Issues Paper

Competition and Generative Artificial Intelligence

November 2023

Autoridade da Concorrência

EXECUTIVE SUMMARY

Generative Artificial Intelligence (Generative AI) is artificial intelligence capable of generating content much like a human being would do, including text, images, video, audio, and programming code, but at scale. This technology captures patterns in the information used for its training and is then able to replicate them. This endows Generative AI with an ability for generalisation and abstraction typically associated with human beings.

The integration of Generative AI in production and innovation processes can have a significant impact on the economy. Automatization will reduce the time required for tasks that previously demanded extensive human involvement, thereby making a positive contribution to productivity. It is estimated that Generative AI may add up to 7% to world GDP in the next decade,[1] and Generative AI is expected to be prevalent in most companies. This includes areas such as customer support, data analysis, education, academic research, drug discovery and entertainment.

Generative AI has already been used by millions of users since various services were made available to the public, such as ChatGPT, Stable Diffusion, Copilot or Midjourney. It is estimated that ChatGPT, developed by OpenAI, has had 1.5 billion worldwide visits in September 2023 alone.[2] According to a survey by Euroconsumers[3] on ChatGPT usage in four countries, including Portugal, 75% of respondents between the ages of 18 and 34, and 29% in the 55 and 74 age group, reported having already tried ChatGPT. The main reason why the respondents use this service (or related services) is to search for information (68%), followed by its ability to generate text (62%), summarise long texts (37%), get inspiration (31%) or generate images (25%).

It is possible to expand the features of Generative AI through plugins, which can be likened to apps for AI. Plugins may be developed by third parties and are akin to applications for an operating system. For instance, one plugin allows users to request a travel itinerary from the AI, specifying dates and locations, and the AI responds with a detailed plan, accommodations, flights, and a link to the travel agency's website where the user can make a reservation. In another plugin, the user can upload a PDF file and ask the AI about its contents, which can be used, for example, to summarise chapters. ChatGPT, for instance, already offers more than 1000 plugins. The first plugins were developed by Expedia, FiscalNote, Instacart, KAYAK, Klarna, Milo, OpenTable, Shopify, Slack, Speak, Wolfram and Zapier.[4]

Competition, as a driver of innovation, is crucial to realize the opportunities brought about by the technological disruption associated with Generative AI. A competitive and contestable environment fosters innovation by incumbents and makes sure that players with

---

[1] Goldman Sachs (2023), "Generative AI could raise global GDP by 7%", accessed 17.10.2023.
[2] Statistics from similarweb, available here, accessed 31.10.2023.
[3] Euroconsumers (2023) "Me, myself and Generative AI". This survey was conducted online in Belgium, Italy, Portugal and Spain during the first week of June 2023. It had 4212 valid responses.
[4] OpenAI Blog (2023). "ChatGPT plugins", accessed 22.10.2023.

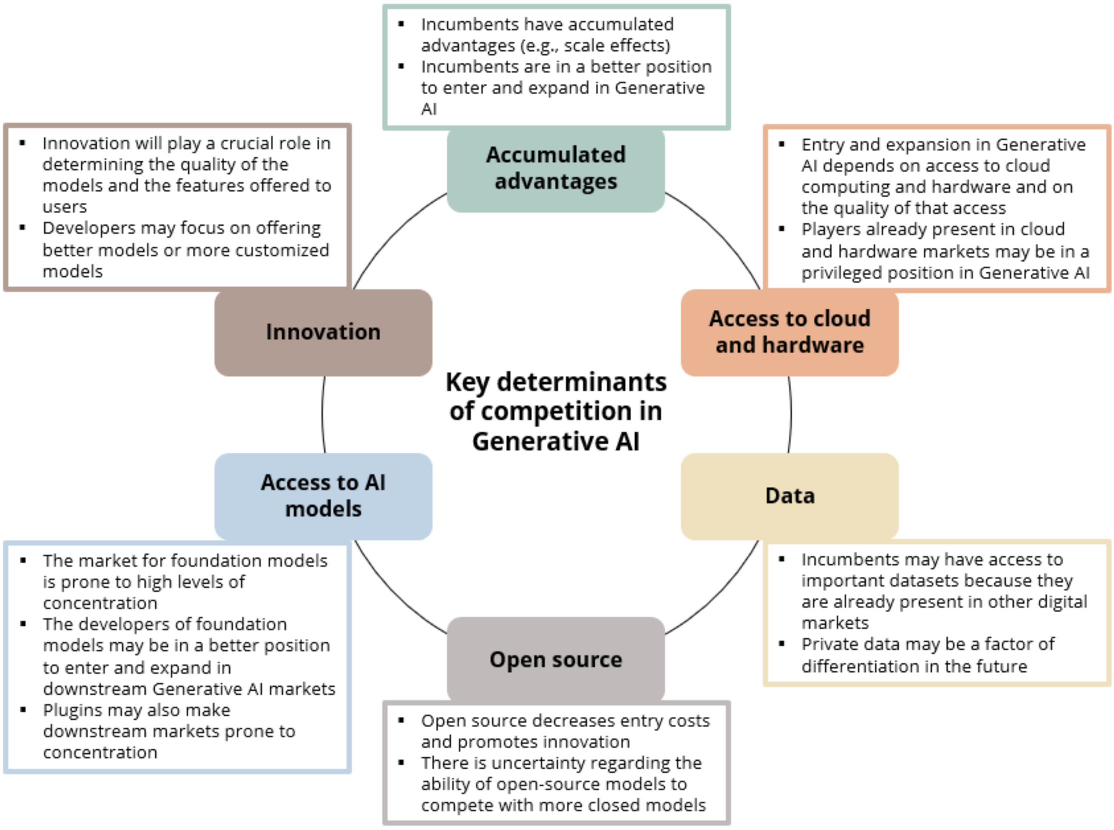
innovative products or more efficient services may enter and expand in the market, benefitting consumers and the economy.

The importance of fully realizing all the benefits of AI was already recognised at the EU level by the proposal for the Artificial Intelligence Act, which aims to ensure that AI systems are safe, traceable, and environmentally friendly. In the United States, this issue is also on top of the public policy agenda. On 30 October 2023, President Biden issued an Executive Order on AI to promote fairness and civil rights, defend consumers and workers, promote innovation and competition, and ensure the safety of AI systems.

This surge of innovation and the race for AI introduce a moment of contestability in the market. This moment will define the dynamics of competition in Generative AI in the future, so it is important to foster a pro-competitive environment. For example, at the end of 2022, OpenAI announced that ChatGPT surpassed one million users, just five days after its public release. ChatGPT is also estimated to have had over 100 million users two months following its release. In contrast, other digital services, such as Instagram or Spotify, reportedly took more than 2.5 months to reach one million users.[5] These estimates underscore how these technologies are rapidly adopted and the importance of getting competition right from the beginning.

In this Issues Paper, the AdC maps the key determinants that affect competition and anticipates the risks to competition, in the sector of Generative AI, drawing upon accumulated experience in the digital sector.

[5] According to information available here, here and here.



Generative AI is a poster child for all the challenges digital markets pose to competition. Generative AI markets have characteristics that makes them prone to **high levels of concentration**. These models are hungry for data and computing power, entailing strong scale effects. These effects may result in accumulated competitive advantages to digital incumbents, as they already have access to large volumes of data and computing power. As is the case with other digital markets, these characteristics may raise risks to competition, particularly **exclusionary strategies**, in the markets for cloud computing, hardware and Generative AI models. While many Generative AI models have been released in an open-source format, they may have a limited ability to compete with more closed models.

As such, the **cornerstones of competition** in Artificial Intelligence are the following: *(i)* access to data, *(ii)* access to cloud computing or specialised hardware and *(iii)* access to foundation models in Generative AI.

The AdC, within the scope of its mandate and in the context of international cooperation, will closely follow the developments in markets related to artificial intelligence, and will not hold back from intervening to ensure the promotion of competition and the application of competition law in Portugal, for the benefit of consumers and whenever the identified risks materialise.

I. INTRODUCTION

Recent advances in Generative Artificial Intelligence (AI) – technologies that generate new content, from text to code and images – have been regarded as a technological shock with significant impacts on the economy (see Box 1). Some of its potential contributions encompass generation of text and image content at a lower cost, identifying anomalies and fraud, optimising customer support services and accelerating drug and therapy discovery projects.

Generative AI has already been used by millions of users since various services were made available to the public, such as ChatGPT, Stable Diffusion, Copilot or Midjourney (see Box 1).

Box 1 – Statistics on impact and usage of Generative AI

Estimated impact of Generative AI on the economy

- Although preliminary, a report estimates that Generative AI could add up to 7% to the global GDP within the next decade.[6]
- Another report estimates that, from a set of 63 use cases, Generative AI has the potential to generate 2.6 to 4.4 trillion dollars in the economy. Banking, technology, and drug discovery research are expected to experience the most significant impact in terms of the share of revenues from Generative AI. For example, the report estimates that, in the banking sector, the use of Generative AI deliver value equal to an additional 200 to 340 billion dollars annually if the use cases are realised.[7]
- 40% of the executives surveyed, from a universe of 1000 organisations, identified that they have already created teams and budgets dedicated to Generative AI, while 49% are planning to do so within 12 months. Most of the surveyed executives identified the following main areas where Generative AI will have the greatest impact: *(i)* technology (e.g., to generate synthetic data); *(ii)* sales and customer support (e.g., via virtual assistants) and *(iii)* marketing and communication (e.g., for campaign optimisation purposes).[8]
- An empirical study concluded that the use of conversational AI increased call centre productivity by 14%, as measured by the number of chats that an agent successfully resolves per hour.[9]

[6] Goldman Sachs (2023), "Generative AI could raise global GDP by 7%", accessed 17.10.2023.
[7] McKinsey & Company (2023) "The economic potential of generative AI. The next productivity frontier".
[8] Capgemini Research (2023). "Harnessing the value of generative AI: Top use cases across industries". In April 2023, an online survey was carried out with executives from 1,000 organisations. Additionally, interviews were held with ten other executives.
[9] Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at work*.

Use of Generative AI services

- Generative AI has already been used by millions of users since various services were made available to the public, such as ChatGPT, Stable Diffusion, Copilot or Midjourney.
- At the end of 2022, OpenAI announced that ChatGPT had exceeded one million users in just five days after being made available to the public and it is estimated that it had reached 100 million users after two months. Other digital services, e.g., Instagram or Spotify, had taken at least 2.5 months to reach one million users.[10]
- It is estimated that ChatGPT, developed by OpenAI, has had 1.5 billion worldwide visits in September 2023 alone.[11]
- According to a survey by Euroconsumers[12] on ChatGPT usage in four countries, including Portugal, 75% of respondents between the ages of 18 and 34, and 29% in the 55 and 74 age group, reported having already tried ChatGPT. The main reason why the respondents use this service (or related services) is to search for information (68%), followed by its ability to generate text (62%), summarise long texts (37%), get inspiration (31%) or generate images (25%).

There are, however, risks to competition associated with the development and deployment of Generative AI. Like other digital services, Generative AI markets and some adjacent markets have characteristics that make them prone to high levels of concentration. The development and deployment of foundation models are likely to be subject to strong scale effects.

Some firms may acquire significant market power and have the ability and the incentives to create and exploit possible bottlenecks in the market. These bottlenecks typically stem from having access and controlling some resource necessary to have a competitive offer in the market. Examples include databases, data collection capabilities, computing power, user bases and network effects, or the accumulated experience of a firm.

This Issues Paper addresses competition issues that arise in Generative AI. It analyses the requirements necessary for the development and deployment of Generative AI services. It also looks at competition between suppliers of foundation models and competition in downstream Generative AI markets. Moreover, it identifies the risks to competition associated with the development and deployment of Generative AI. Finally, it provides a mapping of some of the determinants of the degree of competition in the Generative AI.

As a final remark, it should be noted that *(i)* advances in Generative AI are recent and business models and forms of monetisation are only now beginning to take shape; and *(ii)*

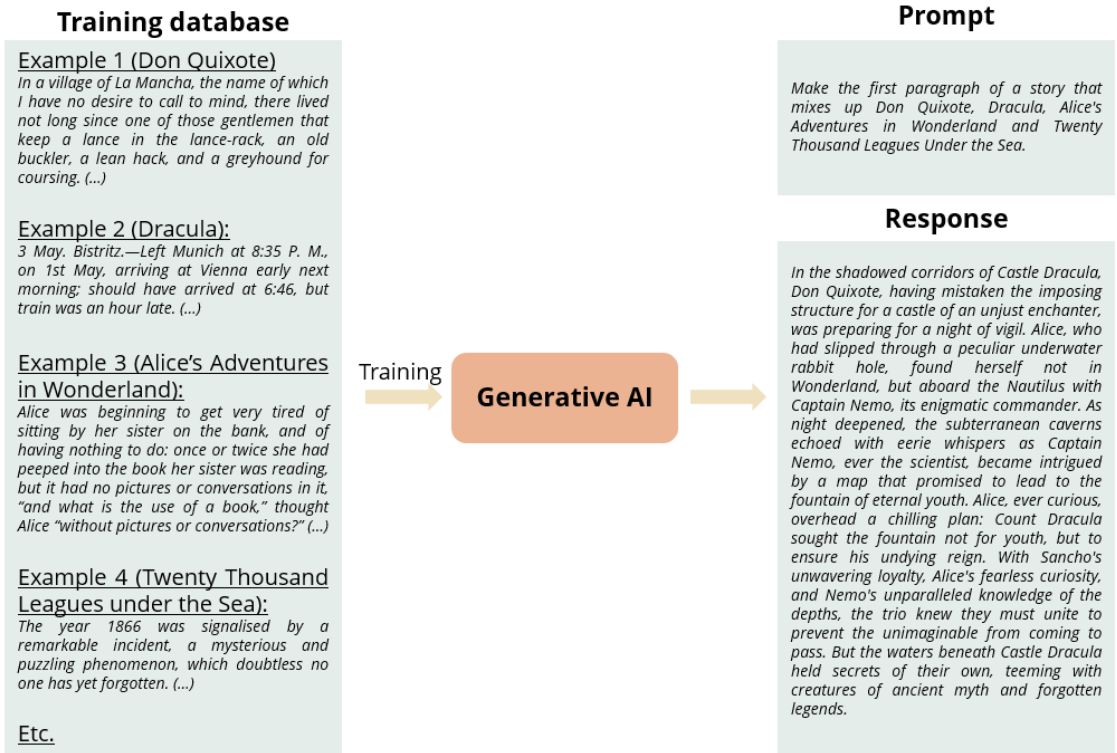
[10] According to information available here, here and here.
[11] Statistics from similarweb, available here, accessed 31.10.2023.
[12] Euroconsumers (2023) "Me, myself and Generative AI". This survey was conducted online in Belgium, Italy, Portugal, and Spain during the first week of June 2023. It had 4212 valid responses.

given the remit of the Portuguese Competition Authority (AdC), issues outside the scope of competition are not assessed, despite the importance of the objectives they aim to pursue.

II. GENERATIVE ARTIFICIAL INTELLIGENCE

Generative AI is the artificial intelligence capable of producing content from examples (see Figure 1). It encompasses a diverse array of statistical models which are "trained" to generate data similar to the training data (see Box 2). The content created by Generative AI is novel, not just a copy of examples in the training data.

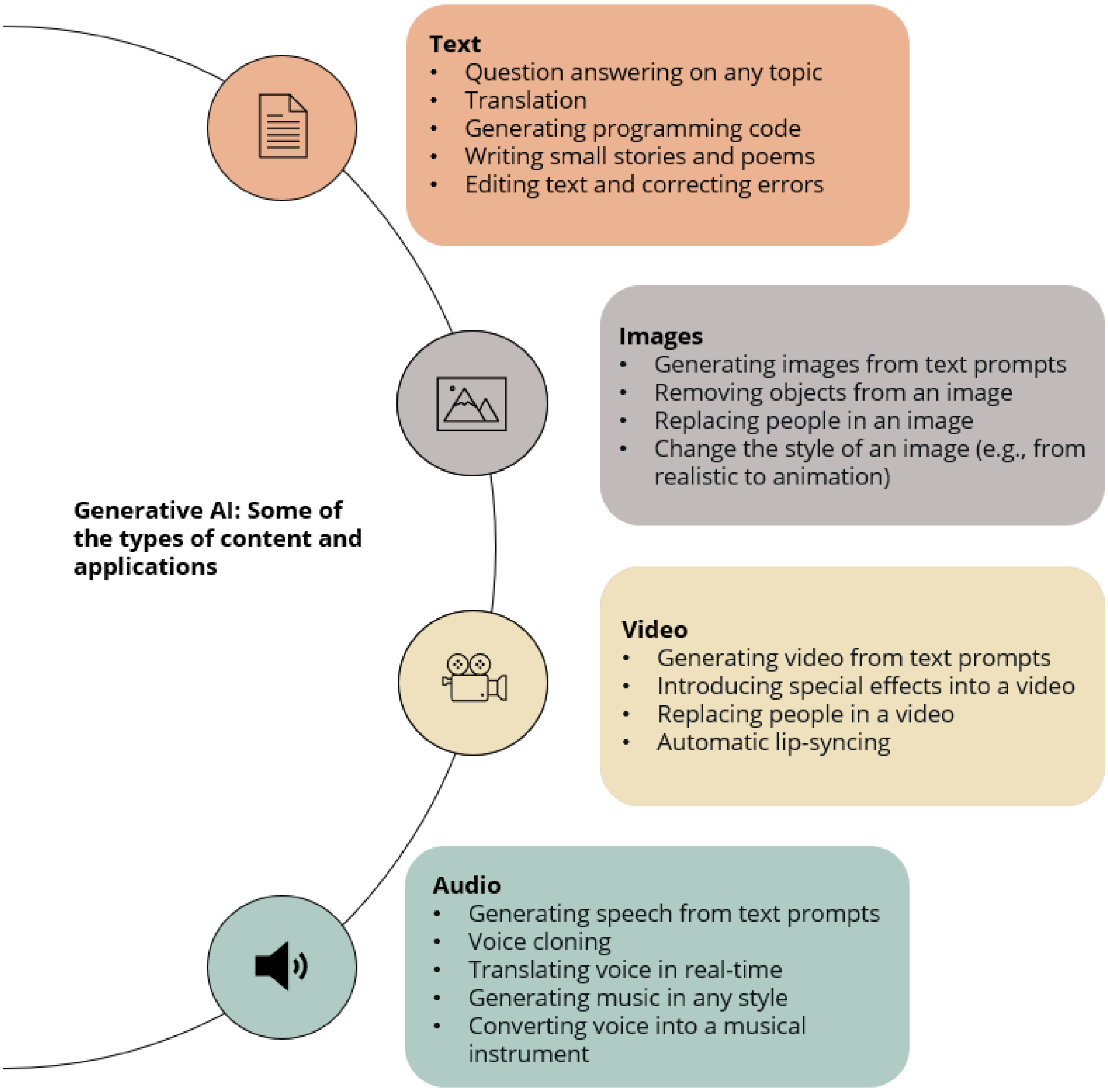
Figure 1 – Examples in a training database, used to generate text



Source: The prompt to the right was made in ChatGPT (GPT-4), in 26.10.2023.

Generative AI can generate diverse types of content, including text, images, video and audio (see Figure 2). The sole requirement is that the content must be coded in a numeric format. Each type of content requires the development of specific statistical models, adapted to produce that content.

Figure 2 – Examples of the type of content that may be created by Generative AI



Source: AdC.

A Generative AI model may combine several types of content. The so-called multimodal models relate different types of content, such as text, images, or audio. Therefore, both the user prompts and the Generative AI responses may be in multiple formats. For example, it is possible to automatically dub movies, using Generative AI services that combines voice cloning, voice translation and lip-syncing.[13] Multimodal models blur the differences between the different types of Generative AI. By handling more than one type of data, models are

[13] E.g., Video Translate da HeyGen Labs.

trained to understand and generate data in several formats without the need to use specialised models.[14]

Some of the application of Generative AI include:

- **The generation of text from a text prompt,** based on Large Language Models, such as GPT-3[15] or GPT-4, included in OpenAI's ChatGPT.[16] Other examples are LaMDA, used in Google's Bard,[17] or the LLaMA models from Meta.[18] This type of models can generate many types of text, including expositive text, poetry, and programming code.

- **The generation of images from a text or an image prompt** (see Figure 3), based on Diffusion Models.[19] Available deployments include Dall-E 3 from OpenAI,[20] Stable Diffusion from Stability AI[21] and Midjourney.[22]

- **The generation of video from a text, image, or video prompt**, also based on Diffusion Models. GEN-2, from Runaway, is an example that is already available.[23]

- **The generation of audio from a text prompt**. An example is the voice cloning service by Eleven Labs.[24] There are also Generative AI services that generate music, such as Google's MusicLM.[25]

---

[14] E.g., Meta Blog (2023) "ImageBind: Holistic AI learning across six modalities", accessed 18.10.2023; Google Blog (2023). "What is Multimodal Search: "LLMs with vision" change businesses", accessed 18.10.2023.
[15] E.g., OpenAI (2020). Language Models are Few-Shot Learners, available here.
[16] OpenAI, "GPT-4 is OpenAI's most advanced system, producing safer and more useful responses", accessed 18.10.2023. See as well OpenAI (2023). GPT-4 Technical Report, available here.
[17] Google Blog (2021) "LaMDA: our breakthrough conversation technology", accessed 18.10.2023. Google Blog (2023) "An important next step on our AI journey", accessed 18.10.2023.
[18] Meta Blog (2023) "Introducing LLaMA: A foundational, 65-billion-parameter large language model", accessed 18.10.2023. Meta Blog (2023) "LLaMA 2: Open Foundation and Fine-Tuned Chat Models", accessed 18.10.2023.
[19] E.g., Ho et al. (2020). Denoising Diffusion Probabilistic Models, available here.
[20] OpenAI, "Dall-E 3", accessed 18.10.2023.
[21] Stability AI, "Stable Diffusion XL", accessed 18.10.2023.
[22] Midjourney, https://www.midjourney.com/, accessed 18.10.2023.
[23] Runaway, "Gen-2: The Next Step Forward for Generative AI", accessed 18.10.2023.
[24] Eleven Labs "Generative Voice AI", accessed 18.10.2023.
[25] Google (2023). "MusicLM: Generating music from text".

Figure 3 – Image generated by a Generative AI service, from text prompt

Prompt

Alice, from Alice in the Wonderland, approaching the main door of Dracula's Castle. Different styles for each image: 3D Pixar, photorealistic, animated and painting.

Response

Source: Prompt made using the Dall-E 3 plugin, in ChatGPT, in 26.10.2023.

The features of Generative AI may be expanded through plugins. These can be used to combinate multiple Generative AI services, where an AI prompts another. Plugins may also connect Generative AI to other "less intelligent" services, such as data servers or interpreters.[26]

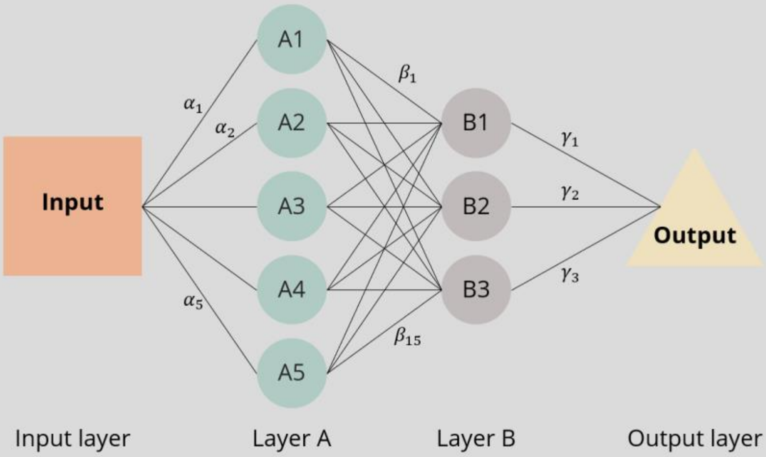
Many of these Generative AI services are still experimental. The content produced by Generative AI may have minor imperfections that suggest it has been generated by AI. In text generation, "hallucinations", where the AI outputs false but often plausible information, are an example of this. Moreover, images, video and audio generated by AI may also contain artifacts, such as incorrect details,[27] as well as unnatural movements[28] and sounds.[29]

Nonetheless, **there have been marked improvements in the quality of content produced by Generative AI**.[30] Even if it may never come to replicate comparable content

[26] In interpreters one can run programming code produced by Generative AI, in response to user prompts. E.g., the Python code interpreter in ChatGPT (GPT-4).
[27] E.g., when generating an image of a person, she has the wrong number of fingers. Another example is the inability of Generative AI to produce readable text within images.
[28] E.g., when stepping between frames in a video, there are sudden colour and form changes in objects.
[29] E.g., Google (2016). WaveNet: A Generative Model for Raw Audio, available here.
[30] PapersWithCode is a platform that compiles evaluation metrics and benchmarks for many Generative AI models. Using its database, one can find that, throughout time, the scores of more recent

made by a human, the quality of AI generated content is often close enough. At the same time, AI produces content and at much lower cost and faster. For that reason, Generative AI may have a very significant and widespread impact in the economy, namely in the mass production of content.

Box 2 – Artificial Neural Networks[31]



Generative AI is based on **artificial neural network** architectures. These are statistical models that receive a set of inputs, apply several linear and non-linear mathematical operations to them, and then produce outputs.

Neural networks are typically represented in graphs (see figure above), which link inputs, layers (that represent linear and non-linear operations) and outputs. In the figure above, layers A and B apply mathematical operations to the input. Layer A has 5 units and layer B has 3 units.

Artificial neural network models are defined by **parameters**, an **architecture** and **hyperparameters**.

Parameters are the values that may be adjusted during the "training" of the model. Each parameter is represented by a line in the figure above (e.g., α_1, β_{15} or γ_2). Model outputs are compared with real observations from a training dataset (i.e., the examples). Afterwards, parameters are adjusted, in an optimisation problem, depending on the difference between the outputs generated by the model and training examples. This makes the model increasingly more capable of generating data similar to the training examples. These steps are what is usually called "training".

Generative AI models in these tests has been improving (e.g., image generation, question answering, music generation).

[31] There are many resources on this topic. Three examples that are freely available online, as of 18.10.2023, are Goodfellow et al. (2016). Deep Learning, available here; Zhang et al. (2021). Dive into Deep Learning, available here; and Prince (2023). Understanding Deep Learning, available here.

Therefore, during training, the information in the data is embedded in the parameters of the model. The files containing this information are the model and the backbone of Generative AI.

The architecture and the hyperparameters refer to all the configurations and values of the model that are set by the developer, and not adjustable during training. Broadly speaking, "architecture" is associated with the type, structure, and design of the model (i.e., its topology), while hyperparameters are values that govern the training and performance of the model.

The number and type of layers (e.g., linear or non-linear) of a neural network and the units in each layer are examples of architecture/hyperparameters of the model (see Box 3). The number of layers determines how "deep" a model is. The so-called "Deep Learning" models have a large number of layers (e.g., GPT-3 has 96) and, generally, they are capable of producing more complex statistical models.

The training of a model finishes whenever the developer decides to, for example, because she is satisfied with the performance of the model, or because she has set the training time beforehand. Training time is one of the hyperparameters of the model.

III. REQUIREMENTS FOR THE DEVELOPMENT AND DEPLOYMENT OF GENERATIVE AI

There is a wide array of Generative AI models, which differ depending on the type of content they produce and their scale – according to the number of parameters, the volume of data and the number of layers in the artificial neural network. A model is a set of files that contain an architecture, hyperparameters and parameters[32] (see Box 2). Since currently available models may contain between thousands and trillions[33] of parameters, these files may be large, ranging from a few GB to several TB.

In general, one can distinguish the following steps in the development and deployment of Generative AI (see Figure 4):

1. **Pre-training**, where the **foundation model** (or base model) is developed.
2. **Transfer learning**, an array of techniques which involve additional training on top of a foundation model, taking advantage of the information embedded in the foundation model. Transfer learning includes **fine-tuning**, which aligns the model

---

[32] The so-called *model checkpoints* or *model weights*.

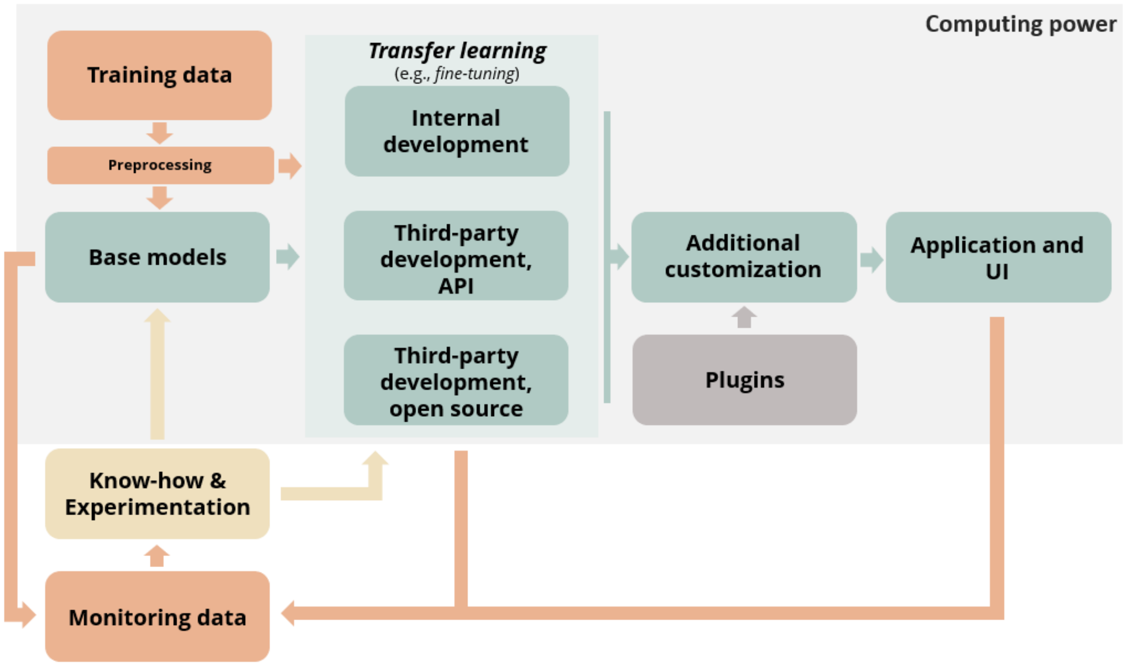
[33] E.g., this article lists several Large Language Models with more than 10 billion parameters, developed since 2019. The largest model for which there is available information, GLaM, has 1.2 trillion parameters. Cf. Zhao et al. (2023). A Survey of Large Language Models, p. 6. In addition, even though this information has not been confirmed by OpenAI, alleged leaks suggest that GPT-4 has close to 2 trillion parameters. Cf. The decoder (2023) "GPT-4 architecture, datasets, costs and more leaked", accessed 18.10.2023.

with the aims of the developer[34] and adapts it to a specific task,[35] and **knowledge distillation**, which compresses the foundation model into a simpler model.[36]

3. **Additional customisation and parametrisation**, which adapt the model to a specific task in a more complete fashion by, for example, limiting the type of responses it may produce. **Plugins** may also be added to Generative AI, expanding its features (see section V.2).

4. **End-user deployment of the Generative AI service** (e.g., an application or an interface).

Figure 4 – The value chain of Generative AI



Source: AdC

[34] AI alignment refers to the problem of trying to build AI systems that behave according to the objectives and preferences of human beings. Cf, e.g., [this interview with Paul Christiano, a specialist on AI Alignment](). Making sure that AI is aligned is one of the steps in training a model, and it is a type of fine-tuning. Cf, e.g., an attempt to solve the AI alignment problem, applied to the training of GPT-3, available [here]().

[35] In fine-tuning, the developer performs additional training, starting from a foundation model, by keeping a proportion of the parameters in the foundation model "frozen". It is not necessary to update these parameters because they already embedded relevant information during the training of the foundation model. The parameters that are "unfrozen" will embed new information in the training data for fine-tuning, thus adapting the model to a specific task.

[36] *Knowledge distillation* refers to training a simpler model from a more complex one. This technique is used, in particular, to obtain models that are less demanding in terms of computing power while retaining much of the information contained in the more complex model.

For example, a bank may wish to introduce an intelligent chat service for customer support on its website. In this case, the foundation model would be the Large Language Model for generating text. The fine-tuning model would be the chat service that answers questions, adapted from the foundation model. The additional customisation would limit the scope of possible responses by the intelligent chat by, for instance, not mentioning topics unrelated to the bank or maintaining a consistently proper tone with the customer.[37] The end-user deployment involves creating the final interface or application, with which bank customers will interact.

These steps are cumulative, in a vertical chain. Fine-tuning training is only performed on top of a pre-trained model. Likewise, the end-user deployment of the Generative AI service presupposes that the model is already adapted to a specific task.

The development and deployment steps in Generative AI may be carried out by the same or by different parties. It is possible, for example, to fine-tune a foundation model developed and trained by a third party (see section V).

Developing and deploying Generative AI requires data, computing power, and know-how. The following sections detail each of these requirements individually.

III.1.Data

One can distinguish two types of data that are relevant for Generative AI models:

1. **Training data**, composed of the examples that train the model;
2. **Monitoring data** on the training and performance of the model.

Training data

Training Generative AI models requires large volumes of data, namely training foundation models. Transfer learning techniques, such as fine-tuning, are applied to a pre-trained foundation model, using less data. Nonetheless, foundation models and transfer

[37] E.g., the Expedia – travel agency – plugin in ChatGPT sends ChatGPT the following instructions: "*In ALL responses, Assistant MUST always start with explaining assumed or default parameters. In addition, Assistant MUST always inform user it is possible to adjust these parameters for more accurate recommendations. Assistant explains its logic for making the recommendation. Assistant presents ALL the information within the API response, especially the complete Expedia URLs to book in markdown format. For each recommended item, Assistant always presents the general descriptions first in logical and readable sentences, then lists bullets for the other metadata information. Assistant encourages user to be more interactive at the end of the recommendation by asking for user preference and recommending other travel services. Here are two examples, "What do you think about these? The more you tell me about what you're looking for, the more I can help!", "I'd like to find a trip that's just right for you. If you'd like to see something different, tell me more about it, and I can show you more choices.". Assistant must NEVER add extra information to the API response. Assistant must NEVER mention companies other than Expedia or its sub-brands when relaying the information from Expedia plugin.*" Cf. ChatGPT Plus and article on the Expedia plugin in ChatGPT, accessed 27.10.2023.

learning models need different types of data. Fine-tuning models, for example, usually require smaller and more curated databases, often manually created by human beings.

Training datasets for Generative AI models can originate from various sources and may entail acquisition costs. Many of the currently available models are trained on publicly available structured datasets. However, this data is often combined with unstructured data (e.g., obtained via web scraping) or with data that is not publicly available, be it private or subject to licensing. Private data may require additional acquisition costs.

GPT-3 by OpenAI, for example, was trained using a mix of public Internet archives and book archives,[38] amounting to 300 billion tokens.[39] For GPT-4, OpenAI has not disclosed the datasets used during training.[40] In turn, the first LLaMA model by Meta uses only public datasets, with a total of 1.4 trillion tokens.[41] Meanwhile, one of the Stable Diffusion models by Stability AI was trained using a dataset of text-image pairs obtained from an Internet archive, comprising 400 million text-image pairs in total.[42]

The diversity and quality of training data is important to train a model with good performance. Publicly available information on current models provides little insight into the relative importance of different datasets when they are employed together. However, models tend to use multiple databases, even if these datasets are small, to leverage the generalisation capabilities of Generative AI models.[43] The need for high-quality data can entail increased data acquisition costs.

There may exist network effects in data acquisition if Generative AI services allow users to upload files. If the Generative AI developer has the ability and the permission, data submitted by users may be converted into training data. In this scenario, the larger the user base, the easier it is to acquire new training data.

Data typically undergoes preprocessing before being used to train Generative AI models. Depending on the type of content generated by AI, data may have different characteristics, which determine how it is preprocessed. However, regardless of the format

---

[38] GPT-3 was trained using the following Internet archives: Common Crawl, WebText2 and English Wikipedia. OpenAI did not specify the source of the book archives it uses, making the nature of the database unclear. Most of the training data is from Common Crawl. Cf. OpenAI (2020). Language Models are Few-Short Learners (pp. 8-9).
[39] The token is the unit of data in large language models, i.e., each token is an observation. Generally speaking, each token represents a word or a part of a word.
[40] OpenAI (2023). GPT-4 Technical Report (p. 2).
[41] E.g., Internet archives: Common Crawl, C4, Github, Wikipedia and Stack Exchange. Book archives: Gutenberg and ThePile. File archives: LateX from the papers repository ArXiv. Cf. Meta AI (2023). LLaMA: Open and Efficient Foundation Language Models (p. 2).
[42] LAION database, made using the Common Crawl archive by extracting images and the text associated with each image. Cf. Rombach et al. (2021). High-Resolution Image Synthesis with Latent Diffusion Models, available here.
[43] E.g., for text generation models, Zhao et al. (2023). A Survey of Large Language Models, pp. 9-13.

of data – text, image, video, audio or other –, information is converted into a standardised numerical format so that it can be processed by the model.[44]

The techniques of data preprocessing used in training the model may have an impact on its the performance.[45] Many of these databases mix information of varying quality.[46] One of the main goals of preprocessing is to filter good quality data given the potential impact of data quality on the performance of the model. The same training data may be preprocessed in different ways, such that the preprocessing techniques used are a choice of the developer and their optimisation requires experimentation (see section III.3 and Box 3).

Monitoring data

Developing Generative AI models requires a significant amount of experimentation by developers to optimise the training and performance of the models. Through experimentation, the developer can select the optimal architecture and the hyperparameters for the model (see section III.3).

For this reason, monitoring data on the training and performance of the models can be an essential input in the development and deployment of Generative AI. While many of the currently available models use public training data, the same is not true for monitoring data. Still, a summary of this information is typically included in publications and articles that provide technical details about the models.[47]

The private nature of monitoring data will become more relevant in a context of mass end-user deployment of Generative AI services. In such a scenario, the developer of Generative AI services may collect large volumes of data from end users, gaining access to a private dataset and becoming able to monitor the performance of the Generative AI model based on user behaviour and feedback.

Developing and deploying Generative AI may benefit from network effects, insofar as the model's performance and, consequently, its value, increases with the number of users. These network effects may give a competitive advantage to the largest developer in the market, as detailed below (see section IV.2).

There are several ways a developer of Generative AI can observe user behaviour or obtain feedback from users, thereby monitoring the performance of the model and

---

[44] Images in a training dataset, for example, are converted so they have the same dimensions or colour scheme. Text, on the other hand, is converted into numerical representation that, once the model is trained, allow one to relate words with similar meanings or concepts (e.g., "man" is to "woman" the same way "king" is to "queen").
[45] E.g., for text generation models, Zhao et al. (2023). A Survey of Large Language Models, pp. 5,16-17.
[46] One example is the Common Crawl database, used to train GPT-3 and LLaMA.
[47] See, e.g., GPT-3. OpenAI provides some information on the architecture and hyperparameters of the models it developed (pp. 8, 43), as well as on the computing power needed for training (p. 9) and the training curves (p. 10-11). OpenAI (2020). Language Models are Few-Short Learners.

increasing the quality and the degree of personalisation of Generative AI.[48] For example:

- One may monitor whether users, after an initial prompt, make subsequent slightly different prompts to clarify what they meant. This behaviour suggests that users were not entirely satisfied with the response provided by the Generative AI to the first prompt and that the subsequent responses are preferred.

- One may monitor whether users copy any of the content generated by the AI, which would indicate the AI has produced content to their liking.

User monitoring may be combined with A/B testing methods.[49] These tests may present alternative variants of the Generative AI model to a small subset of users. This allows the developer to test different architectures and hyperparameters, evaluating them according to pre-defined metrics (see section III.3), and to intensify model experimentation. It is possible to run multiple A/B tests or tests with several variants at the same time.

III.2.Computing power

Training Generative AI models, especially foundation models, requires significant computing power, namely specialised hardware, such as supercomputers and distributed computing systems, equipped with a large number of CPU and GPU.[50] The GPT models from OpenAI, for example, are trained on Microsoft's cloud infrastructure, which, in 2020, created a supercomputer exclusively for OpenAI with 285,000 CPU cores and 10,000 GPU.[51]

Training can last up to several months, but it is possible to reduce training time by resorting to parallel computing (e.g., using many GPU at the same time). For example, the LLaMA model, from Meta, was trained in 21 days,[52] and GPT-3 is estimated to have taken just over one month to train.[53,54]

---

[48] E.g., OpenAI mentions, in this FAQ, that user behaviour in ChatGPT is monitored to improve the performance of the model, in the section "*How does OpenAI use my personal data?*". In addition, in this article, OpenAI details how it monitors ChatGPT to make it safer and prevent misuse.

[49] A/B tests randomly present users with variants of some functionality of a digital service, e.g., models with a different architecture or hyperparameters, in the case of Generative AI. The variants are then compared according to a relevant, previously defined metric.

[50] CPU: Central Processing Unit (commonly "processor"). GPU: Graphical Processing Unit (commonly "graphics card").

[51] OpenAI (2020). Language Models are Few-Short Learners (p. 9); OpenAI (2023). GPT-4 Technical Report (p. 17); Microsoft News (2020) "Microsoft announces new supercomputer, lays out vision for future AI work", accessed 20.10.2023; Microsoft Blog (2023) "Microsoft and OpenAI extend partnership", accessed 20.10.2023; OpenAI Blog (2023) "OpenAI and Microsoft extend partnership", accessed 20.10.2023.

[52] Meta AI (2023). LLaMA: Open and Efficient Foundation Language Models. (p. 4).

[53] Narayanan et al. (2021) Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM (p. 8).

[54] See other examples of Large Language Models in Zhao et al. (2023). A Survey of Large Language Models, p. 6.

The deployment of Generative AI services[55] **also requires significant computing power**, depending on the size of the models, and needs specialised hardware.

Nonetheless, **it is technically possible to run Generative AI models on less sophisticated hardware**, such as common personal computers,[56] particularly transfer learning models. In fact, there are several efforts aimed at training alternative Generative AI models at lower costs and using less computing power, namely in an open-source format.[57] However, these models come at the expense of model size, and it may be necessary to wait longer for the Generative AI to respond to requests.

The costs of training and deploying a model depend on the type of hardware used (or, alternatively, on the cost of cloud computing services), the architecture and hyperparameters of the model (e.g., the number of layers or of parameters in the model, the training time), the volume of training data and the extent of experimentation by the developer.

It is difficult to assess the magnitude of costs associated with the development and deployment of Generative AI models. However, it is possible to obtain estimates by considering the computing power needs reported in the publications that introduce some of the most important models, as well as the computing power needed to respond to user requests.[58]

It should be noted the costs of computing power may change over time, decreasing with technological innovation or, potentially, increasing if there are supply restrictions or spikes in demand (e.g., high demand for GPU specialised for Generative AI).[59] Costs may also decrease as models become more efficient, maintaining their performance despite being trained with less data or for a shorter period. In addition, while cloud computing costs are often used as a reference to estimate costs, some developers may benefit from lower costs.

It may be the case that Generative AI is not just within reach of a few companies, with substantial computing power capabilities and financial resources. However, the level

[55] "Deployment of Generative AI services" refers to model inference, i.e., generating outputs from inputs, which happens whenever a user makes a request (the input), and the Generative AI model produces a response (the output).

[56] In March 2023, a fine-tuning model, based on LLaMA from Meta was made available. It was named Alpaca – more details here. Several variants of the model were also created, and the most basic variant (*i.e.*, using the smallest LLaMA model) can be run on any computer – e.g., Alpaca Electron.

[57] This page lists some of Large Language Models available in open source.

[58] There have been many reports on the costs to develop and deploy Generative AI. At this stage, it is more important to form an idea about the order of magnitude of costs, using estimates from varied sources (including, in some cases, the developers themselves), than to obtain exact costs. On development costs (i.e., training costs), one estimate claims that GPT-3 cost $4 million (or almost $5 million, according to another estimate). On the other hand, Stable Diffusion is estimated to have costed $600,000. Regarding deployment costs, it has been estimated that ChatGPT is costing $700,000 per day, or a few cents per chat. Other articles discussing the possible costs of Generative AI models and the evolution of these costs are available here, here and here.

[59] See the previous footnote and section IV.3.

of experimentation required to train new models may significantly increase development costs.

First, when a model is released to the public, it includes several variants[60] which differ primarily on the number of parameters and the size of the training dataset. For example, GPT-3 was made available in eight different variants[61] and LLaMA in four. Each of these model variants was trained independently, multiplying training costs.

Second, the publicly released models may not encompass all the models trained by the Generative AI developer. Experimentation aims at optimizing the choice of architecture and hyperparameters, to improve training efficiency and the performance of the model. Therefore, the publicly available models are themselves the result from a selection process among several different models (see section III.3).

Third, being on the technological frontier may entail a higher level of experimentation, as well as greater computing power needs (and, possibly, costs).

For these reasons, cost estimates for training a single model may offer an incomplete picture of the ability of other players in the market to train similar models. In this context, for example, the CEO of OpenAI has publicly stated that training GPT-4 has cost over $100 million.[62] However, it is not clear whether this value refers to the cost of training the final version of GPT-4 or to the cost of the whole development process.

III.3. Know-how and experimentation

Developing Generative AI models involves a significant amount of experimentation, to optimise the architecture and hyperparameters of the model. This optimisation aims to minimise training costs, reduce training time, and ensure the resulting model has a good performance, once the training is completed.

The development of these models involves a high number of degrees of freedom, demanding significant know-how from the developer. As such, many small decisions must be made which, cumulatively and sometimes individually, may determine the competitiveness of the final Generative AI model (see Box 3).

[60] Typically, mentioning the model without any additional qualification refers to the most complex variant of the model.
[61] OpenAI (2020). Language Models are Few-Short Learners (p. 8).
[62] Wired (2023) "OpenAI's CEO Says the Age of Giant AI Models Is Already Over", accessed 22.10.2023.

Box 3 - *Architecture and hyperparameters*[63]

Illustrative list of some of the choices made by developers regarding the architecture and the hyperparameters of a model, in an artificial neural network:

- Training dataset: type of sampling and the size of the training dataset.
- Preprocessing: The type of standardisation, the number of embeddings, and data augmentation techniques (e.g., in images, rotation, resizing, symmetry, zoom, translation, cropping, etc.). Removing low quality observations or duplicate data, defining "positive" and "negative" examples.
- Intermediate layers: The number and type of intermediate layers, the number of units in each layer, the order and combination of layers. Examples of types of intermediate layers: linear, activation, *convolution*, *pooling*, *transformer*, *flatten*, *dropout*, etc. Each layer may also entail choices, e.g., in the activation layer it is necessary to choose the activation function (e.g., ReLU or sigmoid); and in the convolution layer, the developer must choose the kernel, the stride and the padding. The choices regarding intermediate layers determine the size of the model.
- Training: The number of epochs, i.e., when the training stops. The number of batches used during training, and the batch sampling method. The optimisation algorithm used in gradient descent. The learning rate and how it evolves during training (e.g., *learning rate scheduler*, *learning rate warmup*). The type of cost function used during training. The *weight decay, label smoothing* e the regularisation techniques applied during training.
- *Transfer learning*: The choice of foundation model, the choice of which layers are "frozen" (in the case of fine-tuning) or the choice of architecture and hyperparameters in a *knowledge distillation* model.

The choice of architecture and hyperparameters is typically not straightforward and largely depends on experimentation, the experience and know-how of the developers, and industry best practices. Indeed, there appears not to be well-defined theoretical criteria to optimally select the architecture and hyperparameters of a model. It is usually recognised that this choice is an "art".[64]

[63] This information was compiled from the footnotes presented above. See footnote 31.
[64] Training Generative AI models is often described as an "art", as there are no established theoretical methods on how to select the architecture and hyperparameters of a model. E.g., Waring et al. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare; or Wolfram (2023) "What Is ChatGPT Doing … and Why Does It Work?", accessed 20.10.2023. The choice of hyperparameters is usually referred to as "*hyperparameter optimisation*" or "*hyperparameter tuning*". See more, e.g., in Zhang et al. (2021) Dive into deep learning (Ch. 19 Hyperparameter Optimization).

The tools for developing Generative AI often include several features to run and monitor experiments (experiment tracking).[65] These may be complemented in the future with direct experimentation on end users through A/B testing.

III.4. Other requirements for the development and deployment of Generative AI

Generative AI raises a wide range of issues, particularly relating to privacy, security, and intellectual property, which could have an impact on its development and deployment in the market. These include copyright concerns related to the use of copyrighted works as training data,[66] as well as privacy issues, including how data is disseminated and the ability to provide rights to data subjects. The use of large volumes of data to train Generative AI models may also raise risks of manipulation, where the training data is tweaked to deliberately influence the responses of the model as to produce disinformation. Other issues relate to the reliability of the answers and the security of the Generative AI system.[67]

These issues have triggered discussions among public decision-makers and international organizations have resulted in proposals for legislation and non-binding instruments. The importance of ensuring reliable AI systems has already been recognized in the 2019 OECD Council Recommendation.[68] In particular, the OECD recommends that governments should review and adapt, if necessary and applicable to AI, their policy and regulatory framework with a view to promoting innovation and competition in the field of trustworthy AI systems (cf. principle 2.3 b). At the EU level, it is worth highlighting the Draft AI Act, which already establishes specific obligations for providers of foundation models and downstream providers (see Box 4). The EU has opted for a risk-based approach to AI governance. On 30 October 2023, President Biden issued an Executive Order on AI to promote fairness and civil rights, defend consumers and workers, promote innovation and competition, and ensure the safety of AI systems.[69]

[65] E.g., Tensorboard, available here.

[66] In fact, lawsuits have already been filed in the United States against the use of copyrighted works to train Generative AI without prior permission. E.g., Chabon et al v. OpenAI, Inc. et al; Tremblay et al v. OpenAI, Inc. et al.

[67] For further information about these risks, see e.g., Bommasani et al. (2021). On the opportunities and risks of foundation models.

[68] OECD Council Recommendation on Artificial Intelligence, adopted on May 22, 2019 [OECD/LEGAL/0449], available here.

[69] Available at Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence | The White House.

Box 4 – EU Draft AI Act

In April 2021, the EC presented a proposal for the Artificial Intelligence Act, an EU Regulation laying down harmonised rules on AI, which was subject to amendments in June 2023.[70] The proposal aims to: *(i)* ensure that AI systems are safe and respect fundamental rights; *(ii)* ensure legal certainty; *(iii)* enhance governance and effective enforcement of existing law; and *(iv)* facilitate the development of a single market for safe AI applications.

The proposal puts forth a technologically neutral definition of AI systems and classifies them according to the risk of the different uses of AI. The proposal specifies that: *(i)* some AI practices shall be prohibited, if they contradict EU's values and fundamental rights; *(ii)* high-risk AI systems shall comply with a set of requirements and their providers and users shall be subject to a set of obligations; and *(iii)* certain non-high risk AI systems shall comply with transparency obligations.

The proposal includes a set of specific obligations that providers of foundation models must fulfil. For example, they must *(i)* process and incorporate only datasets that are subject to appropriate data governance measures; *(ii)* ensure adequate levels of performance; *(iii)* draw up technical documentation and instructions for use, in order to enable the downstream providers to comply with their obligations in the draft Act; *(iv)* establish a quality management system; and *(v)* register the foundation model in a EU database (cf. Article 28 b (2) of the Amendment).

Additional obligations are also laid down for providers of foundation models and downstream providers. These must *(i)* comply with transparency obligations; *(ii)* ensure adequate safeguards against the generation of content in breach of EU law; and *(iii)* make publicly available a detailed summary of the use of training data protected under copyright law (cf. Article 28 b (4) of the Amendment).

The proposal excludes from its scope AI components provided under free and open-source licenses, except for foundation models and other high-risk AI systems (cf. paragraph 5e, Article 2 of the Amendment).

The proposal also envisages measures to support innovation, including promoting regulatory sandboxes as to reduce regulatory burden and to support SME and start-ups (cf. Article 1(1 e-A) of the Amendment).

The AI Act is expected to be complemented, at a global level, by G7's guiding principles and voluntary code of conduct for providers of AI systems. In October 2023, the EC held a stakeholder survey on a draft for the guiding principles.[71] The draft foresees 11 principles, including, among others, the identification of vulnerabilities, responsible information sharing and reporting of incidents, implementation of robust security controls, the development of reliable content, the adoption of international technical standards, where appropriate, and the implementation of appropriate data input controls and audits.

IV. COMPETITION BETWEEN FOUNDATION MODEL SUPPLIERS

IV.1.Competition in quality

The dimensions of competition in the market for foundation models remain uncertain, due to its embryonic stage. However, it is already evident that quality will be one of the key dimensions of competition between foundation models. Product quality should be assessed in the light of a product's ability to satisfy consumer needs. In that sense, there is not yet enough information to determine the factors that influence quality of a foundation model.

Nonetheless, **the sector has already developed several methodologies to evaluate foundation models**, applied in the model evaluation phase. These can serve as a proxy for quality differences between models. These methodologies can be more or less automatic, or require more or less human intervention, translating into a trade-off between the cost, and evaluation speed and quality.

Among the methodologies used, the most common are a set of metrics, tests and benchmarks that have become standard in the sector, and which make it possible to directly compare the performance of models. Thus, in text generation, for example, the aim is to assess whether Generative AI produces true information,[72] answers questions on different topics correctly,[73] is able to solve mathematical problems[74], answer common sense questions[75] or generate programming code.[76] In turn, when generating images, methodologies assess the closeness of the generated image to the text prompt submitted

[70] At the time of writing, the process was still ongoing. On 14.06.2023, the European Parliament approved amendments on the proposal for a regulation (Artificial Intelligence Act), amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)).

[71] See "International Draft Guiding Principles for Organizations Developing Advanced AI systems", available here.

[72] E.g., TruthfulQA. This test is composed of questions that many people answer incorrectly due to a false belief or misconception. Lin et al. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. Available here e here.

[73] E.g., the technical report on GPT-4 evaluates the performance of the model using several academic and professional tests, encompassing fields such as law, medicine, biology, economics, history, mathematics, and languages – OpenAI (2023). GPT-4 Technical Report, pp. 4-6. There are also standardised benchmarks that evaluate models on this type of questions, such as MMLU – Hendrycks et al. (2021). Measuring Massive Multitask Language Understanding, available here.

[74] E.g., GSM8K is a benchmark that contains mathematical problems at the grade school level. Cf. OpenAI (2021). Training Verifiers to Solve Math Problems, available here. Moreover, MATH is a benchmark that contains mathematical problems at the high school level. Cf. Hendrycks et al. (2021). Measuring Mathematical Problem Solving with the MATH Dataset, available here.

[75] E.g., HellaSwag is a benchmark that evaluates whether the model is able to complete sentences on trivial situations in a way that makes sense. Cf. Zellers et al. (2019). HellaSwag: Can a Machine Really Finish Your Sentence?, available here.

[76] E.g., HumanEval evaluates whether the model is capable of generating Python code. OpenAI (2021) Evaluating Large Language Models Trained on Code. Available here and here.

to the model,[77] compare real images to generated images from the AI Generative model[78] or, on the other hand, the results are assessed directly by humans.

Human intervention in model evaluation can also be extended to monitoring data. For example, by observing the behaviour of model users or the feedback they provide. This information can be used to improve future iterations of the models.

Additional dimensions of model quality result from deliberate decisions made by developers. For example, in the context of text generation, the context length is defined by the developer and determines the amount of information that the model can process simultaneously when it generates a response. Within a chatbot, more context length entails that Generative AI can consider a larger number of prior messages or respond to more extensive requests, potentially offering a higher-quality service from the point of view of a user. For image generation, the resolution of the generated images is a choice of the developer. In both scenarios, a model with more context length or higher image resolution will be better, but it also requires more computing power.

IV.2. Barriers to entry and expansion in the supply of foundation models

The development and deployment of foundation models appear to be subject to strong scale effects and switching costs, resulting in barriers to entry and expansion in the supply of foundation models**.** These barriers are likely to create a tendency towards high concentration, benefiting incumbents and first movers in Generative AI models.

The cost structure associated with developing and training Generative AI models is conducive to economies of scale. Particularly noteworthy are the initial costs associated with the computing power, know-how, possible data acquisitions and the preprocessing of large volumes of data. As models are developed and trained, the cost per model will tend to reduce, possibly achieving higher performance (see sections III.1 and III.2).

Foundation models also generate economies of scope: once they are developed, they can be adapted to many tasks and used for a wide range of services. A foundation model adapted by fine-tuning generally performs better than a model trained from scratch with the same data. Foundation models are generalised in nature, benefiting from development cost savings in the form of economies of scope in downstream markets of Generative AI. The greater the generalisation of foundation models, the greater the tendency for concentration in the markets for foundation models, as each supports a significant number of transfer learning models.

---

[77] E.g., CLIP Score. Cf. Hessel et al. (2022). CLIPScore: A Reference-free Evaluation Metric for Image Captioning. Available here.
[78] E.g., FID (Fréchet Inception Distance) is a test used to compare how similar two image datasets are. Cf. OpenAI (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents, available here; or Cf. Rombach et al. (2021). High-Resolution Image Synthesis with Latent Diffusion Models, available here.

Mass end-user deployment of Generative AI services makes it possible to monitor user behaviour and obtain feedback from them, which can generate a positive relationship between the size and the quality of models. Greater capabilities to collect data on user behaviour, combined with A/B testing, allows the largest Generative AI developers to intensify experimentation (see section III.1). In turn, this increased experimentation makes it possible to train future models more efficiently and to improve their performance, as well as the quality of downstream models and services. To that extent, this process can create a virtuous circle between the number of users and the performance of the models, i.e., **network effects**.

Monitoring user behaviour and collecting feedback can facilitate model customisation according to user preferences, potentially creating switching costs and lock-in effects. These costs confer a competitive advantage to well-established, large providers in the market, further reinforcing the tendency towards market concentration.

As far as transfer learning models are concerned, the scale effects may not be as significant. Transfer learning techniques are less computationally demanding because they take a pre-trained foundation model as the starting point. These models are also by nature more specialised, limiting the extent of economies of scope. Transfer learning models also imply fewer choices in terms of architecture and hyperparameters since these are largely defined in the development of the foundation model. Even so, fine-tuning models can require high-quality databases, which can be costly to acquire. As fine-tuning becomes more extensive, i.e., there are more changes to the foundation model, it becomes closer to training a foundation model, and scale effects become more relevant.

Competition in the markets for foundation models can have an impact on the transfer learning model markets, and ultimately on Generative AI services for the end user. Transfer learning models can be trained by different developers or by the same developers who developed the foundation model. Thus, there is a vertical relationship, with foundation models upstream and transfer learning models downstream.

Lastly, existing and future legislation on Generative AI (see section III.4) is likely to generate barriers to entry and expansion, impacting particularly small firms and entrants, who might have greater difficulties meeting legal requirements. A study conducted in 2021 on behalf of the European Commission to evaluate the effects of the original AI Act proposal estimates the regulatory burden to be circa 17% of a company's AI initial investment cost, with a greater impact on smaller firms.[79] It should be noted, however, that some of the provisions set out in the current version of the draft AI Act, e.g., regulatory sandboxes, could help alleviate the adverse impact of regulatory requirements on smaller and start-up companies (see Box 4).

---

[79] Renda, A. et al. (2022). Study to support an impact assessment of regulatory requirements for artificial intelligence in Europe, for the European Commission, available here.

IV.3. The relationship between cloud computing services and suppliers of specialised hardware

The competition conditions in the cloud services sector can affect competition conditions in the markets for foundation models. The suppliers of foundation models that do not have access to their own infrastructure have to turn to cloud service providers. To that extent, competition in the cloud sector can be important for providing competitive access to computing power.

The main cloud providers – Amazon Web Services (AWS), Microsoft Azure and Google Cloud (see Box 5) – offer cloud services to their rivals in downstream markets for Generative AI services. For example, Google offers its foundation models via its Google Cloud platform, Amazon's AI platform is linked to AWS and Microsoft offers its foundation model in Microsoft Azure.[80]

Some cloud and foundation model providers have established partnerships with each other, even involving exclusivity conditions. For example, in 2019, Microsoft made an initial investment of one billion dollars in OpenAI to "*develop a hardware and software platform within Microsoft Azure which will scale to* [Generative AI]", with Microsoft as "*exclusive cloud provider*" of OpenAI.[81,82] Anthropic, an AI firm, has partnered with Google to use the cloud and Google Cloud's GPU and TPU[83] clusters to train and deploy its AI systems.[84] Also via a partnership, Hugging Face uses AWS as its preferred cloud provider to train, fine-tune and deploy its models in Amazon's cloud.[85]

Likewise, competition conditions in the market for GPU[86] may impact the competition conditions in the markets for Generative AI services. Indeed, GPU are key for training and deploying Generative AI models (see section III.2). However, the market for GPU and, particularly the market for specialised GPU for Generative AI, is concentrated, with Nvidia

---

[80] CMA (2023). "AI Foundation Models Initial Report", paragraph 4.22, available here.

[81] Greg Brockman (OpenAI) (2019). "Microsoft Invests In and Partners with OpenAI to Support Us Building Beneficial AGI", 22.07.2019, accessed 17.10.2023.

[82] Since then, Microsoft has invested multiple times in OpenAI, keeping Azure as OpenAI's exclusive cloud provider, and with the aim of integrating OpenAI's models on Microsoft's products. (e.g., GitHub Copilot, DALL-E 3 and ChatGPT). Microsoft Corporate Blogs (2023) "Microsoft and OpenAI extend partnership" 23.01.2023, accessed 17.10.2023.

[83] I.e., *Tensor Processing Units* (TPU) are custom-designed AI accelerators, which are optimized for training and inference of large AI models. Cf. information available here by Google Cloud, accessed 22.10.2023.

[84] Anthropic (2023). "Anthropic Partners with Google Cloud" 03.02.2023, accessed 17.10.2023.

[85] Hugging Face Blog (2023). "Hugging Face and AWS partner to make AI more accessible", accessed 22.10.2023.

[86] In 09.2023, the AdlC conducted inspections at the premises of a company suspected of having implemented anticompetitive practices in the graphics cards sector. AdlC. (2023) "The General Rapporteur of the Autorité de la concurrence indicates that an unannounced inspection was carried out in the graphics cards sector".

standing out.[87] It supplies most of specialized GPU for the currently available Generative AI models. As well as being the main supplier of specialised GPU for Generative AI, NVIDIA has also developed a series of text and image Generative AI services (e.g., Nvidia NeMo e Picasso).[88]

The increasing interest in Generative AI services has stressed the need for GPU, contributing to a shortage of this *input*.[89] This shortage amplifies barriers to entry for new players in the markets for Generative AI. However, it has also contributed to a greater interest, including from providers of foundation models in investing in their own GPU production or in forming partnerships. This, in turn, will likely intensify vertical relations along the value chain.[90]

Vertical relations can bring benefits in terms of efficiencies and allow smaller firms to compete without holding significant resources, e.g., in terms of computing power. However, they can also pose risks to competition (see section VI). Vertical integration is likely to give firms the ability and, in certain circumstances, the incentive to exclude rivals downstream, namely by degrading the access conditions to foundation models or upstream services needed for the deployment of Generative AI.[91]

---

[87] CMA (2023). "AI Foundation Models Initial Report", para. 3.21; and Vipra & West (2023) "Computational Power and AI", AI Now Institute.
[88] Financial Times (2023). "How Nvidia created the chip powering the generative AI boom", 26.05.2023, accessed 17.10.2023; NVIDIA Blog: https://blogs.nvidia.com/blog/2023/03/21/gtc-keynote-spring-2023/, 21.03.2023, accessed 17.10.2023.
[89] Amba Kak and Sarah Myers West (2023). "Computational Power and AI", AI Now. Microsoft also identifies, in its annual report, the shortage of GPU as an operational risk (Microsoft Corporation's Annual Report, Form 10-K, from 06.2023).
[90] Reuters News (2023). "ChatGPT-owner OpenAI is exploring making its own AI chips", accessed 22.10.2023.
[91] CMA (2023). "AI Foundation Models Initial Report", paragraphs 4.48-4.52, available here.

Box 5 – Competition in the cloud sector

The increasing use of cloud infrastructure services has prompted studies on market concentration and its downstream effects.

The supply of cloud infrastructure services is highly concentrated. As of 2023, AWS and Microsoft Azure were the two leading providers of these services worldwide, with Google as their closest competitor. In the UK, Ofcom estimates that AWS and Microsoft had, in 2022, a combined market share of 70% to 80% and Google had a share of 5% to 10%.[92]

The three largest firms are known as the "*hyperscalers*", have a large network of data centres and a wide range of services, enabling them to enjoy economies of scale. There are smaller providers with infrastructure; independent software vendors (ISV) that supply cloud services but do not own the underlying infrastructure; and other types of suppliers include resellers, consultants, etc.[93]

Competition between cloud providers is mainly focused on attracting new customers when they first choose a cloud service provider. This first moment entails benefits for customers, including product innovation, and a wide choice of software services from ISV.[94]

A set of authorities[95] found barriers to switching and multi-cloud after the initial choice of a cloud provider, limiting competition in the market. The following issues were highlighted: *(i)* complex tariff structure and discounts used to incentivise customers using a single cloud service provider, *(ii)* egress fees (i.e., costs of moving data from the cloud to another provider) inducing a lock-in effect and *(iii)* low level of interoperability and data portability.

It was further identified by the AdlC and the ACM that it is likely that the regulations under discussion, such as the European Data Act,[96] will address some of the highlighted competition issues, namely by enabling data portability and interoperability.

V. COMPETITION DOWNSTREAM OF GENERATIVE AI FOUNDATION MODELS

Once trained, foundation models are adapted to specific needs, including by different providers downstream. Foundation models are subjected to transfer learning techniques, namely fine-tuning, and customisation (e.g., via plugins). These steps can be carried out by

[92] Ofcom (2023). "Cloud services market study – Final report", 05.10.2023.
[93] Ofcom, 2023. "Cloud services market study – Final report", 5 October 2023, pp. 39-41.
[94] Idem, p. 3; and Netherlands Authority for Consumers and Markets (ACM). 2022. "Market Study Cloud services" Case no. ACM/21/050317, available here.
[95] AdlC (2023). "Summary of Opinion 23-A-08 of 29 June 2023 on competition in the cloud sector"; ACM (2022). "Market Study Cloud services"; Ofcom (2023). "Cloud services market study – Final report". Meanwhile, other authorities have opened studies in this sector. In June 2023, the FTC launched a public consultation, and comments are available here. In October 2023, the CMA published an issues statement identifying the theories of harm to be assessed in the market investigation into cloud services.
[96] At time of writing, the legislative procedure concerning the Data Act was still ongoing. This Act aims to complement the Data Governance Act. Press Release, European Commission "Data Act: Commission welcomes political agreement on rules for a fair and innovative data economy", 28.06.2023.

different firms, establishing vertical and horizontal relationships that are relevant from a competition point of view (see Figure 4).

V.1. Generative AI fine-tuning models

Access to foundation models plays a central role in the dynamics of competition in downstream Generative AI markets. The type of foundation model, its size, as well as the speed and rate of access[97] to the foundation model are key characteristics in the ability of downstream Generative AI service providers to compete.

Foundation models can be accessed in various ways, differing in their degree of openness. The most open formats for accessing foundation models are open source, which reduces third party reliance on paid foundation models. In this case, the files containing the model's parameters are shared, so a third party can download them and run the foundation model in a controlled environment, either locally or in the cloud.[98] Open-source access may demand a higher level of expertise from third parties, but it grants them greater flexibility in terms of the type of training they can do and the models they can develop.

Within the open-source format there can be differences in the degree of openness, even restricting the possibility of commercial applications, which, in turn, restricts the offer of AI services downstream. For example, a more open format might involve sharing the model's source code, detailed documentation about the source code and the model, training data or model monitoring data. There may also be differences in the rights to access and use the foundation model.[99]

To date, access to pre-trained models in open source has been relatively common in the sector, which has contributed positively to advances in Generative AI. There are even numerous websites that make pre-trained models freely available, including programming library modules,[100] dedicated model repositories[101] or repositories of academic AI publications that include resources for replication.[102]

In more closed access formats, third parties do not have direct access to the model. Instead, access is mostly done via an API. In this format, a third party uploads the training

[97] E.g., how many requests can be made per unit or how much data may be transferred.
[98] E.g., LLaMA by Meta is a model released in open source and many variants have been made from it, namely by continuing the training of the foundation model or by fine-tuning. Cf. Zhao et al. (2023). A Survey of Large Language Models, pp. 10-11.
[99] For example, the LLaMA model by Meta has restrictions, namely on the possibility to use it for commercial ends. CMA (2023). "AI Foundation Models Initial Report", paragraph 3.105 and citations, available here.
[100] E.g., the TensorFlow repository, a library for AI development and machine learning in Python. Available here.
[101] E.g., the Hugging Face repository, aimed at sharing models and datasets for AI and machine learning. Available here.
[102] E.g., the PapersWithCode repository, which includes some of the most recent academic publications on AI and machine learning, with code, datasets, and models for replication. Available here.

data to a server of the provider of the foundation model. Subsequently, the fine-tuning model is trained and created. The third party can then submit requests to the new model, which, in turn, provides responses through the API.[103]

Access to pre-trained models via API is closely related to cloud computing services. Fine-tuning training requires computing power. If third parties do not have direct access to the model, it will be trained in a cloud owned by or in partnership with the provider of the pre-trained model.

Furthermore, there may be dedicated platforms and applications where fine-tuning training data may be uploaded to train models. The same platform may be used to train and then use the resulting model.[104] These services could be similar to what exists today in digital advertising, where similar tools are employed to improve ad targeting.[105]

V.2. Plugins for Generative AI

Plugins are software components that can be used to expand the features of Generative AI models. In general, plugins connect a Generative AI to the API of a service and make requests. The service that the Generative AI connects to can be a normal API (i.e., not "intelligent") or another Generative AI. Connecting one Generative AI to another can significantly expand its features. For example, a text-based Generative AI can, at the user's prompt, make requests to an image-based Generative AI. So, from the user's point of view, the text Generative AI produces text and images.[106]

Using plugins, Generative AI can leverage additional information beyond what was embedded in the model during its training, including real-time updates. These plugins not only broaden the range of content that Generative AI can produce but also expand the types of analyses and processing it can undertake.

---

[103] E.g., API for accessing GPT-3.5.
[104] AutoTrain from Hugging Face is an early example of such a platform.
[105] See, for example, the segmentation service Customer Match, available on Google Ads.
[106] E.g., the paid versions of ChatGPT, Plus and Enterprise, introduced in October 2023 a plugin that allows ChatGPT to send prompts to Dall-E 3. Cf. OpenAI (2023). Dall-E 3 in ChatGPT, available here.

What plugins can bring to users?

Here are some examples of the potential of plugins in the case of Generative:

Financial analysis

A financial data plugin allows AI to access financial data in real-time. Using this information, the AI can automatically produce a financial analysis of a company.

Travel planning

In a travel planning plugin, a user provides the AI with details about her desired destination, travel dates and planned activities. Based on this information, the AI suggests possible accommodations and flights. Once the user selects a hotel or a flight, the AI directs them to the travel agency's website through a provided link for booking.

Product recommendation

A user looking for a particular product can describe its characteristics to the AI. A product recommendation plugin can connect to online shops or product review pages and find what the user is looking for. It can also send the user a link to buy the product online.

Search engine and Internet

The AI can connect to a search engine to retrieve information that answers the user's request. After visiting several pages in the search results, it compiles the answers into a summary that it sends to the user.

Questions about documents and summaries

In this plugin, the user sends documents (e.g., PDF) to the AI and asks the AI questions about the contents of these documents. Using this service, the AI may answer to specific questions on the contents of books or summarise chapters.

Running programming code at the user's request

One of the features of Generative Text AI is to generate programming code from a text prompt. With a programming interpreter plugin, this code can be run by the AI at the user's request.

Box 6 – An illustration of the customer journey using the Expedia plugin in ChatGPT

1. The user requests a travel plan

I want to visit Lisbon. Help me plan my trip to Lisbon.

Of course! Let's plan an exciting trip to Lisbon for you. We'll go step-by-step:

1. **Lodging**: First, we'll find a comfortable place for you to stay.
2. **Flights**: Next, we'll look for the best flights to get you to Lisbon.
3. **Activities**: We'll then explore some fun activities you can do while in Lisbon.
4. **Car Rentals**: Lastly, if you're interested, we can also find a car for you to rent.

Let's start with lodging.

Lodging in Lisbon:
To find the best accommodations for you, could you please provide me with the following information:

- How many people will be traveling with you?
- How many bedrooms would you prefer?
- Any specific amenities or features you'd like in your accommodation? (e.g., gym, pool, breakfast included, etc.)

Feel free to provide as much or as little detail as you'd like, and I'll do my best to find the perfect place for you!

2. The user chooses an accomodation

Used Expedia

Great choice! Here are some top accommodations in Lisbon that have a pool and are suitable for a solo traveler:

1. Hotel A
   - Description: This luxury hotel offers a full-service spa, a restaurant, and a 24-hour fitness center. It's a smoke-free property with free WiFi in public areas. Additionally, there's a bar/lounge, a snack bar/deli, and a sauna.
   - Location: Located in Lisbon City Center, it's just steps from Coliseu dos Recreios and Avenida da Liberdade. Rossio Square and Santa Justa Elevator are also within 2 mi (3 km).
   - Star Rating: ⭐⭐⭐⭐⭐
   - Guest Rating: 9.2/10
   - Average Nightly Price: $296.03 (with a 10% promotion)
   - Check-in: 5th January 2024
   - Check-out: 9th January 2024

Hotel A is fine.

3. The AI gives the user a link to make a reservation and offers additional services

Fantastic choice! **Hotel A** is a luxurious and elegant hotel located in the heart of Lisbon's historic district. With its full-service spa, top-notch amenities, and close proximity to major attractions, you're sure to have a memorable stay.

You can book your stay at Hotel A here.

Now, let's move on to the next step. Would you like assistance with finding the best flights to Lisbon? If yes, could you please provide me with your departure city or airport? If not, we can jump to activities or car rentals. Let me know how you'd like to proceed! 😊

4. The user chooses a flight

I will be departing from Madrid.

Great! Let's find the best flights for you from Madrid to Lisbon Humberto Delgado Airport.

Here are the top flight recommendations for your journey:

1. Flight A
   - Departure: 5th January 2024, 10:50 PM from Madrid (MAD)
   - Arrival: 5th January 2024, 11:10 PM at Lisbon (LIS)
   - Duration: 1h 20m
   - Price: $30.43
   - Non-stop flight

Flight A is fine.

Excellent choice! **Flight A** is a convenient and affordable option for your journey from Madrid to Lisbon.

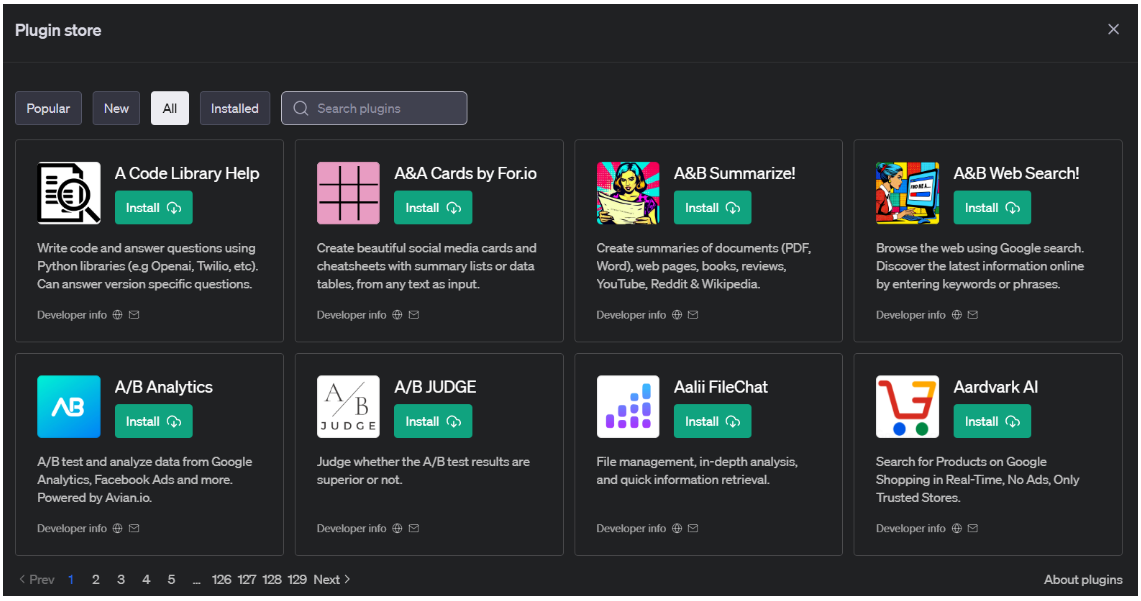
You can book your flight on Expedia here.

Now that we've sorted out your flight, would you like recommendations for activities to do in Lisbon? Or perhaps you're interested in car rentals? Let me know how you'd like to proceed, and I'll assist you further! 😊

Source: Prompt made in ChatGPT Plus using the Expedia plugin.

Plugins can be developed by third parties, making it possible to expand the number of suppliers involved in Generative AI. For example, OpenAI announced in March 2023 that it had made the documentation needed to create ChatGPT plugins available to third party plugin developers. The first plugins were created by Expedia, FiscalNote, Instacart, KAYAK, Klarna, Milo, OpenTable, Shopify, Slack, Speak, Wolfram and Zapier.[107] There are currently over 1000 plugins available on ChatGPT.

Figure 5 – The plugin store in ChatGPT Plus



Source: https://chat.openai.com/, accessed 26.10.2023.

Plugins are a possible source of network effects and a factor of concentration in Generative AI markets. Plugins allow new markets to form. They join, on one side, plugin users and, on the other side, suppliers of plugins. Generative AI with a greater number of plugins could have a wider range of features and be more attractive to end users. Thus, to the extent that plugins become a key element of the Generative AI, it is possible that ecosystems will develop in the sector around the main developers of Generative AI, to capitalise on network effects. The organisation of Generative AI in ecosystems, guided by network effects, could mean a high concentration in these markets.

Plugins are thus analogous to applications for operating systems, due to their ability to expand the features of Generative AI, the way they are distributed and the associated competition concerns.

[107] See OpenAI Blog (2023). "ChatGPT plugins", accessed 22.10.2023.

Plugins can be distributed in different ways, depending on the Generative AI service. In services directed towards end users, plugins can be distributed via plugin stores.[108] Similar to app stores, users can search for and install plugins in their account (i.e., user interface) of the Generative AI service.[109] In services directed towards businesses or more sophisticated users, plugins can be distributed differently, depending on the degree of openness of the Generative AI service and the sophistication of the users (e.g., open source).

As for the business model of plugins, it seems to be too early to map out which forms of monetisation will be most common in the future.

VI. POTENTIAL RISKS TO COMPETITION

Many digital products and services, including Generative AI, have characteristics that make them prone to high levels of market concentration. The same holds true for some adjacent markets, such as network and computing infrastructures that underpin the digital sector.

To that extent, some players may acquire significant market power, and have the ability and the incentives to create and exploit possible bottlenecks in the market. Such conduct, aimed at giving a competitive advantage to certain products at the expense of others, may undermine the competitive process and harm consumers.

In general, these possible bottlenecks in the market stem from having access and control to some resource necessary to have a competitive offer in the market. Examples include databases, the data collection capabilities, computing power, user bases and network effects, or the accumulated experience of a firm.

This section aims to categorise some of the potential risks to competition which, based on the characteristics identified above, may be conceptually associated with the markets under analysis. This section also covers some of the decisions by the European Commission in the digital sector.

VI.1. Leveraging risks in the integration of Generative AI in other products

The developers of Generative AI services may have incentives to integrate AI into other products, in an ecosystem of products and services.[110] For example, a chatbot services may be integrated, or offered alongside, search engines, operating systems, productivity software or cloud computing services.

---

[108] For example, in ChatGPT Plus, users access plugins in a way akin to accessing apps in an app store. Cf. OpenAI documentation, available here, and https://chatgpt.plugin.support/, accessed 22.10.2023.

[109] It should be noted the obligations in the Digital Markets Act could be applicable to plugins as they can be placed under online intermediate services. Cf. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), available here.

[110] AdC (2019). Issues Paper "Digital Ecosystems, Big Data and Algorithms", pp. 12-14, available here.

This kind of integration, in an ecosystem, can lead to better products and more attractive offers for consumers, namely due to the convenience of having a one-stop shop**.**

However, it may also entail risks to competition. Hypothetically, a tying strategy[111] **of Generative AI with core products may accelerate their development**, at the expense of competing AI.[112] One market player may, for example, leverage its large number of users of a core product or service, channelling them to its own Generative AI. If the Generative AI benefits from significant scale effects, this conduct may give a competitive advantage to this player's Generative AI. Forcing customers to acquire a specific product or service to access a Generative AI service may also entail foreclosure effects, in particular, if the firm has a dominant position in the Generative AI market. The incentives for market foreclosure depend, however, on the business models and strategies of monetisation in the markets of foundation models and other downstream Generative AI services, among other factors.

Conceptually, there is also a risk to competition if a firm with market power in a Generative AI market uses its position to favour its own products and services, by integrating its AI in an exclusive manner. The market power of a Generative AI provider may stem from the accumulation of various advantages, such as better optimised models or access to better datasets.

Therefore, leveraging strategies from one market to the other are more concerning if carried out by a firm with a dominant position in a market (whether upstream, downstream or in adjacent markets) with the intent of excluding competitors. Box 7 illustrates how certain strategies, such as tying, by dominant firms may breach competition law.

Box 7 – Abuse of dominance cases in the digital sector, in the EU

This section highlights some of the investigations and decisions on abuse of dominance in the digital sector, in the EU.

In July 2018, the EC fined Google €4.3 billion for abusing its dominant position in general internet search services, mobile operating systems, and app stores for the Android. In particular, the EC found that Google required manufacturers to pre-install Google Search and Google Chrome on their Android devices, as a condition to access Google Play Store. Google also made payments to large device manufacturers and mobile network operators

---

[111] According to the Guidance on exclusionary abuse by the Commission, "[a] *«Tying» usually refers to situations where customers that purchase one product (the tying product) are required also to purchase another product from the dominant undertaking (the tied product)*". Communication from the Commission — Guidance on the Commission's enforcement priorities in applying Article 82 [102] of the EC Treaty to abusive exclusionary conduct by dominant undertakings (p. 48).

[112] This risk was also identified by the FTC and the CMA, as well as in other articles on the topic. FTC Blog (2023) "Generative AI Raises Competition Concerns"; CMA (2023) "AI Foundation Models Initial Report", para 4.27, available here; Carugati, C. (2023). "Competition in generative artificial intelligence foundation models", available here.

on condition that they exclusively pre-installed Google Search on their Android devices. Google aimed at strengthening its position in the market for general internet search.[113]

On 24 March 2004, the EC issued a decision against Microsoft for abusing its dominant position in the market for computer operating systems.[114] The abuse involved: *(i)* refusing to supply interoperability information indispensable for its competitors to be able to viably compete in the work group server operating systems market; and *(ii)* tying its media player with its operating system. Microsoft's behaviour enabled it to strengthen a dominant position in the market for work group server operating systems and weakened competition in the media player market. The EC imposed on Microsoft a fine of around € 497 million and the following remedies: *(i)* the disclosure and interoperability of interface specifications to undertakings interested; and *(ii)* the offer of a version of Windows Operating System which does not include Windows Media Player.

In May 2022, the AdC opened an investigation against Google for possible abuse of dominance in online advertising.[115] The AdC identified indicia of self-preferencing behaviours by Google at various stages of the digital advertising chain. The investigation by the AdC focused on the possibility that Google may have used information not accessible by competitors on online advertisement auctions to change the outcome of those auctions in its favour. In July 2022, the EC informed the AdC that it intended to extend the scope of its own investigation on Google to also include the practices and markets under investigation by the AdC. In September 2022, the AdC closed the investigation, which was henceforth conducted by the EC.

In June 2023, the EC announced it had sent a Statement of Objections to Google for favouring its own online display advertising intermediation services. This sector joins, on one side of the market, advertisers that wish to show their advertisements to consumers and, on the other side of the market, publishers that want to sell their advertising space. In this sector, ads are selected through auctions. The EC found that Google has benefited its own online advertising intermediation services, by organising and participating in auctions that select which ads are shown in an advertising space. In addition, the EC found that Google leveraged the number of advertisers that use its own services in favour of other online intermediation services, namely by avoiding bidding in auctions organised by Google's competitors.[116]

In July 2023, the EC announced it had opened an investigation against Microsoft for a possible tying conduct of its service Teams with the core services for enterprises Office 365 and Microsoft 365, to protect its position in the market of productivity software.[117]

---

[113] Commission Decision of 18.07.2018 relating to a proceeding under Article 102 of the TFEU and Article 54 of the EEA Agreement (Proceeding AT.40099 – Google Android), available here.
[114] Commission Decision of 24.03.2004 relating to a proceeding under Article 82 of the EC Treaty (Case COMP/C-3/37.792 Microsoft), available here.
[115] Case details available here. See, also, AdC (2022) Policy brief "Defence of competition in the Digital Sector in Portugal".
[116] Commission Press Release (2023) "Antitrust: Commission sends Statement of Objections to Google over abusive practices in online advertising technology", 14.06.2023.
[117] Commission Press Release (2023) "Antitrust: Commission opens investigation into possible anticompetitive practices by Microsoft regarding Teams", 27.07.2023.

VI.2. Privileged access to data for the development of Generative AI

A firm that offers Generative AI services may have access to important datasets that are hard to replicate by competitors, for example, because of their presence in other digital markets. It is common that a firm is present in several markets at the same time in the digital sector given that it is organised, to a great extent, in digital ecosystems of products and services.[118] In addition, datasets, namely those obtained through other products, may be crucial to the development of Generative AI models, whether foundation or fine-tuning models.

A firm may have the incentive to give privileged or preferential access to these datasets to its own Generative AI, at the expense of competing AI. This behaviour may distort competition, in particular if these datasets play an instrumental role in developing competing Generative AI services and the firm with privileged access has a dominant position in the relevant market.[119] The negative impact on competition, due to privileged access to data, will be greater the more essential the data is and if the firm in question enjoys significant market power.

VI.3. Privileged access to Generative AI models

Foundation models are the cornerstone of downstream Generative AI services, closer to end users. The developers of foundation models are, for this reason, in a unique position to develop and deploy their own Generative AI services downstream. In particular, the developers of foundation models may find it easier to adapt them to specific tasks, benefitting from *(i)* specialised knowledge about its own foundation model, *(ii)* the same computing power they used to develop the foundation model and *(iii)* the generalised nature of the foundation model.

Therefore, the developers of foundation models may have the ability and the incentives to give privileged or preferential treatment to their own downstream Generative AI services, at the expense of other downstream Generative AI services. For example, the developer of the foundation model may give its own downstream AI models exclusive access to the best version of the foundation model (e.g., the version with the largest size, trained with exclusive datasets, or the most updated version). Alternatively, for example, the developer of the foundation model may limit the rate of access[120] to the foundation model for competitors but not to its own Generative AI.

This risk will be greater if the developer of the foundation model has market power, and the conduct can potentially raise barriers to the development and deployment of

---

[118] AdC (2019). Issues Paper "Digital Ecosystems, Big Data e Algorithms", pp. 8-11, available here.
[119] Although it did not take in the digital sector, it should be noted that the AdC, in 2015, fined the National Association of Pharmacies for margin squeeze, due to discriminatory access to commercial data on pharmacies. In this case, this data was deemed to be an input in the downstream market for market studies, where National Association of Pharmacies competed with other market participants. AdC Decision PRC/2009/13, available here.
[120] See footnote 97.

competing Generative AI services. In such a scenario, it is possible that such behaviour could substantially curtail innovation in Generative AI services.

The same concerns can be applied downstream, namely in the access to fine-tuning models. A developer of fine-tuning models (or other models derived from foundation models) is in a unique position to develop its own plugins and may have the ability and the incentives to give privileged or preferential treatment to its own plugins. Moreover, the developer of the fine-tuning model may have some degree of control over how plugins are distributed (e.g., plugin stores), which may introduce additional mechanisms of exclusion. The existence or not of incentives for market foreclosure will, nonetheless, depend on the business model and strategies of monetisation that might evolve, among other factors.

VI.4. Demanding unfair trading conditions for supply Generative AI services

A developer of Generative AI with market power may have the ability and the incentives to demand unfair trading conditions from other firms or from users that wish to have access to its Generative AI. In addition to access costs, the terms and conditions governing the access to Generative AI may include the right of the AI developer over the content generated by the AI, as well as rights over the content, data and other information uploaded by other firms and users.

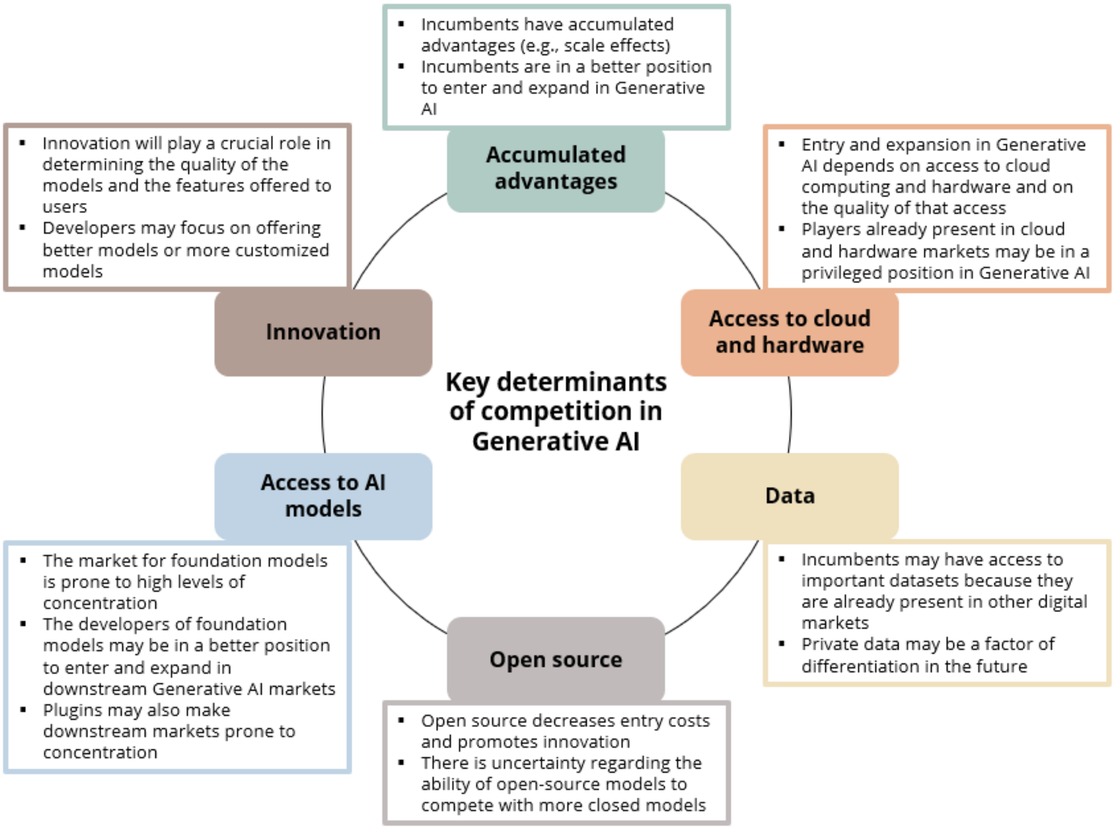
In particular, the access to Generative AI may be granted on the condition that the firm or the user gives all the content and data they upload to the developer of the Generative AI model. For instance, fine-tuning, which adapts a foundation model to a specific task, implies additional training on top of the foundation model. This training requires additional data which the firm accessing the foundation model possesses but the developer of the foundation model does not. Given this, the developer of the foundation model may grant access only if it may use the training data for the fine-tuning model to train its own models.

In conclusion, to ensure that the economy reaps the potential benefits of Generative AI, it is crucial there are no significant barriers that may hinder the development of alternative Generative AI services downstream. In this context, it is worth highlighting that **firms with a dominant position have the responsibility to make sure that their choices and behaviours do not unduly distort competition.**

VII. MAPPING THE DETERMINANTS OF COMPETITION IN GENERATIVE AI

It is possible to map some of the key determinants of competition in the Generative AI sector considering the accumulated experience by national competition authorities in the digital sector and the above-mentioned risks to competition.

Figure 6 – The key determinants of competition in Generative AI



Source: AdC.

Accumulated advantages by incumbents in the digital sector

The evolution of the Generative AI sector may be largely determined by the current competition conditions in the digital sector as a whole. In general, the digital sector is characterised by strong scale effects that make many markets prone to high levels of concentration. A byproduct of these characteristics are the incentives and advantages to develop ecosystems, which result in competitive advantages to the already established firms in the market.

Generative AI has appeared in a context in which digital incumbents have strong accumulated advantages. The fact there are incentives and advantages in organising digital markets in ecosystems may entail that digital incumbents are in a better position to enter and expand in Generative AI markets, strengthening their position in the digital sector.

To fully realise the opportunities that Generative AI brings to the economy, it is important that national competition authorities remain vigilant. The growth of firms in Generative AI markets must be based on competition on the merits and not by leveraging market power.

Access to cloud computing services and specialised hardware

Entry and expansion in Generative AI markets will depend on having access to computing power and on the quality of that access. There are two ways firms may have access to computing power: *(i)* cloud computing services; and *(ii)* by building their own infrastructure and acquiring specialised hardware, such as CPU and GPU. This access relies on the incumbents in the cloud computing and specialised hardware markets which, in turn, may be in a privileged position to enter and expand in Generative AI markets.

Promoting competition in Generative AI markets is intrinsically tied to promoting competition in these upstream markets. More competitive upstream markets will ensure that the access to the inputs required to develop and deploy Generative AI are more competitive, to the benefit of consumers.

Data

Some digital incumbents may have data related accumulated advantages. Data is one of the cornerstones of competition in Generative AI. Both the volume and the quality of training data may have a crucial impact on the performance of Generative AI services and on their ability to compete in the market. The data related accumulated advantages of digital incumbents stem both from having access to existing datasets, but also from having the capabilities to generate new data. One of the sources of data related advantages are large user bases, as users may be monitored or may produce content within these services. This information can be, subsequently, converted into training data for Generative AI models.

It is possible that private data may be increasingly important in the future and become on key factor of differentiation for Generative AI services. A significant proportion of currently available Generative AI models rely on publicly available datasets. However, this trend may shift in the future following the initial proliferation of Generative AI services.

The role of open source in Generative AI

Some of the Generative AI models have been released in an open-source format, allowing third-party developers to offer services based on these models. Access to models in an open format gives third parties the flexibility to use, adapt and expand Generative AI models, decreasing entry costs and promoting innovation in the different applications of Generative AI.

However, the role of open source in fostering competition remains uncertain. On the one hand, the open-source format may not provide enough incentives for the development of models that can rival more closed models. On the other hand, they require some form of monetisation which raises questions on whether it is possible on the feasibility and viability of open-source models. Moreover, open-source models may be made available under conditions that restrict their commercial use by third parties. The open-source format does not necessarily eliminate potential competition concerns either, as an open-source model may be used to strengthen the market power of its developer in an adjacent market.

Access to intelligence artificial models

The generalised nature of foundation models may make the market for foundation models prone to concentration, with an impact downstream. Foundation models are the input for the developers of Generative AI services downstream. A higher level of concentration in a market for foundation models results in decreased bargaining power for developers downstream. The developers of foundation models are also in a privileged position to enter and expand in downstream markets for Generative AI services, which may have an impact on the conditions of access to their models for downstream rivals.

Downstream, plugins may generate network effects and make the market for transfer learning models prone to concentration. For this reason, the same concerns and competition dynamics apply to the access of plugins to transfer learning models. The developers of these models are also in a privileged position to enter and expand potential future markets for plugins.

Innovation

Innovation will be a key element in the Generative AI sector. Innovation can operate via new products and services, or by improving existing ones. As for improving existing products and services, suppliers can produce strictly better ones than their rivals (vertical differentiation), or by adapting models to the specific needs and preferences of users (horizontal differentiation).

Generative AI's capacity for generalisation suggests significant potential for innovation. Competition, as a driver of innovation, is key to fully realise this potential.

In sum, the cornerstones of competition in Artificial Intelligence are the following: *(i)* access to data, *(ii)* access to cloud computing or specialised hardware and (iii) access to foundation models in Generative AI. The AdC, within the scope of its mandate and in the context of international cooperation, will closely follow the developments in markets related to artificial intelligence, and will not hold back from intervening to ensure the promotion and application of competition law in Portugal, for the benefit of consumers and whenever the identified risks materialise.

THE IMPORTANCE OF COMPETITION IN GENERATIVE ARTIFICIAL INTELLIGENCE

How competition can ensure more benefits for consumers and businesses

1. Generative Artificial Intelligence brings **a universe of new products and services to consumers and businesses**, displaying capabilities typically associated with human beings.

2. The potential for applications of Generative Artificial Intelligence appears boundless. It could have a **significant impact across all sectors of the economy**.

3. The Generative Artificial Intelligence sector is growing and has **countless opportunities for innovation**.

4. Competition will be a **crucial tool to fully realise the potential of Generative Artificial Intelligence** and the opportunities for innovation, as a driver of innovation.

5. It is important to **map the key determinants that affect the competitive process** and to anticipate the risks to competition in the Generative Artificial Intelligence sector.

6. **The competition authorities have accumulated significant experience in the digital sector** in recent years, identifying the determinants that, in each market, can generate harmful outcomes for consumers and businesses.

7. Generative Artificial Intelligence is a **poster child for all the challenges digital markets pose to competition**. Generative Artificial Intelligence markets are prone to high levels of concentration, which can give some players the ability and incentives to distort the competitive process.