thought to exist even where a company or entity does not enjoy a 'dominant position'—a traditional requirement in EU competition law to apply Article 102 TFEU on abuse of dominance. Article 3 of the DMA, further discussed in the next subsection, sets clear and unambiguous criteria for the designation of gatekeepers. If a company is designated as a gatekeeper, they are automatically obligated to comply with the various rules imposed by the DMA for their relevant core platform services.

Meanwhile, the AI Act is the first comprehensive AI regulation in the world, targeting a broad range of AI applications. The EU legislature reached a political agreement on the AI Act in December 2023, and it is expected to formally become law in the first half of 2024. The AI Act, as proposed by the Commission, was focused on mitigating risks associated with some applications of AI based on a four-tiered risk framework. Under this framework, which is retained in the latest version of the AI Act, the use of AI with unacceptable risk, such as government use of automated social scoring of its citizens, is banned. The framework establishes specific obligations for AI applications that present high, limited, and minimal risks to health, safety, and fundamental rights of EU citizens. While the high-risk category applications need to conduct conformity assessments to ensure all the obligations are met, the last two categories are left mostly unregulated.

Reflecting on the enormous success of OpenAI's ChatGPT in November 2022, the European Parliament proposed new AI Act provisions that address AI systems which have multiple use cases and hence cannot be easily categorised within the existing four tiers. This category now features in the AI Act after lengthy deliberations with the Council and is called 'general-purpose AI'. While the original proposal from the Commission was focused on determining the risk of different applications based solely on their purpose (for example, applications that were built for the purposes of employment or law enforcement were classified as high risk), following the amendments to the draft, the potential capabilities of general-purpose AI systems will also need to be considered by their providers. The many use cases of generative AI models motivated the MEPs to expand the four-tiered approach. The new section addressing general-purpose AI includes obligations for the providers of foundation models to guarantee 'robust protection of fundamental rights, health and safety and the environment, democracy and the rule of law', as well as a 'need to assess and mitigate risks, comply with design, information and environmental requirements and register in the EU database' (European Parliament, 2023). Generative AI built on top of foundation models would also have to comply with transparency requirements to mitigate the risk of misinformation from generating 'deep-fakes' in image and video outputs.

The deliberate inclusion of general-purpose AI and foundation models in the AI Act aims to mitigate its distinctive, systemic risks to EU citizens and institutions. It does not, however, address the risks of contestability and any imbalances of bargaining power between generative AI companies and their business users, as this is not within the scope of the AI Act.

2.2 Core Platform Services and Gatekeeper Obligations

The DMA targets 'gatekeeper' companies. A gatekeeper is defined under Article 3(1) as an undertaking that: *(i)* 'has a significant impact on the [EU's] internal market', *(ii)* 'provides a core platform service which is an important gateway for business users to reach end-users', and *(iii)* 'enjoys an entrenched and durable position in its operations, or it is foreseeable that it will enjoy such a position in the near future'. 'Gateway' is broadly construed in the DMA: it covers any platform service used by a business to provide goods or services to any end users. The latter excludes business users.

Article 3(2) provides the corresponding quantitative criteria. If an undertaking meets the quantitative thresholds for any of its core platform services, it is presumed to satisfy the qualitative criteria of Article 3(1) and is therefore designated a 'gatekeeper'. The designation concerns only the core platform services meeting the gatekeeping criteria.

The current list of 'core platform services' provided in Article 2(2) covers ten established digital services.[1] This list of core platform services does not explicitly cover generative AI services. Certain cases of generative AI applications will indirectly fall within the remit of the DMA when gatekeepers integrate proprietary or third-party generative AI into their core platform services, such as search engines. However, the DMA's gatekeeper obligations (Articles 5, 6 and 7) designed to counteract contestability and fairness issues will not apply to their generative AI offering when it is provided as a standalone platform service. In the remainder of this paper, we present ways in which generative AI is provided as a platform service and argue for its explicit integration into the DMA.

Nearly two years passed between the Commission's DMA proposal in December 2020 and its entry into force (this was relatively quick for the EU's law-making process). The European Parliament and the Council of the European Union, the EU's other two legislative bodies, decided their positions largely in 2021. The DMA reached its near-final shape in March 2022 (Council of the EU, 2023). Generative AI models have been gaining traction since 2018, but arguably it was not until the launch of ChatGPT in late 2022 that the public at large and policymakers became aware of the significance of generative AI. By then, the DMA was already in place.

The drafters of the regulation foresaw that the DMA would need to be amended with technological developments and put in place mechanisms to keep the regulation up to date. Recital 77 states: 'The services in the digital sector and the types of practices relating to these services can change quickly and to a significant extent', which has been the case for generative AI services. Article 19 provides the procedure that the Commission needs to follow to add/drop core platform services. We explain this process below and lay out the conditions under which a new core platform service should be added to the DMA.

[1] Such as online intermediation services; online search engines; online social networking services; video-sharing platform services; number-independent interpersonal communications services; operating systems; web browsers; virtual assistants; cloud computing services; and online advertising services. The first round of designations by the Commission includes six gatekeepers and eight different categories of core platform services.

2.3 Expanding the DMA's List of Core Platform Services

The DMA does not define 'core platform service'. Article 2 ('Definitions') merely provides the list of core platform services without a specific definition. Likewise, Article 19 gives a lot of leeway to the Commission in its examination of potential new core platform services and does not provide any specific criteria. The Commission, in its press release concerning the first designation of six gatekeepers, offered the following definition:

> Core platform services are those services in the digital economy that exhibit certain features and where absent regulatory intervention the identified failures would effectively remain un-addressed. Such features entail highly concentrated services, where usually one or very few large digital platforms set the commercial conditions with considerable autonomy and where few large digital platforms act as gateways for business users to reach their customers. (European Commission, 2023)

Furthermore, while DMA does not provide a functional definition of core platform services, it does delineate characteristics of the core platform services that it aims to address (Recitals 2 and 14). These characteristics include 'extreme scale economies', 'very strong network effects', 'a significant degree of dependence of both business users and end users', 'lock-in effects', 'a lack of multi-homing', 'vertical integration', and 'data driven-advantages' (Recitals 2 and 14). As such, the DMA arguably conflates 'core platform service' and 'gatekeeper': many of these characteristics are also listed under Article 3(8) concerning Commission decisions to designate a gatekeeper where the quantitative thresholds are not met.

The DMA clearly outlines a mechanism to add new or remove existing core platform services in Article 19. For the purpose of examining whether one (or more) digital services should be added to the list of core platform services, the Commission can open a market investigation. The Commission must then publish a report within 18 months and submit a legislative proposal to the European Parliament and the Council. From there, the amendment enters the ordinary legislative procedure. Adding new core platform services to the DMA is 'a core part of keeping the regulation future-proof' (European Commission, 2020a). The dynamic nature of digital markets requires that the Commission maintains regular, continuous monitoring of its impact. Therefore, there are no legislative barriers to adding new core platform services to the DMA.

The expansion of the list of core platform services has already transpired during the inter-institutional negotiations. At first, the DMA proposal (European Commission, 2020c) listed eight out of the current ten core platform services: online intermediation services, online search engines, social networking, video sharing platform services, number-independent interpersonal electronic communication services, operation systems, cloud services, and advertising services. Subsequently, the final text (European Commission, 2022b) included two additional core platform services: web browsers and virtual assistants. The decision to include virtual assistants was based on the Commission's sector inquiry into the consumer Internet of Things (IoT) (European Commission, 2022a), which found competition concerns such as high entry barriers and difficulties in competing with companies that are vertically integrated and form their own ecosystems.

Based on the DMA, the Impact Assessment, and the Commission's press releases, we propose that the following three cumulative criteria need to be met for the Commission to designate a new core platform service.

First, the prospective digital service should constitute a **gateway for business users to reach end users**. The DMA excludes platforms that have no role in bringing together business users with end users. While multi-sided markets such as search engines, video-sharing platforms, and operating systems are canonical examples of gateways that directly intermediate transactions between businesses and consumers, the concept of gateway is construed more broadly than match-making in the DMA. For example, it designates cloud computing service as a core platform service because although it does not match-make transactions, it provides the infrastructure to support and enable business users to reach customers.

Second, the service or the market in which it operates must exhibit certain problematic **characteristics that facilitate the emergence of gatekeepers**. Drawing on the DMA and the Impact Assessment, these characteristics include strong network effects which make markets prone to 'tipping', extreme scale economies, vertical integration, data-driven advantages, and lack of multi-homing. These characteristics are largely drawn from past antitrust cases, sector inquiries, and academic (especially economic) literature.

Finally, at the risk of stating the obvious, the new core platform service should be **sufficiently distinguishable from existing core platform services**. This, however, does not exclude the possibility of distinguishing between different core platform services even when they are, at times, provided in an integrated fashion. For example, Google Shopping is provided both as a distinct service and within Google Search. Alphabet was designated as gatekeeper separately for these two core platform services.

In the two sections that follow, we first discuss the platformisation of generative AI, where certain AI systems, distinct from the DMA's existing core platform services, now act as a gateway between business users and end users (Section 3); we then move on to four main characteristics of generative AI systems which facilitate the emergence of gatekeepers: first-mover advantage, capital-intensive resources, data resources, and system integration (Section 4).

3 Platformization of Generative AI

Generative AI refers to a specific category of AI systems which are able to generate new data based on a previously available training dataset. For example, they can generate types of content, such as text, image, video, audio, and computer code that might be be indistinguishable from content created by humans, experts in specific fields, or those encountered in the real world. For example, AI chatbots such as OpenAI's ChatGPT and Anthropic's Claude can converse with users naturally by generating text in response to user queries. Text-to-image generative models such as Stable Diffusion and Dall-E can generate images based on text descriptions. Code generation tools, such as Amazon CodeWhisperer, are also emerging.

Generative AI has enabled novel digital applications and significantly enhanced existing ones. The technologies underlying generative AI systems and generative AI applications are increasingly used as platform infrastructure by business users. As a result, certain generative AI technologies and applications are starting to act as gateways between business users and end users within the meaning of the DMA. Below, we first discuss the transformation of generative AI into platform infrastructure, followed by use cases where generative AI augments existing services.

3.1 Digital Platform Infrastructure

A few companies in the generative AI industry are developing the critical infrastructure for generative AI-enabled services. Central to this infrastructure is what is known as a foundation model. These are large-scale AI models that have been pre-trained on diverse data at scale and are designed to be versatile. They can be adapted for a variety of specific tasks through fine-tuning, prompt engineering, or other customization techniques (Bommasani et al., 2021). Analogous to how cloud computing provides 'infrastructure to support and enable functionality in services offered by others' (European Commission, 2020a), foundation models function as the infrastructure for a wide range of generative AI applications and services because they provide the fundamental capabilities of general, generative artificial intelligence.

Within the realm of foundation models, a distinction is often made between *frontier* models and *behind-the-frontier* models. Frontier models, as defined by the Frontier Model Forum,[2] are 'large-scale machine-learning models that exceed the capabilities currently present in the most advanced existing models' (Microsoft, 2023b). On the other hand, models that do not reach this level of capability are considered behind-the-frontier models. This distinction is crucial because frontier models can suddenly obtain capabilities that behind-the-frontier models lack, a phenomenon known as emergent abilities (Wei et al., 2022; Schaeffer et al., 2024). Such emergent abilities give frontier models a considerable competitive advantage, positioning them as essential infrastructure that offers functionalities unattainable by other models.

Key to this process is the adaptability of foundation models, meaning that developers of generative AI applications need not develop their own models. Instead, a pre-trained foundation model can be

[2]The Frontier Model Forum was established by OpenAI, Anthropic, Google, and Microsoft. See https://openai.com/blog/frontier-model-forum

built 'on top of' to develop various generative AI applications. For example, OpenAI's Generative pre-trained transformer (GPT) models, besides underpinning OpenAI's own chatbot product ChatGPT, have also been customized by startups like Harvey to assist lawyers in accessing and generating legal content (Grady & Curnin, 2023), and by Jasper to aid in marketing content creation (Jasper, 2023). This paradigm has fostered a delineation between the platform layer—which provides foundational models—and the application layer, which focuses on building user-facing products.

Consequently, the foundation model emerges as a novel type of digital platform that provides essential generative AI functionalities for business users to deliver generative AI applications to end users. The development of (frontier) foundation models is often prohibitively expensive for individual developers, as further discussed in Section 3. For example, the cost of training GPT-4 is estimated to be over 100 million USD (The Economist, 2023). Another major LLM company, Anthropic, has planned to spend one billion USD to build the next generation LLM (Wiggers et al., 2023). This foundation model-as-a-platform approach democratizes access to generative AI technologies for small and medium-sized developers, significantly reducing the up-front cost and the technical complexities associated with the development and integration of foundation models.

Establishing foundation models as platforms is becoming a strategic imperative across the generative AI industry, especially for the developers of frontier models. For example, OpenAI provides (1) API access (OpenAI, 2020) to proprietary GPT models, (2) GPTs, which enable lightweight customization of ChatGPT for specific purposes (further discussed in the next sub-section), and (3) a developer product called Foundry (Wiggers, 2023), which allows for the fine-tuning of GPT models and static allocation of compute resources to support large-scale inference. In a similar vein, both Amazon Bedrock (Amazon, 2023) and Google Vertex AI (Google, 2023) offer a broad collection of foundation models along with the tools required to develop generative AI applications. Other contenders in this space include IBM's Watsonx.ai (IBM, 2023) and Azure Machine Learning (Microsoft, 2023). Scale's generative AI platform targets enterprise customers and positions itself as a full-stack platform for powering Generative AI applications (Scale AI).

3.2 Augmented Services

The capabilities of generative AI come alive when interacting with users. This is possible through a new breed of user interfaces such as Google's Gemini, which integrate inert model capabilities into the behavioral affordances of a system. As these capabilities become dynamic and multimodal, combining various tasks at once, the user experience in digital services is transformed. This qualitatively raises the stakes for legacy providers across various service domains such as search engines.

Chatbots are the prime example of how traditional digital services are being augmented by generative AI. AI chatbots comprise a type of computer interface that leverages AI technology to take user queries as input and return responses in real time, simulating the flow of human conversation. While traditional chatbots were rule-based and served specific use cases, such as the first-ever chatbot Eliza that simulates conversations with a psychotherapist (Weizenbaum, 1976), generative AI-powered chatbots mark a significant leap forward. They display a comprehensive ability to simulate the dynamics of human conversation over a wide range of topic domains, and are adept at generating relevant responses to open-ended queries. ChatGPT exemplifies this new breed of chatbots.

The significance of chatbots lies not just in their technical proficiency at simulating human linguistic patterns, but in the material threat they pose to market actors who offer the services they augment. This is most clear in the panic ChatGPT caused to Google and Microsoft, which had been developing their own generative AI models internally for years but had not released them publicly. OpenAI's success was in developing a user interface (ChatGPT) that, when paired with a large language model, became a compelling alternative to the web search interfaces that have defined the internet experience over the past quarter century. Google acted quickly, merging DeepMind and Google Brain in April 2023 to become Google DeepMind, while Microsoft immediately made a multibillion-dollar investment in OpenAI itself and integrated ChatGPT's underlying language model into its Bing web search tool.

There are signs that generative AI chatbots and, more broadly, generative AI applications are turning into gateways for business users to reach end users. A significant number of institutions have started to employ generative AI models in servicing users for a variety of applications, including in the European Union (Macaulay, 2023). Generative AI has already become commonplace as a tool used by businesses to serve customers in both small businesses (Soni, 2023) and large businesses (Kanbach et al., 2023). For example, Adobe Sensei use genAI integration in creative software tools (Insights, 2023; Adobe, 2023), whereas Siemens integrates genAI in medical diagnoses services (Siemens, 2023).

In keeping with this trend, OpenAI has enhanced ChatGPT by introducing third-party plugins. These plugins are typically developed by businesses that specialize in specific digital services. For instance, Expedia and KAYAK have created plugins to integrate with their own websites, enabling ChatGPT users to search for flights and accommodations. While these plugins extend the capabilities of ChatGPT, they also strengthen its role in directing and managing traffic for business users, making ChatGPT more platform-like. OpenAI has also introduced another product called GPTs, which allows developers to customize ChatGPT for specific purposes. Among the initial GPTs released are the Canva chatbot, which assists users in creating personalized designs, and AI Actions by Zapier, offering access to Zapier's applications for developers who build custom GPTs (OpenAI, 2023b). OpenAI also established a GPT Store for distributing custom GPTs. Against this background, OpenAI's gateway role is twofold: Firstly, the original ChatGPT serves as the generative AI infrastructure, akin to how cloud computing provides the necessary infrastructure to support and enable the functionality of various services. Secondly, the introduction of the GPT Store marks OpenAI's strategic shift towards a platform-centric ecosystem. While the monetization pathways are still being defined, OpenAI's CEO, Sam Altman, has indicated '[w]e're going to pay people who make the most used and most useful GPTs [i.e., customized ChatGPTs for specific purposes] with a portion of our revenue' during OpenAI's first Developers Day (Coldevey, 2023).

Chatbots raise the question of whether they could be categorised as online search engines or virtual assistants, the two most closely related core platform services under the DMA. Art. 2(6) DMA defines 'online search engine' by referring to Art. 2(5) of the platform-to-business relations regulation (European Union, 2019) (P2B Regulation), which in turn defines it as

> A digital service that allows users to input queries in order to perform searches of, in principle, all websites, or all websites in a particular language, on the basis of a query on any subject in the form of a keyword, voice request, phrase or other input, and returns results in any format in which information related to the requested content can be found. (European Union, 2019)

This definition does not restrict the format of the search results and only requires that 'information related to the requested content can be found', thereby probably encompassing the conversational responses of chatbots. However, users do not engage with AI chatbots with the intent to 'perform searches of, in principle, all websites'. Nor do chatbots have to search the web to generate their responses. They can produce relevant information directly through the content generation processes. Although AI chatbots commonly feature integrated web search capabilities, these functions are implemented as distinct services, or plugins (OpenAI, 2023a), which chatbots can utilize to produce up-to-date and comprehensive responses. These distinctions set chatbots apart from the DMA's definition of online search engines.

Although there might be a closer conceptual resemblance, chatbots also diverge from virtual assistants in important ways. According to Art. 2(12) DMA, a virtual assistant is 'a software that can process demands, tasks or questions, including those based on audio, visual, written input, gestures or motions, and that, based on those demands, tasks or questions, provides access to other services or controls connected physical devices'. While chatbots do process user prompts in terms of 'demands, tasks or questions', their primary function is neither to 'provide access to other services' nor to 'control connected physical devices', despite their capacity to perform these tasks. Fundamentally, what generative AI chatbots provide is *the generated responses (or, more broadly, data) themselves*, rather than access to other distinct services. The DMA defines virtual assistants in this way because it is motivated by virtual assistants in the IoT industry, such as Amazon's Alexa and Google's Assistant. The drafters of the DMA did not foresee the emergence of generative AI-based chatbots. There is some room to argue that to the extent a generative AI chatbot is designed to refer users to other services, it falls under the virtual assistant definition. However, to the extent generative AI applications serve as infrastructure to build other generative AI applications, they fall outside the scope of 'virtual assistants'.

4 Generative AI and Core Platform Service Characteristics

The generative AI sector has recently experienced significant growth, with both established tech companies and myriad startups entering the market to develop and/or deploy foundation models and new applications in generative AI. Some of these companies hold a significant competitive advantage due to their extensive data resources, specialized hardware architectures, vertical integration, previous talent acquisition, financial clout, and accumulated knowledge through prior product releases. Many of these advantages—such as data resources, financial clout, integration, and the importance of early movers—have been present in digital markets in general and given rise to the gatekeeping positions that the DMA now seeks to address.

As emphasised in the DMA and the Impact Assessment, these advantages are exacerbated by network effects. Generative AI systems exhibit strong network effects, albeit in ways different than network effects in traditional platforms. The key for any generative AI system to provide high-quality service is data: 'To remain competitive, an AI operator must corral data, analyze it, offer predictions, and then seek feedback to sharpen the suggestions. The value of the system depends on—and increases with—data that arrives from users' (Levine & Jain, 2023).

Data network effects give a significant advantage not just to companies that have legacy and user data from services other than their generative AI offering, but also to those who have the most users and the state-of-the art technology to process the data. 'Kickstarting' the 'virtuous circle' of data network effects requires a combination of technology, including the model and the hardware, and quality data (Vomberg et al., 2023). As end users engage with generative AI models like GPT, they improve the model, increasing its accuracy and quality.

In this section, we catalogue four major advantages exhibited by several generative AI companies: *(i)* first-mover advantage, *(ii)* computing resources, *(iii)* data collection, and (iv) multimodal system integration. They go hand-in-hand with data network effects in the generative AI industry. We explain these dimensions where appropriate by drawing analogies to pre-generative AI instances of gatekeeping. Without regulatory intervention, these new advantages will likely solidify strong positions within the generative AI services landscape, mirroring patterns observed in other digital markets.

4.1 First-Mover Advantage from Novel Product Features

Some generative AI companies, especially those developing large language models, possess a first-mover advantage. First-mover advantage is defined as a competitive advantage gained by being the first one to introduce a product or service in a market, allowing it to build brand recognition and customer loyalty (Lieberman & Montgomery, 1988). A preliminary distinction must be drawn between being the *first* vs. being the *first-mover*. The first company or team to develop a model with new capabilities is not necessarily the first to capitalize on those capabilities. While a first-mover advantage is not problematic as such, in markets exhibiting strong network effects, it can lead to entrenched and durable positions.[3]

The company exhibiting a clear first-mover advantage in the generative AI industry is OpenAI. Google developed the transformers architecture (Vaswani et al., 2017) that became essential in large language models, yet this did not cement Google's position as a first-mover within generative AI. Although Google developed and open-sourced several large language models that used the previously developed transformers architecture, such as BERT (Devlin et al., 2019), LaMDA (Thoppilan et al., 2022), and PaLM (Chowdhery et al., 2023; Anil et al., 2023), neither model took off as a large-scale product used by the public. There are two reasons for this: first, BERT (the first large language model developed by Google) was designed and developed primarily as a scientific tool to demonstrate the possibility of tokenization and not as a large engineering project (Vipra & Korinek, 2023). Second, neither BERT nor LaMDA had a public, low-friction user interface.[4]

OpenAI created GPT-2 in 2019 (Radford et al., 2019), shortly after the release of BERT, with an easy-to-use interface but limited to a subset of users that requested access via their own website. The stated intent of the limited access policy was to prevent 'malicious' use from spreading in the name of due diligence (OpenAI, 2019), the collateral effect of this policy was to induce scarcity around a new kind of resource: the unrivalled capabilities of large language models. GPT-2 was comparable

[3] See paragraph 79 in the Impact Assessment European Commission (2020a).

[4] While BERT has garnered academic interest, it is often criticized for its suboptimal tokenization methods.

to other state-of-the-art models and did not constitute a leap in performance. However, the limited access policy created a new dynamic in which the research community went from 'open-by-default' to scarcity around research access to large language model capabilities (Vincent, 2019).[5] Because OpenAI put the emphasis on safety concerns emanating from the state-of-the-art capabilities of the model, the limited access policy also created an early standard for understanding large language models' capabilities and supposed risks.

GPT-3 introduced a new user interface, ChatGPT, alongside a significant increase in conversational capabilities. Notably, there is a distinction with Anthropic's Claude, which exhibited similar capabilities to GPT-3 but was tested internally within the company's Slack channels rather than made available to the public. Lastly, GPT-4 extended these capabilities by introducing new features, such as larger output sizes and system integration with Bing's browser services. Additionally, it introduced the ChatGPT Plus service, offering priority access to new features for a monthly subscription fee of $20/month.

Each release of OpenAI's GPT models represents an upgrade in their capabilities and a strategic evolution in OpenAI's place in the market. OpenAI carefully plans each model's debut, introducing new features that are initially scarce before becoming widely available and, ultimately, monetized. This strategy highlights OpenAI's efforts to enhance its market influence with each new model. Specifically, GPT-2 laid the groundwork by creating a demand for research access, effectively setting a new standard for accessing and understanding these models in both academic and commercial settings. Following this, GPT-3 focused on capturing a larger share of computing resources, while GPT-4 advanced into monetization, linking directly to the concept of controlling access as outlined by the Digital Markets Act (DMA). OpenAI is actively pursuing regulatory measures to maintain its competitive edge by safeguarding its technology from open-source distribution and reinforcing its market position.

Moreover, securing a first-mover position often means defining the benchmarks for software performance and ethical standards within the industry. For instance, the deployment and assessments of GPT-4 have established the criteria for evaluating the effectiveness of large language models. Such standards are now the baseline for comparing new methodologies, as seen in the widespread reference to GPT-4 in recent studies (Achiam et al., 2023; Saparov et al., 2024; Gemini Team et al., 2023; Li et al., 2024). These comparisons extend to performance on standardized tests, human evaluations, and also the significant expansion of the user community (Duarte, 2024), illustrating GPT-4's role in setting the industry's performance and value expectations.

On a final note, while understanding Google's impact on the generative AI market may not be straightforward, its impact is no doubt significant. On the one hand, it was the first to develop the state-of-the-art architecture needed for large-scale generative AI. Google has repeatedly shaped the AI field with its research and development contributions to AI, from pioneering the transformers architecture (Vaswani et al., 2017) to developing some of the earliest language models such as BERT. On the other hand, it was not the first *mover* in the sense of deploying its technologies at scale by building a user base and integrating them into other systems. However, Google's impact is significant due to its technological advancements and recently released generative AI products. The recent Gemini model includes a larger base of tokens than state-of-the-art models like GPT4 (Gemini Team et al., 2023) and is integrated into Google Workspaces, making it highly competitive. Google's vast data sources, technological infrastructure, and integration of AI across existing and novel products and services give it unique leverage points. Thanks to these advantages, which we further discuss below, Google maintains a powerful position in the AI ecosystem, molding the development and deployment of generative AI technologies.

4.2 Capital-Intensive Resources Required for Performance Optimization

State-of-the-art models in generative AI require a significant amount of computing power, as measured through the number of floating point operations or FLOPs.[6] Computing power is needed to

[5]As a subset of the coauthors can attest, although elements of BERT were language-dependent (it was initially English-only), its architecture was made open-source and was understood by the research community as comprising an ongoing scientific and technical agenda being carried out by relevant teams at Google. As such, ongoing access was informally maintained by Google and, in effect, guaranteed by having a standing research affiliation with one of its labs. For these reasons, despite being led by a corporate entity, research on language models did not face formal organizational barriers before GPT-2. For more information, see Devlin & Chang (2018) and references therein.

[6]Not to be confused with FLOP/s, which is the measure for computing performance defined as the number of floating point operations per second.

process the training data through a model: on the one hand, the model undergoes updates through its weights and parameters; on the other hand, the data that is used for training is being read, processed, and re-used in many iterations of the model updates. Thus, the larger a model is (i.e. the more parameters it has) and the larger the training data is, the more computing power is needed for successful training. As AI models have undergone significant growth in model size, complexity, and training data size, generative AI is at the forefront of computing power usage.

Computing power is critical for two main tasks in generative AI models: training and inference. Training is used to update the model parameter and weights through several iterations by minimizing an objective function, whereas inference is defined as the task where the model outputs a response based on a query (e.g. a prompt given to ChatGPT outputs a response, which is the inference step). Training uses computing power in terms of both required memory storage and FLOPs. For example, as a recent report from Andreessen-Horowitz argues (Appenzeller et al., 2023), the memory requirement during training for a model like GPT-3 that has 175 billion parameters is over a terabyte. State-of-the-art hardware options, such as NVIDIA's A100s, currently offer 80GB of high bandwidth memory, meaning that multiple units are needed for training models of such size (NVIDIA, 2021). For example, ByteDance is training LLMs with GPU clusters at the scale of more than 10,000 GPUs (Jiang et al., 2024). Similarly, the amount of computing power needed for training (FLOPs available) is directly linked to the training time (e.g. a 175 billion parameter model would take decades to train on a single GPU). In contrast, the computing power needed for inference is significantly less (an inference prompt is tiny, incomparable in size to the training data, and would take less than 1 second on a A100 machine).

State-of-the-art models have seen a massive increase in the number of parameters used (Achiam et al., 2023; Ramesh et al., 2022). Furthermore, the size of the model is not the only variable in determining the cost of development, as all models in generative AI (larger or smaller) use very large datasets for training. More specifically, the larger the training dataset, the more epochs a model needs to undergo in the training phase, thus increasing the number of FLOPs used. Thus, it is the training data size and the model size together that create a huge demand for computing power, and the latest hardware specifications are needed to accommodate these evolving computational requirements.

However, acquiring and using the necessary GPUs for training large models is no easy task for two reasons: (1) increasing scarcity in their availability (e.g. Nvidia's A100s have been in short supply in 2023 (Appenzeller et al., 2023)) and (2) large training costs. For example, estimates state that using cloud providers, the cost of training GPT-3 is of the order of \$4 million (without the added inference costs) (Thompson et al., 2021), while the cost of training GPT-4 is close to \$100 million (Knight, 2023). Of course, the exact costs are not publicly available and may be slightly reduced due to the use of in-house infrastructure or optimization methods. On this last point, even with an in–house infrastructure, training GPT-4 is estimated to have needed around 25,000 A100 GPUs (for reference, the latest 8-GPU system from Nvidia, NVIDIA DGX A100 has an estimated cost of $\$200,000$ in 2023). Smaller generative AI models, such as Stable Diffusion, have a reported cost of $\$600,000$ for training, according to the CEO's public statements.[7] These costs do not include the acquiring, storage, and maintenance costs regarding the training dataset, as well as inference costs, personnel costs, etc. It was the need for financial resources to secure the access to computing power that led OpenAI to change their business model and become a capped for-profit company.[8]

It is doubtful that the availability of small models and the decreasing costs of fine tuning are enough to counteract the increasing power of the lucky few who can afford and access the computing power necessary to train large generative AI models. The amount of compute needed in developing AI models has grown not only for large models, as regular-scale models have been doubling their compute usage in just 5.7 months while large models have been doubling their compute usage in 9.9 months (Sevilla et al., 2022). Fundamentally, this represents a change from the 'Deep Learning Era', as argued by Sevilla et al. (2022): on average, large-scale models use two orders of magnitude more computing power than classical deep learning models (see Figs. 1–3 in Sevilla et al. (2022)).

Meanwhile, there is evidence that smaller models cannot compete with the capabilities of state-of-the-art large generative AI models (Gudibande et al., 2023). Despite the decreasing cost of fine-tuning, achieving state-of-the-art performance still requires a high budget, creating an entry barrier for

[7] https://twitter.com/EMostaque

[8] Sam Altman described this shift in an interview with Lex Friedman, available at https://www.youtube.com/watch?v=L_Guz73e6fw. See also (Brockman et al., 2019).

potential players. Coupled with the scarcity concerns emanating from the global supply chain issues for manufacturing and shipping state-of-the-art GPUs, we are at a point where the sheer amount of computing power needed appears to favour the actors with immense financial and reputational resources in acquiring such GPUs. This entry barrier disproportionately affects smaller companies, public institutions, and universities, who often lack the financial resources to establish independent large-scale generative AI systems (Vipra & Myers West, 2023; Murgia, 2023). Consequently, the concentration of cutting-edge computing power and expertise in the hands of a few players risks limiting player diversity and stifling innovation within the generative AI industry.

4.3 Data Resources

High-quality training and evaluation data is the foundation of any generative AI model. Although many datasets that are used for training generative AI models (and other types of models) are publicly available or sold by third parties for general commercial purposes, a number of generative AI companies benefit from expanded data resources along several dimensions, including private, tailor-made corpus of generative AI inputs, outputs, and metadata that includes personal data.

First, established platform service providers hold massive troves of *data from legacy applications* that can be used to train or evaluate generative AI models. For example, Microsoft owns GitHub, and a current lawsuit alleges that Microsoft used private GitHub repository data to evaluate its Copilot generative AI coding tool (Firm & Butterick, 2023). Additionally, as argued by FTC (2023), established players may benefit from 'honed proprietary data collection tools and technologies for acquiring or scraping data'.

Usage data generated from interactions with AI applications (through prompts and other inputs) is also valuable. The more generative AI models are used and interacted with, the more their quirks or issues are tested, understood, optimized, and adapted. Several generative AI providers are in a position to collect usage data from both individuals and business users. While, for example, OpenAI claims to no longer use enterprise or API customer data to train its models (OpenAI, 2024b), it has in the past, possibly up until March 2023. Nor does this preclude the use of data from individuals. Although individuals in the EU retain the right to opt out of this data processing under the GDPR, the process to opt out is not user-friendly (OpenAI, 2024c,a), and it is unclear how many users do so in practice. Other large firms also use input and output data for model training. For example, Anthropic has an opt-in mechanism to use input data for model training (AnthropicAI, 2024).

Even when large firms do not directly use business user input and output data to train models (e.g. because the user has opted out), they still gain access to a vast trove of information and metadata about *how* business users use generative AI systems. For example, OpenAI reserves the right to run business data through content review systems and classifiers (OpenAI, 2024b). Due to the rapidly evolving nature of generative AI models and the large role played by reinforcement learning with human feedback (RLHF) and red teaming in improving generative AI model performance, this type of metadata may be quite useful in establishing entrenched and durable positions. For example, companies may be in a position to leverage the quality and scale of metadata to address concrete challenges in RLHF fine-tuning Casper et al. (2023), unlocking new technical capabilities unavailable to smaller competitors.

Another side-effect of such rich and complex data collection systems is the threat to end user privacy. A recent data breach by Microsoft researchers exposed 38 terabytes of user data to the public (Ben-Sasson & Greenberg, 2023). Furthermore, OpenAI's data collection practices have led to the ban of ChatGPT in March 2023, as it could not reliably prove that it did not collect personal information from children below 13 years of age (Weatherbed, 2023). Later in the year, a privacy researcher, Lukasz Olejnik, filed a complaint before the Polish data protection authority, alleging that OpenAI failed to produce details on how it collects personal data (Lomas, 2023). Since data is the primary raw ingredient in the recipe to create LLMs, generative AI providers might be incentivised to keep their data sources and collection practices out of the public eye.[9] Privacy is not a direct objective of the DMA, and issues of opacity and security are already addressed in the GDPR and now also the AI Act. However, the DMA clearly emphasises the intersection of data collection practices, privacy

[9] Examples of restrictions on access to data or placing databases behind paywalls already exist See e.g. KeyserSosa (2023) and Nylen & Bass (2023)

and contestability, by explicitly mentioning the obligation of the gatekeeper to 'protect privacy' and encouraging 'an adequate level of transparency of profiling practices employed by gatekeepers'.[10]

4.4 Integrated Systems

A growing trend in generative AI systems is the recent influx of integrated services, especially by companies already designated as gatekeepers under the DMA. A variety of services have started to integrate LLMs, including search engines (Microsoft, 2023; Reid, 2023), personal assistants, note-taking (e.g. Notion), and editing. Platform tendency for integration and the associated loss of user autonomy is not new: Google, Microsoft and Apple have all pushed for integrated systems in their now-established services, such as search and email interfaces. As integrated systems that use generative AI become ubiquitous, their convenience and potential for creative endeavours trade off with user autonomy.

This integration goes both ways between generative AI and other services. Both Google Search and Microsoft's Bing search engine have built functionalities drawing upon generative AI models—indeed, 'GPT-4 is already used in many places on Bing, including Copilot, Image Creator from [Microsoft] Designer, and already in regular web result ranking', according to an official Bing blog post announcing yet another GPT-powered integration, Deep Search (Microsoft, 2023). Microsoft's Services Agreement confirms that as part of its AI services, which would include Bing powered by GPT-4,[11] 'Microsoft will process and store your inputs to the service as well as output from the service, for purposes of monitoring for and preventing abusive or harmful uses or outputs of the service' (Microsoft, 2023a). In a similar vein, Google connected its Bard generative AI model to a suite of services such as Gmail, Google Docs, Google Maps, and YouTube with its Bard Extensions update in September 2023 (Pinsky, 2023). From the other end, ChatGPT gained 'browse with Bing' functionality in September 2023 (Arunasalam, 2023).

This integration trend further muddies the distinctions between different services when it comes to data processing practices. As services and data become increasingly intertwined, it becomes difficult for a user to switch between platforms and transport their data and projects. As emphasised in the DMA (Recital 3), control over whole ecosystems allows gatekeepers to leverage their advantages from one area of activity to the other, which in turn reduces contestability.

5 Proposal for a New Core Platform Service: General Purpose AI Systems

The potential harms of AI have been discussed for years, but some of them are now materialising: the current race toward generative AI enhancement might lead to the emergence of a few large players at the expense of contestability and fairness in the generative AI industry. The DMA and the AI Act do not directly address this problem.

Certain cases of generative AI applications might already be indirectly caught by the DMA when current gatekeepers integrate proprietary or third-party generative AI into their core platform services. Their standalone generative AI offering might also be caught. For example, the DMA specifically prohibits certain instances of personal data and service integration across different core platform service and other services provided by gatekeepers (e.g., Article 5(2)(b) and 5(8)).

However, without the explicit inclusion of generative AI as a form of core platform service, any protection provided by the DMA can only have limited effect. Such indirect application does not address generative AI's specific contestability and fairness concerns. For one, unless generative AI is listed as a separate core platform service, any generative AI behemoths that do not themselves provide any other core platform service cannot be designated as gatekeeper under the DMA. Furthermore, the DMA's gatekeeper obligations will not even indirectly apply to existing gatekeepers' generative AI offering in many instances. For example, gatekeepers will remain free to leverage generative AI business user data where the generative AI service is provided separately from other core platform services.[12] Similarly, the DMA's prohibition on forcing business and end users to subscribe to

[10]Recitals 35 and 72

[11]The agreement defines AI services as 'services that are labeled or described by Microsoft as including, using, powered by, or being an Artificial Intelligence ("AI") system'.

[12]See Article 6(2) of the DMA.

or register with any other services of the gatekeeper applies only across core platform services.[13] Without explicit integration of generative AI as a core platform service, gatekeepers remain free to force users to register with their, for example, social networking services as a condition to use their generative AI services.

While the development of large generative AI models is relatively recent in comparison to other core platform services, the potential for some generative AI providers to acquire an entrenched and durable position is significant. At any rate, Article 3(1) allows the Commission to designate a gatekeeper even where an entrenched and durable position is not yet reached, but is foreseeable in the near future, which ensures that the Commission does not have to wait for economic power to settle in order to intervene under the DMA. In a similar vein, the quantitative thresholds under Article 3(2) provide a presumption, but the Commission may still designate a gatekeeper where the qualitative conditions are met but the quantitative conditions are not (Articles 3(8) and 17), relieving the Commission of the necessity to wait three years under Article 3(2)(c).

For example, in terms of user numbers which are indicative of a service's importance under Article 3(1)(b) and 3(2)(b), ChatGPT is the second fastest digital service to reach 1 million users in just five days (Duarte, 2024) (with an estimated 100 million monthly active users in November 2023), only surpassed by Threads (which inherited Instagram's user base; note that Instagram took 2.5 months to reach 1 million users). Other models, such as Stable Diffusion, acquired a user base of 1 million users in just over a month (Douglas, 2022) Furthermore, other state-of-the art generative AI models have been developed by companies that already benefit from an entrenched and durable position, making their generative AI services inherit a reputational advantage and most importantly, a large user base—Google's Gemini, Meta's LLaMA, Microsoft's Bing AI (which is powered by GPT-4), among many others. While user numbers, especially business user numbers of these services are not publicly available, the available data demonstrates their rapid and continuous growth.

In terms of market capitalization which is indicative of 'significant impact on the internal market', several generative AI providers have already been found by the Commission to have a significant impact, such as Google, Meta, and Microsoft, who have been developing or integrating large generative AI models in their services. Among other providers of generative AI models, valuations have been rapidly increasing over the last couple of years (Liu, 2023; Vipra & Korinek, 2023). In its latest fundraising round in February 2024, OpenAI was valued at USD 80 billion, which is approximately EUR 73.7 billion—a figure nearly reaching the DMA's quantitative threshold at EUR 75 billion. While there is always a risk of inflation in private valuations, there is little doubt that OpenAI is fast becoming one of the most valuable companies in the world.

Furthermore, there is evidence that unfair practices are emerging in generative AI. For example, Microsoft reportedly threatened to cut off two business customers' access to its search engine index data, which it licenses to other search engines, if they did not stop using the data as training input for their own generative AI tools (Nylen & Bass, 2023). Such behavior appears to be intended to weaken the contestability of Microsoft's own generative AI services by denying potential competitors the key input for model training. Generative AI providers remain free to leverage their superior bargaining power to engage in such unfair practices, and their already considerable economic power to prevent contestability.

The DMA provides an opportunity to address contestability and fairness issues in the generative AI industry by designating generative AI as a core platform service. To be clear, including a new core platform service in the DMA does not automatically result in the designation of gatekeepers for that core platform service. A gatekeeper can only be designated if the conditions of Article 3(1) are fulfilled. However, once generative AI becomes a core platform service, the Commission can act swiftly if and once gatekeepers emerge.

Our proposal invites the question of how the relevant generative AI core platform service or services should be defined. There is no need to muddy the regulatory waters: a new core platform service to address generative AI can be defined in reference to the AI Act as 'general-purpose AI systems'. As explained above, this category was added to the draft AI Act with generative AI applications in mind following the advent of ChatGPT. This approach would be in keeping with the DMA's definition of core platform services in reference to other EU legislation, to the extent definitions exists in other directives and regulations, e.g. for 'online search engine' and 'cloud computing service'.

[13] See Article 5(8) of the DMA.

An initial assessment of gatekeeper obligations reveals that some of them would already apply to generative AI systems if they were listed in the DMA as a core platform service. For example, under Article 6(2), gatekeepers are prohibited from using, in competition with their business users, the data generated by business users of their core platform services and by these businesses' customers. This provision would prevent generative AI gatekeepers from free-riding on data produced by businesses relying on their API to provide downstream applications, either in generative AI verticals or other industries. Similarly, a generative AI gatekeeper without any other core platform service would be prevented from combining personal data across its services (Article 5(2)). Article 6(13) could be used to ensure that generative AI gatekeepers do not implement disproportionate terms and conditions for terminating the provision of the generative AI service.

6 Conclusion

The DMA should be amended to add generative AI to its list of core platform services. This is the first necessary step to address the emergence of entrenched positions and unfair practices related to gatekeeping in the generative AI industry. We invite the Commission as well as other scholars and practitioners to further analyse the implications of gatekeeping in generative AI with a view to designing adequate gatekeeper obligations in this context. Furthermore, the facts and data that we presented to make our case were obtained from publicly available resources. They strongly point to the risk of gatekeeping in generative AI, but as we have discussed, opacity as opposed to openness is currently the industry norm. We are confident that the Commission, which has extensive investigative powers, can uncover a great deal more information from generative AI providers demonstrating this risk.

While this sector is advancing rapidly, the technological and institutional trends are now clear. It is better to nip gatekeeping in the bud before we find ourselves once again with economic power positions too strong to challenge in any meaningful way.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Adobe. Adobe unveils future of creative cloud with generative AI as a creative co-pilot in Photoshop. Press release, 2023. URL https://news.adobe.com/news/news-details/2023/Adobe-Unveils-Future-of-Creative-Cloud-with-Generative-AI-as-a-Creative-Co-Pilot-in-Photoshop-default.aspx/default.aspx. May 23, 2023.

Amazon. Announcing new tools to help every business embrace generative AI Amazon Web Services. Press release, 2023. URL https://aws.amazon.com/blogs/machine-learning/announcing-new-tools-to-help-every-business-embrace-generative-ai/. September 28, 2023.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

AnthropicAI. Privacy policy. Press release, 2024. URL https://www.anthropic.com/legal/privacy. March 4, 2024.

Appenzeller, G., Bornstein, M., and Casado, M. Navigating the High Cost of AI compute. Web Page, 2023. URL https://a16z.com/navigating-the-high-cost-of-ai-compute/. April 27, 2023.

Arunasalam, S. ChatGPT users can now browse internet, OpenAI says. Web Page, 2023. URL https://www.reuters.com/technology/openai-says-chatgpt-can-now-browse-internet-2023-09-27/. September 27, 2023.

Ben-Sasson, H. and Greenberg, R. 38TB of data accidentally exposed by Microsoft AI researchers. Web Page, 2023. URL https://www.wiz.io/blog/38-terabytes-of-private-data-accidentally-exposed-by-microsoft-ai-researchers. September 22, 2023.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Brockman, G., Sutskever, I., and OpenAI. OpenAI LP. Press release, 2019. URL https://openai.com/blog/openai-lp. March 11, 2019.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Coldevey, D. App Store for AI: Openai's GPT Store lets you build (and monetize) your own GPT. Web Page, 2023. URL https://techcrunch.com/2023/11/06/app-store-for-ai-build-your-own-gpt-and-sell-it-on-openais-gpt-store/. November 7, 2023.

Council of the EU. Digital Markets Act (DMA): agreement between the Council and the European Parliament. Press release, 2023. URL https://www.consilium.europa.eu/en/press/press-releases/2022/03/25/council-and-european-parliament-reach-agreement-on-the-digital-markets-act/. March 25, 2022.

Devlin, J. and Chang, M.-W. Open sourcing BERT: State-of-the-art pre-training for natural language processing. Web Page, 2018. URL https://blog.research.google/2018/11/open-sourcing-bert-state-of-the-art-pre.html. November 2, 2018.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Douglas, W. Generative AI is changing everything. But what's left when the hype is gone? Technology Review, 2022. URL https://www.technologyreview.com/2022/12/16/1065005/generative-ai-revolution-art/. December 16, 2022.

Duarte, F. Number of ChatGPT Users. Web Page, 2024. URL https://explodingtopics.com/blog/chatgpt-users. March 1, 2024.

European Commission. Impact assessment of the Digital Markets Act. Press release, 2020a. URL https://ec.europa.eu/newsroom/dae/redirection/document/72185. December 15, 2020.

European Commission. Shaping Europe's Digital Future. https://commission.europa.eu/document/download/84c05739-547a-4b86-9564-76e834dc7a49_en?filename=communication-shaping-europes-digital-future-feb2020_en.pdf, 2020b. February, 2020.

European Commission. Proposal for a regulation of the european parliament and of the council on contestable and fair markets in the digital sector (digital markets act). *SEC (2020) 437 final*, 2020c.

European Commission. Final report—sector inquiry into consumer Internet of Things. COM(2022) 19 final, 2022a. URL https://ec.europa.eu/commission/presscorner/detail/en/ip_22_402. January 20, 2022.

European Commission. On contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 Digital Markets Act. OJ L 265, 2022b. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R1925. December 10, 2022.

European Commission. Questions and Answers: Digital Markets Act: Ensuring fair and open digital markets. Press release, 2023. URL https://ec.europa.eu/commission/presscorner/api/files/document/print/en/qanda_20_2349/QANDA_20_2349_EN.pdf. September 6, 2023.

European Parliament. AI Act: a step closer to the first rules on artificial intelligence. Press release, 2023. URL https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence. May 11, 2023.

European Union. Regulation (EU) 2019/1150 of the European Parliament and of the Council on promoting fairness and transparency for business users of online intermediation services. OJ L 186, 2019. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?url=CELEX:32019R1150. June 20, 2019.

Firm, J. S. L. and Butterick, M. Github Copilot litigation, 2023. URL https://githubcopilotlitigation.com/case-updates.html. January 25, 2024.

FTC. Generative AI raises competition concerns, 2023. URL https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns. June 29, 2023.

Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Google. Vertex AI Model Garden and Generative AI Studio. Press release, 2023. URL https://cloud.google.com/blog/products/ai-machine-learning/vertex-ai-model-garden-and-generative-ai-studio. March 29, 2023.

Grady, P. and Curnin, C. Partnering with Harvey: Putting LLMs to Work. Web Page, 2023. URL https://www.sequoiacap.com/article/partnering-with-harvey-putting-llms-to-work/. April 26, 2023.

Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., and Song, D. The false promise of imitating proprietary LLMs. *arXiv preprint arXiv:2305.15717*, 2023.

IBM. IBM unveils the Watsonx platform to power Next-generation foundation models for business. Press release, 2023. URL https://newsroom.ibm.com/2023-05-09-IBM-Unveils-the-Watsonx-Platform-to-Power-Next-Generation-Foundation-Models-for-Business. May 9,2023.

Insights, M. T. R. The great acceleration: CIO perspectives on Generative AI. *MIT & Databricks*, 2023.

Jasper. Jasper vs. ChatGPT: Which is right for you? Web Page, 2023. URL https://www.jasper.ai/comparison/jasper-vs-chatgpt.

Jiang, Z., Lin, H., Zhong, Y., Huang, Q., Chen, Y., Zhang, Z., Peng, Y., Li, X., Xie, C., Nong, S., et al. MegaScale: Scaling large language model training to more than 10,000 GPUs. *arXiv preprint arXiv:2402.15627*, 2024.

Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., and Lahmann, A. The GenAI is out of the bottle: generative Artificial Intelligence from a business model innovation perspective. *Review of Managerial Science*, pp. 1–32, 2023.

KeyserSosa. An update regarding Reddit's API. Press release, 2023. URL https://www.reddit.com/r/reddit/comments/12qwagm/an_update_regarding_reddits_api/. November 16, 2023.

Knight, W. OpenAI's CEO says the age of giant AI models is already over. Web Page, 2023. URL https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/. April 17, 2023.

Levine, S. S. and Jain, D. How network effects make AI smarter. HBR, 2023. URL https://blog.google/products/search/generative-ai-search/. March 14, 2023.

Li, H., Su, J., Chen, Y., Li, Q., and ZHANG, Z.-X. Sheetcopilot: Bringing software productivity to the next level through large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Lieberman, M. B. and Montgomery, D. B. First-mover advantages. *Strategic management journal*, 9 (S1):41–58, 1988.

Liu, E. A $90 billion valuation for OpenAI? Tech's new star is hotter than Nvidia. Web Page, 2023. URL https://www.barrons.com/articles/openai-valuation-ai-stock-microsoft-nvidia-a696ebf8. September 28, 2023.

Lomas, N. ChatGPT-maker OpenAI accused of string of data protection breaches in GDPR complaint filed by privacy researcher. Web Page, 2023. URL https://techcrunch.com/2023/08/30/chatgpt-maker-openai-accused-of-string-of-data-protection-breaches-in-gdpr-complaint-filed-by-privacy-researcher/. August 30, 2023.

Macaulay, T. ChatGPT has generated a brand new app sector—and European devs are leading it. Web Page, 2023. URL https://thenextweb.com/news/chatgpt-generated-new-app-sector-for-european-devs-generative-ai. May 2, 2023.

Microsoft. Announcing foundation models in Azure Machine Learning. Web Page, 2023. URL https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/announcing-foundation-models-in-azure-machine-learning/ba-p/3827481. May 23, 2023.

Microsoft. Introducing deep search. Web Page, 2023. URL https://blogs.bing.com/search-quality-insights/december-2023/Introducing-Deep-Search. December 5, 2023.

Microsoft. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. Press release, 2023. URL https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/. May 16, 2023.

Microsoft. Microsoft services agreement. Press release, 2023a. URL https://www.microsoft.com/en-us/servicesagreement. July 30, 2023.

Microsoft. Microsoft, Anthropic, Google, and OpenAI launch frontier model forum. Web Page, 2023b. URL https://blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum/. July 26, 2023.

Murgia, M. Big Tech companies use cloud computing arms to pursue alliances with AI groups. Web Page, 2023. URL https://www.ft.com/content/5b17d011-8e0b-4ba1-bdca-4fbfdba10363. February 5, 2023.

NVIDIA. NVIDIA A100 Tensor Core GPU. Press release, 2021. URL https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf. June, 2021.

Nylen, L. and Bass, D. Microsoft threatens data restrictions in rival AI search. Web Page, 2023. URL https://www.bloomberg.com/news/articles/2023-03-25/microsoft-threatens-to-restrict-bing-data-from-rival-ai-search-tools. March 25, 2023.

OpenAI. Better language models and their implications, 2019. URL https://openai.com/research/better-language-models.

OpenAI. OpenAI API. Press release, 2020. URL https://openai.com/blog/openai-api. June 11, 2020.

OpenAI. ChatGPT plugins. Press release, 2023a. URL https://openai.com/blog/chatgpt-plugins. March 23, 2023.

OpenAI. Introducing GPTs. Press release, 2023b. URL https://openai.com/blog/introducing-gpts. November 6, 2023.

OpenAI. How ChatGPT and our language models are developed. Press release, 2024a. URL https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed. January 10, 2024.

OpenAI. Enterprise privacy at OpenAI. Press release, 2024b. URL https://openai.com/enterprise-privacy. January 10, 2024.

OpenAI. Europe privacy policy. Press release, 2024c. URL https://openai.com/policies/eu-privacy-policy. December 15, 2023.

Pinsky, Y. BARD can now connect to your google apps and services. Web Page, 2023. URL https://blog.google/products/gemini/google-bard-new-features-update-sept-2023/. September 19, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Reid, E. Supercharging search with generative AI. Press release, 2023. URL https://blog.google/products/search/generative-ai-search/. May 10, 2023.

Saparov, A., Pang, R. Y., Padmakumar, V., Joshi, N., Kazemi, M., Kim, N., and He, H. Testing the general deductive reasoning capacity of large language models using ood examples. *Advances in Neural Information Processing Systems*, 36, 2024.

Scale AI. Enterprise generative AI platform. Web Page. URL https://scale.com/generative-ai-platform.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2022.

Siemens. Siemens healthineers shows potential of generative AI in medical technology. 2023. URL https://www.siemens-healthineers.com/press/releases/generativeai. November 26, 2023.

Soni, V. Impact of generative AI on small and medium enterprises' revenue growth: The moderating role of human, technological, and market factors. *Reviews of Contemporary Business Analytics*, 6 (1):133–153, 2023.

The Economist. Large, creative AI models will transform lives and labour markets. Web Page, 2023. URL https://www.economist.com/interactive/science-and-technology/2023/04/22/large-creative-ai-models-will-transform-how-we-live-and-work. April 22, 2023.

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F. Deep learning's diminishing returns: The cost of improvement is becoming unsustainable. *IEEE Spectrum*, 58(10):50–55, 2021.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

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

Vincent, J. AI researchers debate the ethics of sharing potentially harmful programs. Verge, 2019. URL https://www.theverge.com/2019/2/21/18234500/ai-ethics-debate-researchers-harmful-programs-openai. February 21, 2019.

Vipra, J. and Korinek, A. Market concentration implications of foundation models: The invisible hand of ChatGPT. 2023. URL https://www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt/. September 7, 2023.

Vipra, J. and Myers West, S. Computational power and AI. *AI Now Institute*, 2023.

Vomberg, A., Schauerte, N., Krakowski, S., Bogusz, C. I., Gijsenberg, M. J., and Bleier, A. The cold-start problem in nascent AI strategy: Kickstarting data network effects. *Journal of Business Research*, 168:114236, 2023.

Weatherbed, J. OpenAI's regulatory woes have barely started. Web Page, 2023. URL https://www.theverge.com/2023/5/5/23709833/openai-chatgpt-gdpr-ai-regulation-europe-eu-italy. May 5, 2023.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Weizenbaum, J. Computer power and human reason: From judgment to calculation. *WH Freeman & Co*, 1976.

Wiggers, K. OpenAI's Foundry will let customers buy dedicated compute to run its AI models. Web Page, 2023. URL https://techcrunch.com/2023/02/21/openai-foundry-will-let-customers-buy-dedicated-capacity-to-run-its-ai-models/. February 22, 2023.

Wiggers, K., Coldevey, D., and Singh, M. Anthropic's $5b, 4-year plan to take on OpenAI. Web Page, 2023. URL https://techcrunch.com/2023/04/06/anthropics-5b-4-year-plan-to-take-on-openai/. April 6, 2023.