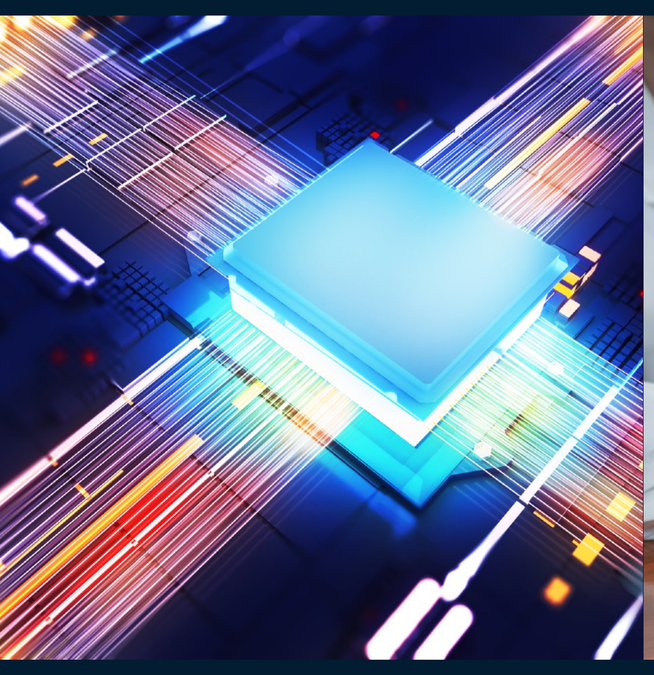
RÉPUBLIQUE FRANÇAISE

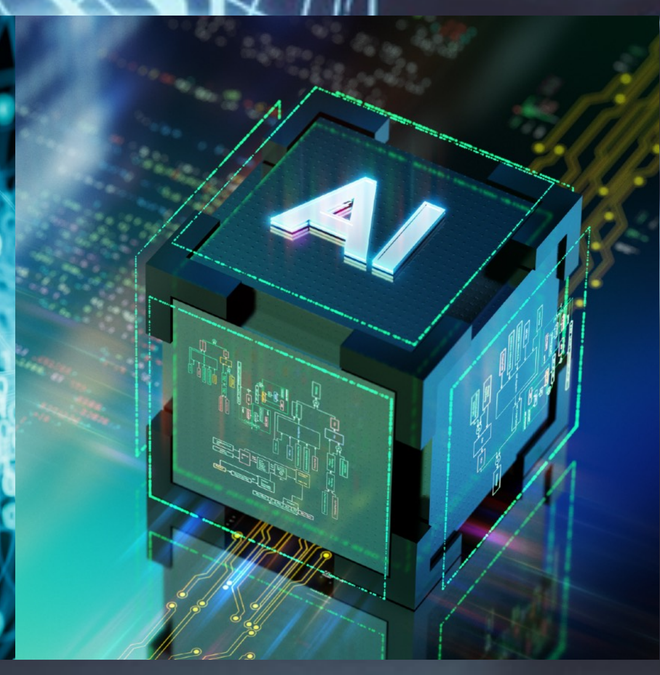
Autorité
de la concurrence

Opinion 24-A-05

of 28 June 2024

on the competitive functioning of the generative artificial intelligence sector





RÉPUBLIQUE FRANÇAISE

Autorité de la concurrence

Only the French version is authentic. In the event of any discrepancy, the French version will prevail over the translation.

Opinion 24-A-05 of 28 June 2024 on the competitive functioning of the generative artificial intelligence sector

The *Autorité de la concurrence* (section IA),

Having regard to Decision 24-SOA-01 of 7 February 2024 registered under number 24/0007 A, by which the *Autorité de la concurrence* decided to start inquiries *ex officio* into the generative artificial intelligence sector with a view to issuing an opinion;

Having regard to Book IV of the French Commercial Code (*Code de commerce*);

Having regard to the public consultation document published by the *Autorité de la concurrence* on 8 February 2024 as part of the *ex officio* inquiry into the generative artificial intelligence sector;

Having regard to the contributions received up to 22 March 2024;

Having regard to the other evidence in the case file;

Representatives of Google, Microsoft, Mistral AI, Orange, France Digitale and the Directorate General for Enterprise (DGE) having been heard on the basis of the provisions of Article L. 463-7, paragraph 2 of the French Commercial Code (*Code de commerce*);

The case officers (*rapporteurs*), the Head of the Digital Economy Unit and the Government Commissioner having been heard at the hearing on 29 May 2024;

Adopts the following opinion:

Summary[1]

Since the public release of the ChatGPT chatbot (created by OpenAI) in November 2022, generative artificial intelligence (hereafter "AI") has taken centre stage in public and economic debate. The questions raised by generative AI range from ethics and respect for intellectual property to the impact on the labour market and productivity. The technology offers numerous possibilities to companies in terms, for example, of content creation, graphic design, employee collaboration and customer service.

The benefits of generative AI will only materialise if all households and companies have access to a variety of different models adapted to their needs. Competition in the sector must therefore be conducive to innovation and allow for the presence of multiple operators.

Generative AI

According to the European Parliament, AI refers to any tool used by a machine "*to display human-like capabilities such as reasoning, learning, planning and creativity*". **Generative AI refers to AI models capable of generating new content such as text, image, sound or video.**

There are two key phases in generative AI modelling:

- ***training*: the initial learning process of a model (often called "foundation model", which includes large language models [LLMs])**, during which its parameters, known as "weights", are determined. Training requires both significant computing power and a large volume of – generally public – data. The training phase may be followed by fine-tuning, during which the model is adapted to a specific task, such as answering end users' questions, or to a specialised dataset (e.g. legal or health-related data). Fine-tuning is generally based on a smaller, proprietary dataset and may involve human expertise;
- ***inference*: the use of the trained model to generate content.** The model can be made accessible to users via specific applications, such as Open AI's ChatGPT or Mistral AI's Le Chat, or APIs for developers. The computing power required depends on the number of users. Unlike many digital services, the marginal cost of generative AI is not negligible, given the cost of the computing power required. New data that was not used for training may be added during the inference phase, in order to ground the model in recent data, such as news articles.

[1] The summary is for information purposes only and provides an overview of the following numbered reasons for the opinion. Only the French version of the opinion is authentic. In the event of any discrepancy, the French version will prevail over the translation.

The participants in the value chain

The generative AI value chain extends upstream from the design, training and inference of models to their use downstream by end users. The operators in the generative AI value chain are:

- **major digital companies**: Alphabet and Microsoft are present across the entire value chain (vertical and conglomerate integration), while Amazon, Apple, Meta and Nvidia are present only at certain specific layers;
- **model developers**: for example, start-ups or AI-focused research labs, such as Anthropic, Hugging Face, Mistral AI and OpenAI. They have often formed partnerships with one or more digital giants, such as OpenAI with Microsoft and Anthropic with Amazon and Google. They may adopt a more, or less, open approach as regards the information available about their models and the possibility of re-using and adapting them.

At the upstream level, several types of operators are involved:

- **IT component suppliers** develop graphics processing units (GPUs) and AI accelerators, which are essential components for training generative AI models. In addition to Nvidia, the sector's leading operator and the world's most valuable publicly-traded company at the date of this opinion, and major digital companies that develop their own AI accelerators, the sector also includes traditional operators like Advanced Micro Devices (AMD) and Intel;
- **cloud service providers** play a key role in the development of new AI technologies, as they provide the storage, data processing and computing capabilities needed, in particular, by language model developers. They include both digital giants, known as "hyperscalers", such as Amazon Web Services (AWS), Google Cloud Platform (GCP) and Microsoft Azure, cloud providers such as 3DS Outscale, IBM and OVHcloud, as well as specialist AI providers such as CoreWeave. The cloud sector was described by the *Autorité* in Opinion 23-A-08 of 29 June 2023. The necessary computing resources may also be provided by public supercomputers (such as Jean Zay in France), which have historically been dedicated to high-performance computing and have diversified to accommodate AI research projects.

At the downstream level, many operators are marketing new services based on generative AI to the general public (like ChatGPT), companies and public authorities and/or integrating generative AI into their existing services (like Zoom).

A growing priority for public authorities

The generative AI sector is attracting growing interest around the world.

In France, the government launched a **national AI strategy** in 2018 aimed at equipping France with competitive research capabilities and deploying AI technologies throughout the economy. In March 2024, the French AI Commission (*Commission de l'IA*) launched by the Prime Minister presented 25 recommendations calling, in particular, for measures to make France a major centre for computing power, to facilitate data access and to establish global AI governance.

At the European level, several legislation governing the development of the AI sector have been adopted over the last two years. In particular, most of the provisions of the **AI Act** (which will soon be published in the EU Official Journal) will be applicable from 2026. Under the Act, providers of large generative AI models will be subject to transparency obligations and must implement policies to ensure compliance with EU copyright law when training their models. The obligations do not apply to free and open-source models, whose parameters are public (unless they present a systemic risk). Although published before the rise of generative AI, the Digital Markets Act (DMA) and the Data Act will have an impact on the sector.

A series of initiatives on AI have been adopted globally, such as the Bletchley Declaration in the United Kingdom in November 2023 at the AI Safety Summit. **The next global summit will take place in France on 10 and 11 February 2025.** Other initiatives have been taken by the G7, the United States, the United Kingdom and China, for example.

High barriers to entry

Access to sufficient computing power for performing a large number of operations in parallel, and with the high precision needed to determine several billion parameters, is essential for developing foundation models. The GPUs developed by Nvidia (combined with its CUDA software) or the AI accelerators developed by major digital companies (such as the tensor processing units [TPUs] developed by Google) are essential for the training, fine-tuning and inference of generative AI models. They are also very expensive. Since 2023, the sector has experienced shortages due to an explosion in demand.

Aside from digital giants and a few companies with sufficiently large in-house data centres (like Meta or Samsung), **the cloud appears to be the only way to access the computing power needed to train models**. The cloud gives developers access to AI-specific infrastructure and platform services adapted to their needs, while avoiding massive initial investment in IT infrastructure. **The cloud is also a vector for distributing models downstream on marketplaces.**

In addition, training large generative AI models requires **large volumes of data**. Most of this data is obtained from publicly-accessible sources, such as web pages, or datasets like the Common Crawl web archive (an organisation that has been providing free data from the Internet since 2008). The cleansing and processing of this data is a differentiating factor, as operators need to filter the data in order to keep only qualitative content.

The stakeholders consulted as part of this opinion expressed **concerns about data access**. On the one hand, models are getting bigger and bigger and training requires more and more data, raising fears that publicly-accessible data will not be sufficient in the future and that proprietary data held by a small number of operators will become more important. On the other hand, access to certain publicly-accessible data is creating **legal uncertainties**, as illustrated by the actions brought by several rights holders, such as the complaint filed by the New York Times against OpenAI and Microsoft.

Lastly, training large models also requires **highly advanced technical skills** in machine learning, as well as empirical experience that can only be acquired by working with the models.

Operators in the generative AI sector require **substantial funding** to meet their computing power, data and skills needs. Investment in the sector increased six-fold between 2022 and 2023, to more than €20 billion.

Barriers to entry potentially limited by technical and organisational developments and certain public policies

First, computing power can be accessed via **public supercomputers**. In return for contributing to open science (for example, publishing work in an academic journal), access to public supercomputers is free, which can help to reduce the barriers to entry for certain operators, in particular in the research world. For example, a team of researchers from the CentraleSupélec university has trained a model called "CroissantLLM" on the French supercomputer Jean Zay. The joint undertaking EuroHPC is working to develop supercomputers throughout Europe and plans to **install a new supercomputer in France in 2025**.

Second, **a number of technological innovations** are already reducing the need for data and computing power:

- **innovations in generative AI model architecture**, which are making the training and fine-tuning phases more efficient and less costly. Examples include Mixture of Experts (MoE) and Low Rank Adaptation (LoRA);
- **smaller models**, which are easier to use for the inference phase and can be used on smartphones, for example;
- **synthetic data** (also generated by AI), which can partially replace real data and reduce the constraints associated with the use of personal data. However, the use of synthetic data entails certain risks, such as bias or a higher error rate.

Lastly, many developers choose an **open-source approach** in order to contribute to overall knowledge about the technology, thereby enabling other operators to re-use or fine-tune the models. However, open source covers a wide range of scenarios, from open-weights models where only the model weights are made public (the most common scenario) to fully-open models where all the code, architecture, training data, weights and learning process are made public. While publishing model weights can have a beneficial impact on competition for fine-tuning and inference, it does little or nothing to reduce the barriers for an operator wishing to train a foundation model. In order to reproduce an AI model, other elements would need to be made public, such as the code and data for training or the data used.

Advantages for some companies linked to their activities in other digital markets

Major digital companies enjoy preferential access to the inputs needed to train and develop foundation models. Developers of competing foundation models, which do not have access to these inputs under the same conditions, cannot easily replicate these advantages.

They have **easier access to computing power** as partners and competitors of AI chip suppliers. On the one hand, they are able to buy in large quantities and negotiate preferential agreements with GPU suppliers like Nvidia. On the other hand, most of them are also developing in-house AI accelerators specifically tailored to their ecosystems, such as Google's TPUs and AWS' Trainium. Major digital companies are also starting to develop alternatives to Nvidia's CUDA software.

They also enjoy **preferential access to large volumes of data** (as an example, YouTube provides Alphabet with a major source of training data for AI models). They can also access **data associated with the use of their services**, as well as use their financial power to enter into agreements with the owners of third-party data, as demonstrated by Google's agreement to pay $60 million (around €55 million) a year for access to data from Reddit, a US social news aggregation and forum social network.

In addition, **many highly-skilled employees** are enticed by the attractive salaries and job prospects offered by major digital companies, given their reputation for innovation, their global positioning and their wide catalogue of services.

In addition to unrivalled access to the inputs needed to train generative AI models, major digital companies enjoy **advantages linked to their vertical and conglomerate integration, which guarantees access to users, companies and consumers**. The sector is characterised by the high fixed costs involved in the initial training of a foundation model, which gives rise to **economies of scale** as operators seek to spread costs over as many users as possible. Generative AI products are also characterised by economies of scope because, once developed, a foundation model can be used for a wide variety of applications. The generative AI sector can also give rise to **cumulative network effects**, with feedback data from users being used to refine future models and improve performance or offer new services.

The *Autorité* also notes that major digital companies are starting to integrate generative AI tools into their **product and service ecosystems**. For example, Microsoft deploys its own models and those of its partner OpenAI in the "Copilot" function to enhance Microsoft Bing's search functionality and offers an AI assistant designed to work with the Microsoft 365 offering, "Copilot for Microsoft 365". In addition, major digital companies' **marketplaces (Model-as-a-Service [MaaS])** provide access to proprietary and third-party generative AI models designed to run in their ecosystems.

Competition risks upstream in the value chain

While it seems premature at this stage to draw definitive conclusions about the definition of relevant markets and the market power of certain operators, vigilance is nevertheless required because major digital companies' access to key inputs and the advantages linked to their vertical and conglomerate integration create the conditions for strong concentration, to their benefit, and reinforce their power on distinct but linked or related markets, such as office productivity software, search engines or online advertising. **In certain cases, it may therefore be useful to perform the competitive analysis in terms of ecosystems being either created or reinforced, rather than market by market.**

The traditional tools of competition law, **such as antitrust law and, above all, abuse of dominant position**, remain fully relevant. Other legal tools could also be used, such **as abuse of economic dependence**, where no position of dominance exists, or, with regard to contractual practices, the **law on restrictive competition practices**, the implementation of which falls mainly within the remit of the Directorate General for Competition Policy, Consumer Affairs and Fraud Control (DGCCRF) and the commercial courts.

Several risks of abuse identified by the *Autorité*

- **The risk of abuse by IT component providers**

France Digitale, an association representing a large number of French digital start-ups and investors, points to potential risks such as **price fixing, supply restrictions, unfair contractual conditions and discriminatory behaviour**. Concern has also been expressed regarding the sector's dependence on Nvidia's **CUDA** chip programming software (the only one that is 100% compatible with the GPUs that have become essential for accelerated computing). Recent announcements of Nvidia's investments in AI-focused cloud service providers such as CoreWeave are also raising concerns.

The graphics card sector, which was the target of a dawn raid in September 2023, is being closely scrutinised by the *Autorité*'s Investigation Services.

- **The risk of lock-in by major cloud service providers**

The *Autorité* notes that several financial and technical lock-in practices, already identified in Opinion 23-A-08 on competition in the cloud sector, appear to remain and even to be intensifying to attract the largest possible number of start-ups active in the generative AI sector.

First, **particularly high levels of cloud credits** are being offered to innovative companies in the sector. **Technical lock-in** practices have also been identified.

Such practices could be assessed under competition law, in particular on the basis of abuse of dominant position. Some of the practices are also governed by French law 2024-449 of 21 May 2024 to secure and regulate the digital space (known as the "SREN Law") or by the EU Data Act.

❖ **The risks associated with data access**

Innovative companies in the sector may be confronted with practices of **refusal of (or discriminatory) access** by companies with significant access to data, such as a web index.

In addition, agreements under which major digital companies impose exclusive access to content creators' data, or pay them substantial remuneration that is difficult for their competitors to replicate, could constitute anticompetitive practices (cartels or abuse).

Access to user data is also a major challenge. Several stakeholders reported that major companies in the sector continue to use various strategies to restrict third-party access to their users' data, by abusing legal rules, such as personal data protection, or security concerns.

Lastly, content publishers are very concerned about the use of their content by foundation model providers without the **authorisation of rights holders**. In Decision 24-D-03 in the "related rights" case, the *Autorité* established that Google had used content from press agencies and publishers to train its foundation model Gemini (a chatbot based on the foundation model of the same name and formerly called "Bard"), without notifying them and without giving them an effective possibility to opt-out. While this question raises issues relating to the enforcement of intellectual property rights that go beyond the scope of this opinion, competition law could, in principle, address these issues based on an infringement of fair trading, for example, and therefore, exploitative abuse.

❖ **The risks associated with access to a skilled workforce**

In competition law, supervisory authorities pay particular attention to practices in the labour markets. In addition to wage-fixing agreements, no-poach agreements may also constitute prohibited anticompetitive practices.

An additional area of concern is the **recruitment by digital giants of entire teams** (such as Microsoft's hiring of most of start-up Inflection's 70-person staff) or **strategic employees of model developers** (such as Microsoft's brief recruitment of Sam Altman, the founder of OpenAI, before he was eventually hired back by OpenAI). While this type of practice may be examined under merger control rules, it can also be analysed as an attempt to exclude competitors from the sector.

While it appears from the preparation of this opinion that such restrictions are not, for the time being, raising any particular concerns for stakeholders, the *Autorité* considers that vigilance is required.

❖ **The risks associated with open-source models**

While open-source models can help to lower barriers to entry, they can also raise competition concerns. In some cases, the conditions of access and re-use of models or some of their components can lead to users being locked-in.

❖ **The risks associated with the presence of companies on several markets**

The vertical integration of certain digital operators and their service ecosystems may give rise to a number of abusive practices.

At the upstream level, model developers could be **denied or given limited access to the chips or data needed to train competing foundation models**. This type of practice could lead to delays or the introduction of less ambitious models, thereby undermining effective competition in the market.

Several stakeholders are also concerned about **exclusivity agreements** between cloud service providers and foundation model developers. In their view, such agreements aim to make the developers exclusively dependent on the cloud service providers for access to the necessary cloud services and for customer distribution, and are therefore likely to have an **impact on innovation** and competition between providers, especially when a particular model occupies a significant position on the market.

Other risks arise from the downstream use of generative AI models, through **practices of tying**. Companies holding pre-eminent or dominant positions in AI-related markets could tie the sale of products or services to that of their own AI solutions. **In particular, the integration of generative AI tools on certain devices, such as smartphones, is raising concerns.** This type of practice could permanently consolidate the generative AI sector around already dominant digital companies.

Downstream competitors could also be harmed by **self-preferencing** practices of vertically integrated operators, affecting the ability of developers of non-vertically integrated models to compete with those operators.

Through any of the above behaviours, certain companies could use their market power in distinct but related markets to the detriment of alternative operators, thereby restricting the choice available to users and the incentive to develop alternative solutions.

Competition concerns about minority investments and partnerships by digital giants

In a sector such as AI, where investment is very high given the cost of access to inputs, only a few major players have the financial capacity to enter into agreements with or invest in innovative start-ups. Investments and partnerships between operators in the sector are not problematic per se. They can give start-ups the opportunity to benefit from the financial and technological resources of major companies, and thus foster innovation. For the buyer, such investments enable diversification or access to innovative technologies to improve the quality of its services. For example, Microsoft has entered into an exclusive partnership with OpenAI in the form of a multi-year investment.

Nevertheless, they present significant risks that call for particular vigilance by competition authorities. They may weaken competition between the two entities, lead to **vertical effects**, **increase market transparency** or **lock-in** some parties.

Minority investments by major companies may be assessed by competition authorities on several legal grounds. On the one hand, the transactions may be subject to prior authorisation under merger control rules if they give investors de facto control and exceed EU and national notification thresholds. They may also be examined, under certain conditions, if they are below said thresholds, or as part of the analysis of a merger. On the other hand, they may be assessed ex post through competition law, on the basis of antitrust law or abuse of dominant position (including collective dominance). However, the *Autorité* notes a **lack of transparency** in agreements, which can make it difficult to determine whether they are likely to harm competition and hence consumers. These concerns are shared by competition authorities around the world, as evidenced by ongoing investigations into Alphabet, Amazon, Anthropic, Microsoft and OpenAI.

The risk of collusion between companies in the sector

While almost all the stakeholders consulted during the public consultation did not express any specific concerns about the risk of collusion, the use of generative AI could potentially give rise to concerted practices that are already known and which were the subject of a joint study in 2019 by the *Autorité* and the German *Bundeskartellamt*, such as the parallel use of separate individual algorithms or the use of machine learning algorithms. Here too, vigilance is essential.

Outlook

The *Autorité* notes that generative AI is far from having reached its potential. Less than two years after the launch of ChatGPT, many established operators have invested in the field and a multitude of start-ups have emerged to accelerate research and deploy the technology to companies and consumers.

The **race to innovate** and develop new generative AI models is likely to continue on two aspects: model size and optimisation at constant size. Model size is also a key factor in the **environmental impact of generative AI**.

The *Autorité* has also observed a trend towards **"platformisation"** in the generative AI sector. MaaS seems to be only way for model developers to reach consumers and AI-using companies.

One of the main challenges for the healthy development of competition in the generative AI sector lies in the deployment of open-source resources. If the sector had more precise criteria for qualifying the degree of openness of a model, operators who so wished could use model openness as a competitive advantage.

Recommendations

Competition in the sector could be strengthened by the following recommendations, most of which do not require new legislative initiative at the French or European level.

The *Autorité* calls for full use to be made of the regulatory framework applicable to the sector.

The Commission should pay particular attention to the development of services that give access to generative AI models in the cloud (MaaS) and assess the possibility of designating companies providing such services as gatekeepers specifically for those services, under the DMA. Some of the problematic behaviours identified above would therefore be prohibited *ex ante*.

In addition, at the French level, the *Autorité* encourages the DGCCRF to pay particular attention to the use of cloud credits in AI, in particular as part of the implementation of the SREN Law.

Lastly, the future EU AI Office and the competent national authority in France, which will be designated in accordance with Article 70 of the AI Act, should ensure, on the one hand, that the implementation of the Act does not hinder the emergence or expansion of smaller operators, and, on the other hand, that the largest operators in the sector do not misuse the text to their advantage.

The *Autorité* also calls for the support of the relevant authorities and for the use of all available tools. The *Autorité* will remain vigilant in the generative AI sector, alongside the DGCCRF, in order to use all their respective tools, if necessary, to act swiftly and effectively.

With regard to access to computing power, the *Autorité*, like many public authorities, supports the **development of public supercomputers**, which are an alternative to cloud providers and give academics, in particular, access to computing power, which is beneficial for innovation. The *Autorité* is also in favour of opening supercomputers to private operators, under certain conditions, for a fee.

With regard to data, public authorities, in particular as part of the mission entrusted by the French Ministry of Culture to the French Higher Council for Literary and Artistic Property, could encourage rights holders to take account of the **economic value of data** according to the use case (for example, by introducing differentiated pricing), and to propose bundled offers to reduce transaction costs, in order to safeguard the innovation capacities of model developers.

Lastly, the *Autorité* calls for **greater transparency** on minority investments in innovative companies, on the basis of Article 14 of the DMA, under which designated companies can be asked for information on their acquisitions.

CONTENTS

INTRODUCTION **15**

I. THE GENERATIVE AI SECTOR **18**

A. Definitions **18**

B. Development of a generative AI model **19**

1. Training phase of a generative AI model ***19***

a) Initial training of the model 19

Extremely high computing power 20

Vast general datasets 20

Model evaluation 21

b) Model fine-tuning 21

2. Inference or content production ***22***

C. The operators in the generative AI value chain **24**

1. The major digital companies in the sector ***24***

a) Alphabet 24

b) Amazon 25

c) Apple 26

d) Meta 26

e) Microsoft 27

f) Nvidia 28

2. Developers of generative AI models ***28***

3. Partnerships between major companies and model developers ***29***

4. The operators upstream in the value chain ***32***

a) IT component suppliers 32

b) Cloud service providers 32

c) Public supercomputers 33

5. The main operators downstream in the value chain ***34***

a) Major technology companies integrating generative AI tools 34

b) Providers of products and services for users, companies and the public sector 34

D. A growing priority for public authorities **35**

1. France's AI strategy ***35***

2. At European level ***36***

a) The EU Artificial Intelligence Act (AI Act) 36

b) Other EU regulations likely to have an impact on AI 38

The Digital Markets Act (DMA) 38

The EU Data Act 38
3. Rules implemented in the rest of the world 39
II. COMPETITIVE ANALYSIS 41
A. A sector characterised by high barriers to entry 41
1. The inputs required to develop foundation models can act as barriers to entry 41
a) The need to use GPUs or other AI-specialised processors 41
b) The cloud: the key to computing power 42
Costly on-site infrastructure 42
The cloud is the preferred solution for training or fine-tuning models and also facilitates downstream deployment 43
c) Model training requires large datasets 43
Large amounts of data of sufficient quality are needed to train generative AI models 44
Generative AI models are mainly trained using public data 44
Access to public data faces uncertainties 46
d) Rare, highly sought-after technical skills 47
2. Significant investment requires agreements between major companies and foundation model developers 48
3. Developments that could limit barriers to entry 49
a) Public supercomputers, an alternative for model training 49
b) Technological innovations that reduce computing power and data requirements 50
c) Open-source models help reduce barriers to entry 51
B. Some companies may benefit from advantages linked to their activities in other digital markets 54
1. Preferential access to the inputs needed to train and develop foundation models 54
a) Easier access to computing power 54
b) Preferential access to data 55
c) The ability to attract highly-skilled employees 58
2. The advantages of vertical and conglomerate integration for major digital companies 58
a) Economies of scale, scope and network effects 58
b) The gradual creation of ecosystems 60
C. Competition risks upstream in the value chain 62
1. Practices that could be sanctioned under competition law 62
a) Introductory remarks 62
b) Risks of abuse upstream in the value chain 64
The risk of abuse by IT component providers 64
The risk of lock-in by major cloud service providers 64

The risks associated with data access 67
The risks associated with access to a skilled workforce 69
The risks associated with open-source models 72
c) The risks associated with the presence of companies on several markets 72
2. Competition concerns about minority investments and partnerships by digital giants 74
a) The need for particular vigilance in the generative AI sector 74
b) Certain investments may be examined *ex ante* under merger control rules 75
Transactions are subject to prior authorisation if they give investors *de facto* control and exceed EU and national notification thresholds 75
An investment may also be examined by competition authorities below the notification thresholds 77
Investments may also be examined as part of the analysis of a merger 77
A lack of transparency regarding investments and partnerships 78
c) Investments may be covered by competition law 79
3. The risk of collusion between companies in the sector 81
III. OUTLOOK AND RECOMMENDATIONS 83
A. The generative AI sector is far from having reached its potential 83
B. Recommendations 85
1. Proposals that aim to make the regulatory framework applicable to the sector more effective, with no change to existing regulation 85
2. Using the tools of competition law and the law on restrictive competitive practices 87
3. Ensuring access to computing power to encourage innovation 89
4. In the data market, ensure a balance between fair remuneration for rights holders and access for model developers to the data needed to innovate, taking into account the diversity of data usage cases 90
5. Greater transparency on investments by digital giants in innovative companies in the sector seems justified 91
CONCLUSION 93
GLOSSARY 95

Introduction

1. Since the public release of the ChatGPT chatbot (created by OpenAI) in November 2022, generative artificial intelligence (hereafter "AI") has taken **centre stage in public and economic debate**. The questions raised by generative AI range from ethics and respect for intellectual property to the impact on the labour market and productivity. The technology offers numerous possibilities to companies in terms, for example, of content creation, graphic design, employee collaboration and customer support. For the French Artificial Intelligence Commission (*Commission de l'intelligence artificielle*) set up in France by the Prime Minister (hereafter the "French AI Commission"), "*generative AI is a major turning point in the history of innovation*". "[The] *characteristics of generative AI* [realism, simplicity, speed, ability] *enable the automation of a number of tasks that were previously difficult to automate. For example, they facilitate the personalization of commercial offers, simplify the analysis of financial data, speed up scientific research, and so on. These same characteristics suggest that AI could take over from personal computers, social networks and smartphones as* the *dominant digital platform, the technological layer on which all other new services are built*"[2].

2. According to a study by the French Treasury[3] (*Direction générale du Trésor*), it is still **too early** to identify a macroeconomic impact of AI on growth. Numerous studies have attempted to estimate the impact of AI on labour productivity. According to some, although the impact of generative AI is uncertain and conditional on technological advances, this innovation alone could increase US labour productivity by almost 1.5 percentage points per year in the 10 years following widespread adoption[4]. Other authors expect a more modest impact, of less than one point of cumulative productivity over 10 years[5]. For specific tasks, initial studies suggest that AI (especially generative AI) could have a positive impact on the individual productivity of certain workers. For example, one study in the customer support profession found an average productivity gain of 14% for customer support agents with access to a conversational assistant, and an even higher gain for less experienced workers[6]. However, higher-skilled jobs may also be impacted by generative AI, and service jobs more so than industrial jobs, thereby distinguishing AI from previous waves of innovation that first affected low-skilled industrial jobs.

[2] French AI Commission, IA : notre ambition pour la France, March 2024. (English translation also available: AI: our ambition for France).

[3] French Treasury, *Trésor-Eco* study: Les enjeux économiques de l'intelligence artificielle, April 2024. (English translation also available: The economic implications of artificial intelligence).

[4] Goldman Sachs (2023), "The potentially large effects of artificial intelligence on economic growth", Global Economics Analyst.

[5] Daron Acemoglu, The simple macroeconomics of artificial economics, MIT, April 2024.

[6] Erik Brynjolfsson & Danielle Li & Lindsey R. Raymond, 2023. "Generative AI at Work," NBER Working Papers 31161, National Bureau of Economic Research, Inc.

3. Against this backdrop, the *Autorité de la concurrence* (hereafter the "*Autorité*") decided on 8 February 2024 to **start inquires *ex officio*** into the competitive functioning of the generative AI sector with a view to issuing an opinion[7], on the basis of Article L. 462-4 of the French Commercial Code (*Code de commerce*).

4. This opinion aims to provide stakeholders with a competitive analysis of the fast-growing generative AI sector, with a particular focus on the strategies implemented by major digital companies aimed at consolidating their market power **upstream in the generative AI value chain (i.e. the design, training and fine-tuning of large language models)** or at leveraging this market power in order to expand in this booming sector. The *Autorité* looks in particular at practices implemented by operators already present in cloud infrastructure and at issues relating to access to cloud infrastructure, computing power, data and a skilled workforce. It also examines **investments and partnerships by major digital companies**, in particular in innovative companies specialised in generative AI. Accordingly, the *Autorité* only incidentally addresses the practices implemented by operators downstream in the value chain (i.e. in contact with the end consumer) and does not touch on the consequences of AI for the competitive functioning of the economy as a whole – an issue of major importance that will merit further analysis in the future.

5. The purpose of such an opinion is not to classify market behaviour under Articles 101 and 102 of the Treaty on the Functioning of the European Union (TFEU) and Articles L. 420-1 and L. 420-2 of the French Commercial Code (*Code de commerce*), but rather to improve understanding of the sector, propose elements for analysis, outline the potential risks from a competition perspective and, where applicable, make recommendations for improving how the sector operates.

6. The *Autorité* launched a public consultation, open from 8 February to 22 March 2024, aimed at deepening its understanding of the sector.

7. The public consultation document invited stakeholders to comment on the resources required to develop foundation models, the competitive landscape and the practices likely to be implemented by operators, as well as minority investments and market prospects. **Around 40 stakeholders and 10 stakeholder associations**, of a variety of sizes and from different sectors, responded to the consultation to express their position and any potential competition concerns.

[7] *Autorité* press release, Generative artificial intelligence: the Autorité starts inquiries *ex officio* and launches a public consultation open until Friday 22 March, 8 February 2024.

8. At the same time, the *Autorité* conducted a large number of interviews, on the basis of Article L. 450-3 of the French Commercial Code (*Code de commerce*). It spoke to French and international private operators (suppliers, customers, associations and others) and institutional parties (government departments, sector-specific regulators, international competition authorities, etc.). In particular, the *Autorité* talked to authorities that have conducted in-depth work on the competition issues raised by the generative AI sector, including:

   – the Portuguese Competition Authority (*Autoridade da Concorrência*), which published a study on the generative AI sector on 5 November 2023[8];

   – the European Commission (hereafter the "Commission"), which launched a call for contributions on generative AI on 9 January 2024[9];

   – the US Federal Trade Commission (FTC), which launched investigations into investments and partnerships in the generative AI sector on 25 January 2024[10];

   – the UK Competition and Markets Authority (CMA), which published an initial report on foundation models on 18 September 2023, followed by an update on 11 April 2024[11].

[8] *Autoridade da Concorrência*, AdC warns of competition risks in the Generative Artificial Intelligence sector, 5 November 2023.

[9] European Commission press release, Commission launches calls for contributions on competition in virtual worlds and generative AI, 9 January 2024.

[10] FTC, FTC Launches Inquiry into Generative AI Investments and Partnerships, 25 January 2024.

[11] CMA, AI Foundation Models: Initial report, 18 September 2023 and AI Foundation Models: Update Paper, 11 April 2024.

I. The generative AI sector

A. DEFINITIONS

9. The European Parliament defines AI as any tool used by a machine "*to display human-like capabilities such as reasoning, learning, planning and creativity*"[12]. This definition includes many tasks that can be automated, such as classification, content recommendation (common on social media), prediction and data generation.

10. The forthcoming EU AI Act defines a "*general-purpose AI model*" as "*an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market*"[13].

11. According to the Act, large generative AI models are a typical example of general-purpose AI models "*given that they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks*"[14]. General-purpose AI models (including generative AI models) are frequently called "**foundation models"** (see glossary).

12. In France, the French AI Commission considers that "*AI is called generative because it can generate new content in the form of text, image, sound, video or code*"[15].

13. Generative AI models differ according to the type of data accepted as input and the data produced as output. The purpose of the models is to produce content – text, images or videos, for example –, generally based on a query. In the context of text generation, reference is frequently made to **large language models** (LLMs, see glossary), but generative AI is not limited to text content.

14. Generative AI models can also be multimodal, capable of combining different types of input and/or output data. For example, image generation models often take text as input and produce images as the output. Other models can accept a combination of text and image as input.

[12] European Parliament, What is artificial intelligence and how is it used?, 4 September 2020.

[13] AI Act, Article 3(63), 16 April 2024 (definition unchanged in 13 June 2024 version).

[14] Above-cited AI Act, recital 99, 16 April 2024.

[15] Above-cited French AI Commission report.

B. DEVELOPMENT OF A GENERATIVE AI MODEL

15. There are two key phases in generative AI modelling. The purpose of the training phase is to teach the model the general capabilities required to produce the content (text, images or other) that is the most likely answer to a given question. It may be followed by fine-tuning (see glossary), during which the model is adapted to a specific task. Lastly, the production of content from the trained model, also known as "inference", involves making the model available to end users. Each phase requires its own computing power and input data.

1. TRAINING PHASE OF A GENERATIVE AI MODEL

16. A generative AI model is first trained to learn general capabilities and can then be adapted to a specific task.

a) Initial training of the model

17. According to the French data protection authority (*Commission nationale de l'informatique et des libertés* – CNIL), "*training is the machine learning process during which the artificial intelligence system builds a model from data*"[16]. This initial training is used to determine the model's parameters, also known as "weights" (see glossary).

18. According to one stakeholder, "[the] *training of the model is based on repeated evaluation of current predictions against target values. The parameters are then adjusted by measuring the results, the aim being to progressively build an increasingly efficient model. The higher the complexity of the tasks to be performed, the more parameters the model requires*".

19. As generative AI models are very complex, they can have several hundred million to several hundred billion parameters. They use techniques such as deep learning (see glossary) and neural networks (see glossary).

20. Most current foundation models for text generation are developed using a deep learning algorithm called "Transformer", introduced in 2017 by a team of Google researchers[17]. This algorithm improved the existing techniques of the time by adding self-attention mechanisms, which allow for better understanding of the sequential nature of certain types of data, notably natural language.

21. Image generation models can use other architectures, such as generative adversarial networks (GANs) or diffusion models. With the rapid evolution of these technologies, however, new training methods and architectures could emerge and replace existing model architectures.

22. Whatever the model and architecture chosen, the initial training phase requires **significant computing power** and **vast general datasets**, often from public sources.

---

[16] CNIL glossary, "training" definition.

[17] Vaswani et al., Attention is all you need, June 2017.

Extremely high computing power

23. Computing encompasses "*all services offering the capacity to process large amounts of information at the same time*"[18]. Generative IA requires significant computing power and the use of special IT hardware that is capable of performing a large number of high-precision operations simultaneously. Graphics processing units (GPUs, see glossary) are frequently used for this task.

24. Originally designed for image display calculations, GPUs have evolved in recent years to optimise AI computing tasks. They are particularly well-suited to AI-related tasks, as they are capable of performing several thousand mathematical operations in parallel, making them more powerful than central processing units (CPUs).

25. Other processors, called "AI accelerators", can also perform these AI-related computing tasks, namely application-specific integrated circuits (ASICs) that are designed and optimised for AI workloads. The term "AI chips" is commonly used to describe GPUs and AI accelerators.

26. There are three main ways for companies to access computing power: cloud computing (the most widely used route, see below), the development of on-site infrastructure and the use of shared computing resources (such as a public supercomputer). Computing power needs depend not only on the number of model parameters but also on the amount of data used to train the model.

27. Model training requires the use of substantial IT hardware that is capable of performing a large number of high-precision operations simultaneously. According to one stakeholder, the order of magnitude of computing power is around "*1,000/2,000 GPUs for a few weeks for state-of-the-art models (around 70 billion parameters)*".

Vast general datasets

28. Training a generative AI model also requires large quantities of data, which must be sufficiently qualitative and diverse in order to avoid the appearance of bias, as any bias in the data can be reflected in the model. For example, one stakeholder said that "[m]*ore than for other types of AI, data is the most critical element for training and developing GPAI* [general-purpose AI] *models, as massive amounts of data are required to train an algorithm*". Data frequently used in foundation model training can be differentiated according to type (text, image, video, etc.) or source (public, proprietary or third-party, see below).

29. The above applies regardless of the type of model trained: language, image or video generation. The differences in how the models are trained concern the data used for the training, as well as the model architecture. Accordingly, a text-image model that produces images from text queries will require appropriate training data to learn how to perform this task, for example in the form of annotated images.

[18] See *Autorité* Opinion 23-A-08, paragraph 28, page 25.

Model evaluation

30. Once trained, the foundation model can be evaluated and then compared with other existing models. Operators use a range of different tests (benchmarks). These tests evaluate a model's capability on a variety of tasks, such as general knowledge of history, law or computing, mathematical problem-solving or scientific reasoning.

31. Beyond their raw performance on the benchmark tests, generative AI models are also pitted against each other on a number of other characteristics, such as the size of the model, represented by the number of weights, its multimodal capabilities to accept as input or produce as output different types of data (text, image, video, etc.) or the context window, which indicates the maximum amount of content that a model can receive in the initial query for the production of an answer.

b) Model fine-tuning

32. After the initial training, the model can be adapted to specific tasks. The aim of the fine-tuning phase, which consists of adjusting the parameters obtained at the end of the initial training, is to improve the model's capabilities for a specific usage without affecting its overall capabilities. According to the CNIL, fine-tuning is a "*technique that consists of adapting a pre-trained AI model to a specific task, which generally involves training the model as a whole, or just certain layers of a neural network, for a small number of iterations on a specific dataset corresponding to the target task*"[19].

33. Fine-tuning can be done by the operator that did the initial model training, or by any other operator with access to the model weights determined during the initial training.

34. Fine-tuning can take several forms, such as:

- **specialisation in a specific sector**. Specific new data is provided to the model in an extension to its training, to improve its efficiency for sector-specific needs. For example, a general text generation model can be specialised on a corpus of legal texts to answer questions requiring in-depth knowledge of the law;
- **reinforcement learning from human feedback** (RLHF, see glossary). RLHF is a "*reinforcement learning approach* [learning from experience] *that uses feedback and evaluations from human users to guide the learning of an AI model*"[20]. RLHF therefore aims to teach the model to produce content that best matches what humans expect, based on human evaluators who rate the content generated by a generative AI model. As indicated by the French Centre of Expertise for Digital Platform Regulation (*Pôle d'expertise et de régulation du numérique* – PEReN) in its report on generative AI[21], thanks to RLHF, "*conversational LLMs are optimised to satisfy human users insofar as possible*", which improves the content of responses.

[19] CNIL glossary, "fine-tuning" definition.

[20] CNIL glossary, "RLHF" definition.

[21] PEReN, Éclairage sur…n°6 – ChatGPT ou la percée des modèles d'IA conversationnels, 6 April 2023. (English translation also available: Shedding light on…n°6 – ChatGPT and the rise of conversational AI models).

35. The fine-tuning phase generally involves a smaller volume of data and therefore requires less computing power than the training phase. According to one stakeholder, "*this phase consumes a moderate amount of computing resources. With the emergence of large general-purpose models, this phase is expected to become the most important in the spread of AI in business and society*".

2. INFERENCE OR CONTENT PRODUCTION

36. A trained – and, where applicable, fine-tuned – generative AI model is then used to generate content. The final stage of content generation is also known as "inference" (see glossary) and involves making the model available to end users.

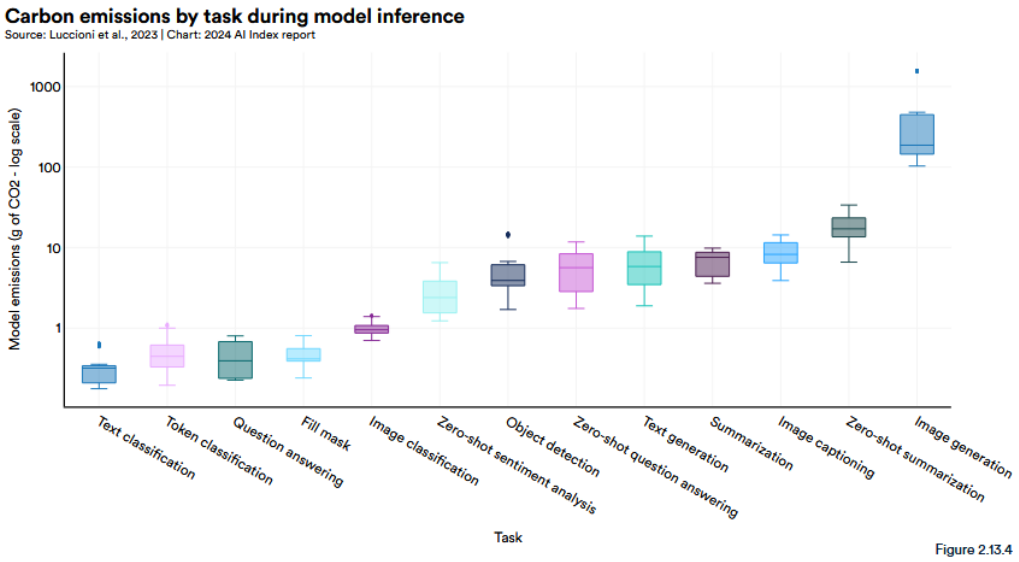
37. There are many ways of deploying generative AI models to end users, generally depending on the degree of openness desired by the developer. Proprietary models can be reserved for internal use or made accessible via specific applications (Internet or mobile, such as OpenAI's ChatGPT or Mistral AI's Le Chat), dialogue windows in applications (e.g. office or collaborative applications), voice assistants or application programming interfaces (API, see glossary) for developers. Users generally do not have access to the model itself, and so cannot reuse or modify the model.

38. Although operators frequently use the term "open source", "open" models are most often made available through the publication of their weights ("open-weights" approach, see glossary), which enables other operators to reuse and/or modify the models, sometimes under certain licensing conditions. All or some of the resources used for model training (code, data, etc.) may also be made available, thus coming close to fully-open models. The challenges raised by the lack of an agreed definition of open source for AI is discussed in paragraphs 179 *et seq*.

39. The computing power required for inference depends on the number of users and the size (number of weights) of the model: for a small model offered to a small number of users, a few GPUs may suffice. However, operating costs can multiply as the size of the model and the number of users increase. According to one stakeholder, "*unlike other major digital innovations, the provision of LLM of generative AI products or services involves meaningful marginal cost, largely because of the cost of compute*".

40. The higher the computing power, the higher the energy consumption. In the inference phase, the carbon footprint varies according to the application used, and is much higher for image than text generation, for example.

Figure 1: Carbon emissions by task during model inference



22 Research also suggests that the reporting of carbon emissions on open model development platforms, such as Hugging Face, is declining over time.

Source: Stanford University, Artificial Intelligence Index Report 2024, page 156.

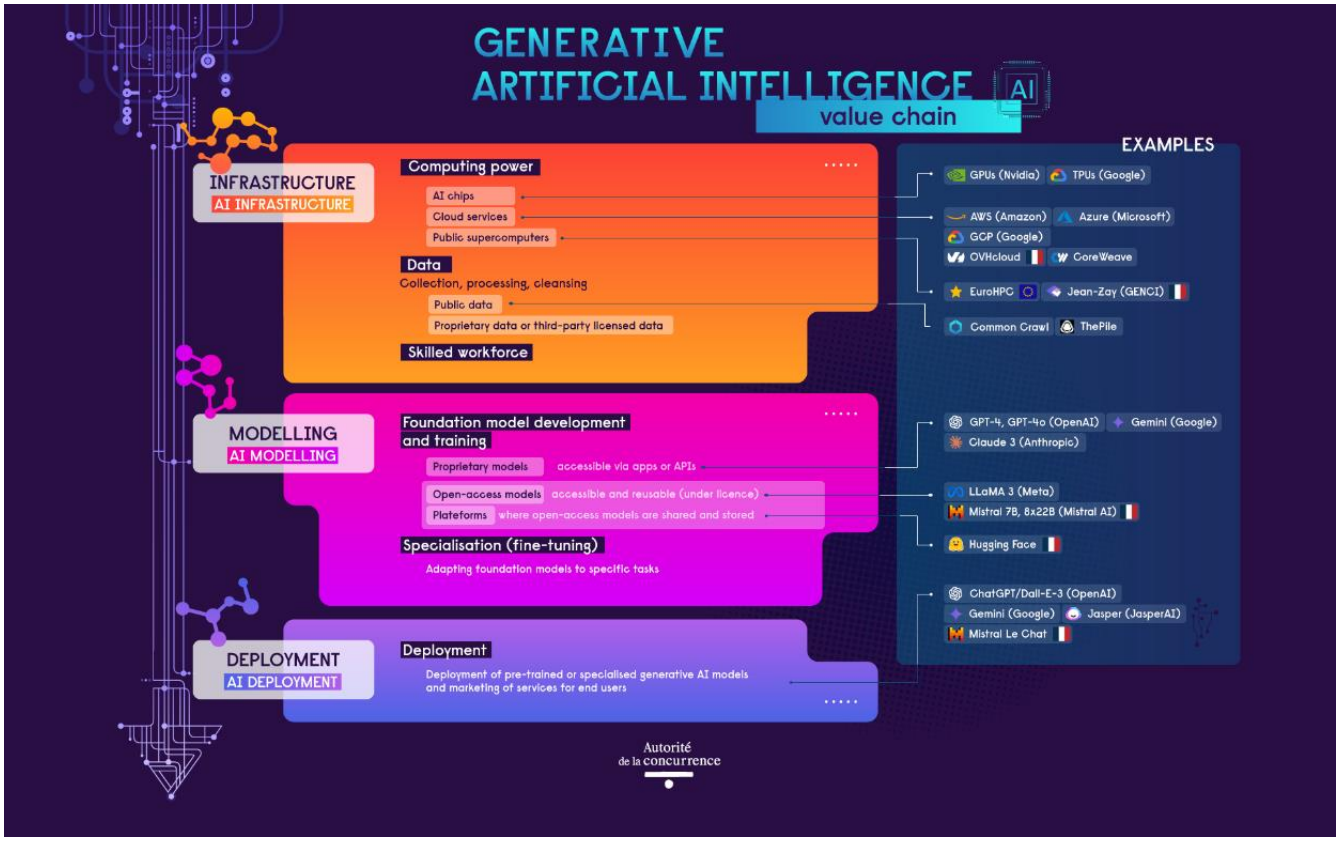
41. In its simplest form, inference requires no additional data beyond that provided in the query, for example the question posed by a user to a chatbot. However, there are techniques for using new data during inference, such as **retrieval-augmented generation** (RAG, see glossary), which aims to improve the result produced by a language model by including reliable external data in the query to add context or provide elements of an answer to the question posed by the user. Specific knowledge bases, such as internal company data, can therefore be used at this stage, providing the model with the most relevant data in addition to the question asked. This technique is used in particular by chatbot-type applications, to provide the robot with knowledge of new data that was not used during initial training, such as news data. For its Gemini application, Google describes a **grounding** stage, which "*consists of Bard* [now Gemini] *sending a query to Google Search to obtain useful information to answer the question posed by the user*"[22].

[22] See *Autorité* Decision 24-D-03, paragraph 166, page 37.

C. THE OPERATORS IN THE GENERATIVE AI VALUE CHAIN

42. As the competitive landscape of the generative AI sector is evolving very rapidly, the presentation of operators below is valid on the date of publication of this opinion. Figure 2 shows the generative AI value chain.

Figure 2: The main operators in the generative AI value chain



Source: Autorité de la concurrence, *inspired by ChatGPT, Bard & Co.: An introduction to AI for competition and regulatory lawyers, Thomas Höppner and Luke Streatfeild, 23 February 2023.*

43. The major digital companies seem to have adopted different strategies for generative AI. Alphabet and Microsoft are present across the entire value chain, thanks in part to partnerships with foundation model developers, while Amazon, Apple, Meta and Nvidia are present at certain layers of the value chain. In addition to these major companies, other operators are present both upstream and downstream.

1. THE MAJOR DIGITAL COMPANIES IN THE SECTOR

a) Alphabet

44. Google is a US company created in 1998, when its founders invented the eponymous search engine, which is the most widely used in France and globally. In 2015, Google combined all its activities into the conglomerate Alphabet, which includes the provision of online search services and operating systems, online advertising and cloud services.

45. Alphabet provides cloud services via Google Cloud Platform (GCP). It has also developed its own **tensor processing units** (TPUs, see glossary), manufactured by Broadcom and which are, according to Google, "*custom-designed AI accelerators, which are optimised for training and inference of large AI models*"[23]. TPUs have been used for Alphabet's internal needs in its data centres since 2015 (although they were only officially announced in 2016), and have been offered and marketed to GCP customers on virtual machines since 2018. While the first generation of TPUs enabled inference only, the fifth version introduced in December 2023[24] enabled the training of the Gemini foundation model, and a sixth, more powerful, version was announced in May 2024[25]. GCP also offers **Model Garden**[26], a platform that provides access to over 130 foundation models, including Google's proprietary generative AI models and other third-party models.

46. In terms of data, Alphabet owns the largest search index for its Google search engine, as well as the world's largest video database with YouTube, which is said to host over 10 billion public videos[27].

47. Alphabet has been active in AI for many years, particularly since its 2014 acquisition of DeepMind, an AI research laboratory founded in 2010. Now called Google DeepMind, it is best known for the launch of AlphaGo (the first AI to beat the world Go champion) and AlphaFold (protein structure prediction). In 2023, it merged with Google Brain, the laboratory that created the TensorFlow development framework (see glossary), widely used in AI. Alphabet's various research laboratories have developed a number of foundation models, such as Bert (2018), Imagen (May 2022), PaLM 2 (May 2023), Gemini (December 2023) and Gemma (February 2024).

48. Alphabet launched its universally accessible Bard chatbot (renamed **Gemini** in February 2024) worldwide in March 2023, and in Europe in July 2023. At its "Google I/O 2024" presentation, it announced the addition of generative AI functionalities to other services, such as its search engine (via AI Overview) and its Workspace office tools, the integration of Gemini into its Android mobile operating system[28], and a Nano version of Gemini in Pixel 8 Pro phones.

b) Amazon

49. Founded in 1994, Amazon is a US company whose main activity is e-commerce via its amazon.com marketplace. It has since diversified into the provision of cloud computing services, via its subsidiary Amazon Web Services (AWS), and connected objects such as Alexa. AWS is one of the leading cloud service providers in France and worldwide[29].

---

[23] Google, Accelerate AI development with Google Cloud TPUs.

[24] Enabling next-generation AI workloads: Announcing TPU v5p and AI Hypercomputer, 7 December 2023.

[25] Google, Introducing Trillium, sixth-generation TPUs, 15 May 2024.

[26] Google Cloud, Model Garden on Vertex AI.

[27] McGrady, R., Zheng, K., Curran, R., Baumgartner, J., & Zuckerman, E. (2023). Dialing for Videos: A Random Sample of YouTube. *Journal of Quantitative Description: Digital Media 3*, 20 December 2023.

[28] Frandroid, La Google I/O 2024 résumée en 15 annonces : Gemini 1.5 Pro, Project Astra, AI Overview, Gmail, Android 15, Veo, etc., 14 May 2024.

[29] See *Autorité* Opinion 23-A-08, paragraphs 91 to 94.

50. AWS has been providing specialised AI chips called "**Trainium"** (for training) and "**Inferentia"** (for inference) to its cloud service customers since 2018. These chips are designed to improve the performance of AI models on AWS while reducing cost and energy consumption[30].

51. Amazon also offers several cloud services specifically for generative AI. For example, **Amazon SageMaker** provides customers with tools to create, train and deploy their own foundation models. Amazon is also positioning itself as a model developer, via its Titan model range, which is accessible via its **Amazon Bedrock** platform where developers can access multiple foundation models developed by Amazon and third parties. Amazon Data Exchange is an AWS cloud service that comprises a number of third-party datasets that Amazon makes available to its cloud users. This service covers many types of data, including financial and sector-specific data.

52. Amazon integrates generative AI systems into several of its products, such as the Alexa voice assistant. Other products have been developed around generative AI, such as Rufus, an AI-powered shopping assistant for its e-commerce platform, and Amazon Q, a chatbot that can write and test code and solicit other cloud services such as AWS S3 for developers on AWS.

c) Apple

53. Apple is a US company founded in 1976 and specialising in the design, manufacture and marketing of electronic products (iPhone, iPad and Mac) and software.

54. Although its activity in the generative AI sector is less developed than that of the other major digital companies, Apple's CEO, Tim Cook, has stated the company's desire to position itself in this market[31]. In the first half of 2024, Apple presented its first models developed in-house, with the large proprietary **MM1** model and a range of small open-source models, OpenELM.

55. With a long history in electronics, Apple has a preferential route to market for its generative AI products. At its Developers Conference on 10 June 2024, Apple announced the launch of generative AI-based features called "Apple Intelligence" in the latest versions of its products (iPhone, iPad and Mac), as well as a partnership with **OpenAI**. In Apple Intelligence, the simplest queries can be processed directly on the device by a small Apple foundation model, while complex queries will be processed in Apple cloud servers (equipped with Apple chips) with the help of its most powerful models and, for certain tasks, OpenAI models[32]. Given the computing power required by generative AI, only iPhones equipped with A17 Pro Bionic chips can support Apple Intelligence.

d) Meta

56. Founded in 2004, Meta (formerly Facebook Inc. until October 2021) is a US company specialising in Internet-related services and products. Meta operates several social networks such as Facebook, Instagram and WhatsApp.

---

[30] Amazon, 2023 Annual Report.

[31] Reuters, Apple to disclose AI plans later this year, CEO Tim Cook says, 28 February 2024.

[32] Apple, Introducing Apple's On-Device and Server Foundation Models, 10 June 2024.

57. Unlike other large vertically integrated US companies (Amazon, Google and Microsoft), Meta does not provide cloud services. However, it has extensive IT infrastructure and **its own data centres** for operating its platforms.

58. Meta develops and designs its own AI accelerators (called "Meta Training and Inference Accelerators [MTIA]") designed to improve the efficiency of workloads such as content recommendation in the news feed, advertising or generative AI. Meta announced the first version of its accelerators in 2023. Given its activity on social networks, Meta also has vast databases of images and videos.

59. Within the Meta group, Facebook AI Research (FAIR) is a consortium of research labs dedicated to fundamental AI research, with the aim of advancing open science. These laboratories have contributed to the development of Meta's language models, the first version of which, **Llama**, was released in February 2023, followed by Llama 2 in July 2023 and Llama 3 in April 2024. These models were all announced and released as open-weights, with a licence allowing for their commercial reuse, except by services with more than 700 million users.

60. In April 2024, Meta also launched **Meta AI**, a chatbot based on the Llama model range. Meta also plans to integrate generative AI tools into its main platforms and, in May 2024, announced the launch of a "sandbox" to test the integration of generative AI into its advertising tools[33].

e) Microsoft

61. Microsoft Corporation (hereafter "Microsoft") is a US company offering a wide range of technology products. Microsoft is a long-standing provider of computer operating systems (Windows) and office software (the Microsoft 365 suite, formerly Office) and also owns the Bing search engine.

62. Microsoft Azure is one of the world's leading cloud service providers. In 2023, it announced the release of AI accelerators called **"Maia"**, which will be available in 2024. These chips have been specially optimised for use in Microsoft Azure cloud infrastructure. Azure also offers a model delivery platform, **Azure AI Model Catalog**.

63. Microsoft has been involved in training LLMs, such as Megatron-Turing-NLG (530 billion parameters) in cooperation with Nvidia in 2021. Following its partnership and investment in OpenAI (see below), Microsoft seems to have made a transition towards the development of small language models (SLMs). In addition to the models developed by OpenAI, Microsoft offers the commercial use of tools based on its proprietary models for image generation (Florence range, including Florence 2 released in November 2023) and text generation (Orca and Phi, including Phi-3 announced and released as open-weights in April 2024).

64. Microsoft integrates generative AI tools into several of its historical products and services. The Bing search engine has offered a Bing AI assistant since 2023. GitHub, a development and code-sharing platform acquired by Microsoft in 2018, has been marketing GitHub Copilot, a development assistant for code generation, since the end of 2021. Microsoft 365 suite software also includes **Copilot** assistants based on generative AI. Most of these tools are based on OpenAI foundation models as well as Microsoft's proprietary models.

[33] Meta, Introducing the AI Sandbox for advertisers and expanding our Meta Advantage Suite, May 12, 2024.

f) Nvidia

65. Nvidia is a US company specialising in the development and design of GPUs and integrated circuits for computing. Founded in 1993, it has been developing a range of specialised GPUs for data centre computing since 2018. Nvidia has seen strong growth since the emergence of AI and, in mid-June 2024, became the world's most valuable publicly-traded company at the date of this opinion, ahead of Microsoft, Apple, Amazon and Alphabet, with a valuation of **over $3.4 trillion** (around €3.14 trillion)[34].

66. Nvidia does not produce GPUs itself but uses the Taiwanese company TSMC. Its current most powerful GPUs are the "**A100"** and **"H100"**. A new generation called **"Blackwell"** was announced in March 2024. Nvidia is also known for its **CUDA** software (see glossary), which enables programming on its own GPUs.

67. Nvidia has extended its presence in the upstream part of the value chain by forging partnerships with a number of cloud service providers, including AWS, GCP and Microsoft Azure. It has contributed to the development of foundation models, in collaboration with Microsoft (Megatron 530B), and has a platform, **Nvidia AI Foundation**, where developers can access its proprietary models and other third-party models. On 14 June 2024, Nvidia also announced the release of a range of models called "Nemotron-4" that can be used to generate synthetic data[35].

2. DEVELOPERS OF GENERATIVE AI MODELS

68. Model developers, which design and train generative AI models, are at the heart of the value chain. Apart from the major companies presented above, developers are mainly AI research labs or innovative AI-native companies.

69. The main developers are as follows:

- **Anthropic** is an AI research laboratory founded in 2021 by former OpenAI members, which aims to offer more secure and responsible AI tools. Anthropic has developed a range of LLMs marketed under the name "Claude". The third version was announced in March 2024;
- **Hugging Face** is a French-US company that provides a hosting and collaboration platform that makes available to developers the vast majority of open-source and open-weights models, as well as shows how those models are reused. As of 5 June 2024, its platform included over 700,000 models and 158,000 datasets[36]. In particular, Hugging Face led the BigScience initiative in 2021 (in collaboration with other players such as the French National Centre for Scientific Research [*Centre national de la recherche scientifique* – CNRS]) to develop a model called "**Bloom**", which was trained on a public supercomputer and is now available to all as open source;

[34] L'Opinion, L'entreprise américaine d'IA Nvidia devient la première capitalisation mondiale en Bourse, 18 June 2024.

[35] Nvidia, NVIDIA Releases Open Synthetic Data Generation Pipeline for Training Large Language Models, 14 June 2024.

[36] Hugging Face website (models and data), consulted on 5 June 2024.

- **Mistral AI** is a French start-up founded in April 2023 and valued at almost €6 billion following its third round of financing in June 2024[37]. It specialises in the development of generative AI models, such as Mistral 7B, Mistral 8x7B and Mixtral 8x22B, which are all open-weights models, and Mistral Large. In February 2024, it also announced the launch of Mistral Large, a proprietary model that in part powers its "Le Chat" chatbot[38];
- **OpenAI** was the first to offer its generative AI model to the general public via the "ChatGPT" chatbot and is the company behind the explosion of interest in this technology. Founded in 2015 as a non-profit association, in 2019 OpenAI created a "capped-profit" branch[39]. The generative AI models powering ChatGPT are GPT3.5, GPT3.5 Turbo and GPT-4[40]. OpenAI is also developing image and video generation models with DALL-E and Sora (announced in February 2024), respectively.

70. Many other operators are developing their own foundation models, such as **Aleph Alpha** (Germany), **Cohere** (United States), **LightOn** (France), **Stability AI** (United States) and **xAI** (United States).

3. PARTNERSHIPS BETWEEN MAJOR COMPANIES AND MODEL DEVELOPERS

71. The generative AI sector is characterised by a large number of agreements between its various stakeholders. These agreements can take several forms, such as agreements to supply computing resources and/or licensing agreements with foundation model suppliers.

72. A number of agreements also include **minority investments** and exclusive or non-exclusive commercial agreements for the development or marketing of foundation models within their platforms.

73. The main agreements, classified by investment amount, are as follows:

- an exclusive partnership between Microsoft and **OpenAI** in the form of a "*multi-year, multi-billion dollar investment* [by Microsoft] *to accelerate AI breakthroughs*" in January 2023[41]. Microsoft provides OpenAI with high-performance computing capabilities to accelerate OpenAI's research, and deploys OpenAI's models in its products, notably via the Azure cloud. Microsoft has also become OpenAI's exclusive cloud service provider;

---

[37] Le Monde Informatique, En levant 600 M€, Mistral AI frère les 6 Md€ de valorisation, 11 June 2024.

[38] Mistral AI, Le Chat, 26 February 2024

[39] In a blog post dated 11 March 2019, Open AI said: "[w]*e want to increase our ability to raise capital while still serving our mission, and no pre-existing legal structure we know of strikes the right balance. Our solution is to create OpenAI LP as a hybrid of a for-profit and nonprofit – which we are calling a 'capped-profit' company. The fundamental idea of OpenAI LP is that investors and employees can get a capped return if we succeed at our mission, which allows us to raise investment capital and attract employees with startup-like equity. But any returns beyond that amount – and if we are successful, we expect to generate orders of magnitude more value than we'd owe to people who invest in or work at OpenAI LP – are owned by the original OpenAI Nonprofit entity*".

[40] Generative pre-trained transformers (GPTs).

[41] Microsoft blog post, Microsoft and OpenAI extend partnership, 23 January 2023. The investment was initiated in 2019 when Microsoft invested $1 billion in OpenAI.

– investments by Amazon ($4 billion, or around €3.7 billion, for a non-exclusive "*minority ownership position*"[42]) and Google ($2 billion, or around €1.8 billion, for an estimated 10% stake, according to some articles[43]) in **Anthropic** in 2023[44]. In particular, these partnerships give Anthropic access to AI chips and customers from AWS and Google;

– the investment of $1.3 billion (around €1.2 billion) in **Inflection**, an innovative US company aiming to offer personalised AI services, in June 2023 by companies such as Microsoft and Nvidia[45]. In March 2024, two of Inflection's co-founders, and some of the staff, left the company to head up the newly-created Microsoft AI division, which will be in charge of Copilot, Bing and Edge[46];

– investments by major companies such as Amazon, Google and Nvidia of $235 million (around €220 million) in **Hugging Face** in August 2023, to accelerate the training, fine-tuning and deployment of large models used to create generative AI applications. On 25 January 2024, Hugging Face also announced a "*strategic partnership*" with Google Cloud[47];

– the €15 million investment by Microsoft in **Mistral AI** in 2024 through a bond convertible into shares and a partnership[48] enabling Mistral AI to access the Azure AI supercomputer infrastructure and offer its premium models in the Azure AI Studio and Azure Machine Learning model catalogue as a Model-as-a-Service (MaaS). Mistral Large, its latest text generation model, will also be available to Microsoft customers. In addition, companies such as Nvidia and Salesforce participated in its latest fundraising round in June 2024.

[42] Amazon press release, Amazon and Anthropic Announce Strategic Collaboration to Advance Generative AI, 25 September 2023.

[43] Le Monde informatique, Après Amazon, Google va investir jusqu'à 2 milliards de dollars dans Anthropic, 30 October 2023.

[44] TechCrunch, AI's proxy war heats up as Google reportedly backs Anthropic with $2B, 27 October 2023.

[45] TechCrunch, Inflection lands $1.3B investment to build more 'personal' AI, 29 June 2023.

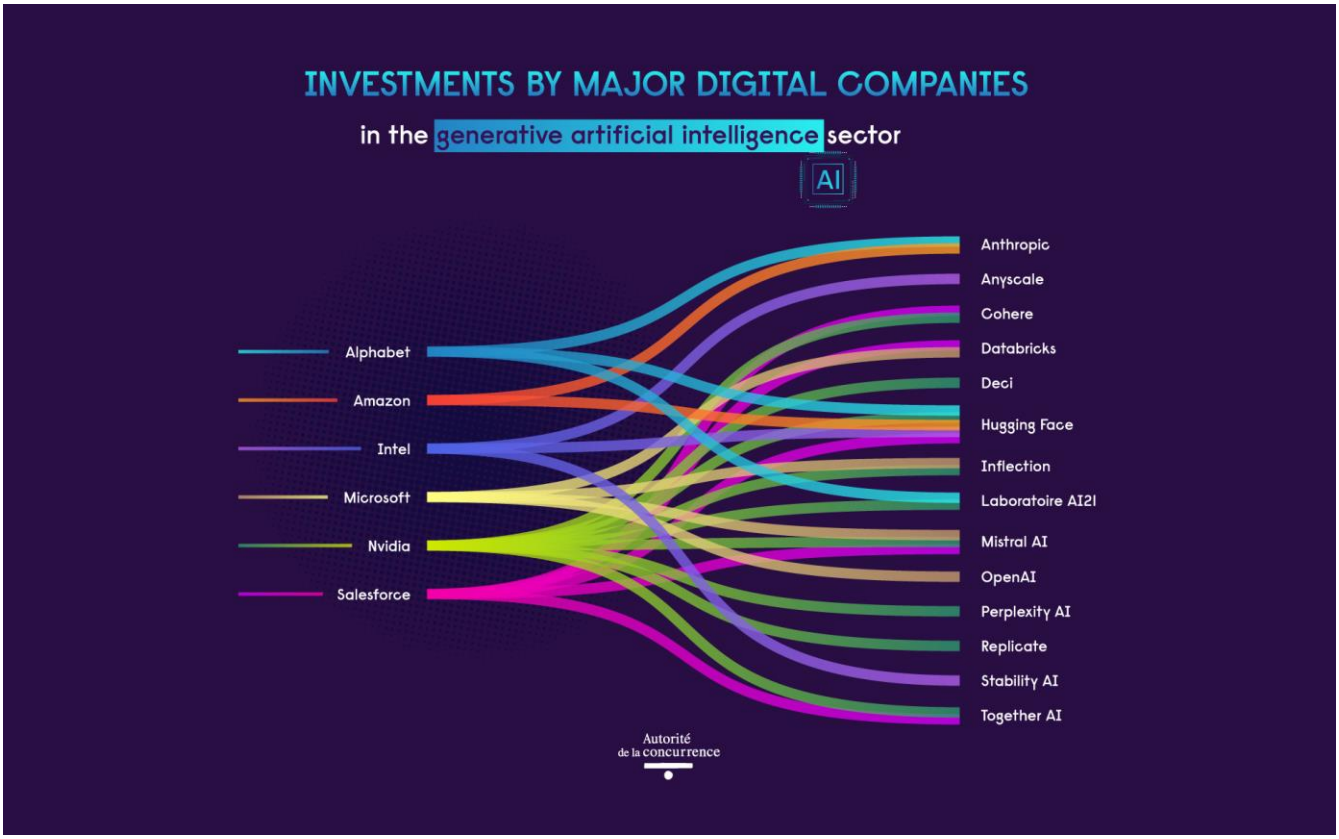
[46] Inflection press release, The new Inflection: an important change to how we'll work, 19 March 2024

[47] Hugging Face press release, Hugging Face and Google partner for open AI collaboration. The partnership will enable Google Cloud customers to "*easily train and deploy Hugging Face models within Google Kubernetes Engine (GKE) and Vertex AI. Customers will benefit from the unique hardware capabilities available in Google Cloud, like TPU instances, A3 VMs, powered by NVIDIA H100 Tensor Core GPUs, and C3 VMs, powered by Intel Sapphire Rapid CPUs*".

[48] Microsoft, Microsoft and Mistral AI announce new partnership to accelerate AI innovation and introduce Mistral Large first on Azure, 26 February 2024.

74. The following diagram illustrates the many cross-investments between major digital companies and innovative companies in the sector, notably by Nvidia, which has invested in a number of companies in the generative AI sector, and in particular model developers (Mistral AI, Inflection, Deci, Hugging Face, AI21 Labs, etc.).

Figure 3: Investments by major digital companies in innovative companies in the sector (May 2024)



Source: Autorité de la concurrence, *inspired by S&P Global, Untangling the web of strategic tech investments in generative AI, 22 February 2024*[49].

[49] The diagram, which is not intended to be exhaustive, is based on a list of the main generative AI operators in France provided by a major stakeholder. The list of investments is also based on research conducted by the *Autorité* and may omit investments that have not been publicly disclosed. Lastly, some innovative companies operating in France do not appear to have received investment from major digital companies (but may receive funding from institutional investors, for example) and are therefore not included in the diagram. This is the case, for example, of model developers such as Eleven Labs, LightOn and Naver.

4. THE OPERATORS UPSTREAM IN THE VALUE CHAIN

75. Other operators are present upstream in the generative AI value chain, for example IT component suppliers, cloud service providers and public supercomputers.

a) IT component suppliers

76. IT components such as GPUs are essential for training generative AI models. In addition to Nvidia and major digital companies (presented above), the sector includes a number of incumbent operators and new entrants.

77. Nvidia's main historical competitors are **Advanced Micro Devices (AMD)** and **Intel**.

78. AMD is a US company founded in 1969. It has been active in the production of chips for graphics cards since the acquisition of ATI Technologies in 2006. The Instinct MI300X graphics processors are AMD's latest GPU release[50], ahead of the MI325X scheduled for late 2024.

79. Intel is a US company founded in 1968, specialising in the development and production of motherboards and chips (known as "x86") for computers. It is also active in GPUs and AI accelerators, including the Gaudi 3 AI accelerator unveiled in April 2024[51].

80. Several other companies, such as Cerebras (United States), Graphcore (United Kingdom), SambaNova (United States) and Groq (United States), are gradually moving into the sector, offering specialised chips for AI.

b) Cloud service providers

81. In addition to the three major cloud service providers (Microsoft Azure, AWS and GCP) presented above, many other companies such as 3DSOutscale, Alibaba Cloud, IBM, Oracle Cloud, OVHcloud and Scaleway provide cloud services that can be used both upstream for model training and fine-tuning, and downstream for generative AI model inference. These cloud operators are described in detail in paragraphs 103 to 117 of Opinion 23-A-08, to which this opinion refers.

82. Other specialised operators are also emerging to meet the sector's specific computing resource needs. An example is US supplier **CoreWeave**, which specialises in the provision of high-performance computing services and, according to the company, is Nvidia's "*elite partner*"[52]. Its last fundraising round in May 2024 raised over $1 billion (around €930 million), increasing its valuation to $19 billion (around €17.5 billion)[53].

83. Other operators such as Denvr Dataworks (Canada), Lambda Labs (United States) and TensorWave (United States) are positioned in the segment for the provision of specialised AI cloud services, often in partnership with IT component operators such as Nvidia, AMD or Intel.

---

[50] L'Usine Digitale, AI générative : Les profits d'AMD tirés sur le haut par son accélérateur Instinct MI300X, 31 January 2024.

[51] L'Usine Digitale, Intel dévoile Gaudi 3, sa dernière arme pour se lancer dans la bataille de l'IA générative, 10 April 2024.

[52] CoreWeave, CoreWeave Becomes NVIDIA's First Elite Cloud Services Provider for Compute.

[53] TechCrunch, CoreWeave's $1.1B raise shows the market for alternative clouds is booming, 5 May 2024.

c) Public supercomputers

84. The French Atomic Energy Commission (*Commissariat à l'énergie atomique* – CEA) defines a supercomputer as "*a very large computer, combining several tens of thousands of processors and capable of performing a very large number of simultaneous computing or data processing operations*"[54]. Supercomputers have traditionally been used for fundamental research, and tasks such as weather and climate forecasting, as well as simulations in materials science, chemistry and medicine. Their aim is to provide computing resources to researchers free of charge.

85. There are many public supercomputers in the world. The TOP500 ranking[55], drawn up by a team of researchers from Lawrence Berkeley National Laboratory and the Universities of Tennessee and Mannheim, ranks supercomputers according to several performance criteria, including computing power and environmental performance. According to the June 2024 ranking, of the ten most powerful supercomputers, five are located in the United States (including one owned by Microsoft Azure and another by Nvidia), four in Europe (in Finland, Switzerland, Italy and Spain) and one in Japan.

86. Within the European Union, eight supercomputers have been installed (in addition to the three above-mentioned supercomputers in Finland, Italy and Spain, five others are located in Luxembourg, Portugal, the Czech Republic, Bulgaria and Slovenia) thanks to the European High Performance Computing (EuroHPC) joint undertaking between public and private partners. A new exaflop supercomputer, i.e. capable of performing one billion billion computing operations per second, is currently under construction in Germany.

87. Another exaflop supercomputer called **"Jules-Verne"** has been announced by EuroHPC in France for 2025. Jules-Verne will be the most powerful supercomputer in France and will supplement the existing range of three supercomputers: **Jean Zay** (at the CNRS' Institute for Development and Resources in Intensive Scientific Computing [*Institut de développement et des ressources en information scientifique* – IDRIS] in Orsay, ranked 190th in the TOP500), **Adastra** (located at the French National Computer Centre for Higher Education [*Centre informatique national de l'enseignement supérieur* – CINES] in Montpellier, 20th in the TOP500) and **Joliot-Curie** (at the High Performance Computing Centre at the Atomic Energy Commission [*Très grand centre de calcul du Commissariat à l'énergie atomique* – TGCC-CEA]) in Bruyères-le-Châtel, 132nd in the TOP500). These supercomputers are managed by the French National High-Performance Computing Agency (*Grand équipement national de calcul intensif* – GENCI)[56] and the partner research organisations (CNRS, CEA, etc.) that host these data centres.

---

[54] CEA, L'essentiel sur les supercalculateurs, 7 March 2022.

[55] TOP500 ranking, June 2024.

[56] GENCI is a non-trading company created in 2007. It is 49% owned by the French State via the Ministry of Higher Education and Research and by the CEA (20%), the CNRS (20%), France Universités (10%) and the French National Institute for Research in Digital Science and Technology (*Institut national de recherche en sciences et technologies du numérique* – INRIA) (1%).

5. THE MAIN OPERATORS DOWNSTREAM IN THE VALUE CHAIN

88. Downstream, operators offer generative AI-based products and services to end users or integrate these products or services into companies' workflows.

a) Major technology companies integrating generative AI tools

89. In addition to the above-mentioned major digital companies, other operators in the technology sector have begun to integrate these new tools into their existing products and services. For example:

- Adobe enables the use of generative AI features in its Photoshop tool, using its own proprietary Firefly model and others such as DALL-E;
- in January 2024, Samsung launched its Galaxy S24 smartphone range, which includes a number of generative AI tools (real-time translation, photo editing, instant search, etc.);
- Zoom Workplace has integrated generative AI tools (Zoom AI Companion) to offer automated reports of calls and meetings made through its platform.

b) Providers of products and services for users, companies and the public sector

90. Generative AI tools can be used by companies, developers and the general public.

91. Most model developers (see above) also offer an Internet interface for testing and using their generative AI products, most often in the form of a chatbot (for text or image models) accessible free of charge or by paying to access more advanced features. The generative AI applications for the general public include **OpenAI's ChatGPT, Google's Gemini** and **Mistral AI's Le Chat** for text generation, and **MidJourney** and **StableDiffusion** for image generation.

92. Globally, many companies offer applications, tools or platforms based on generative AI, leading one stakeholder to say that "*the deployment market is much more competitive, as many companies are now using foundation models to design their own specialised systems, either by fine-tuning or simply by optimising prompting or using RAG*".

93. Several hundred of them offer applications for companies in a wide range of activities (such as sales, marketing, human resources, finance and legal) and in all sectors, including banking, insurance, healthcare, transport, agriculture and industry[57]. These applications cover a wide range of modalities, from text, image and video generation to computer code generation. In France alone, over 130 start-ups offer tools in different categories, including design, productivity, customer relations, sales, healthcare, cybersecurity, and knowledge management[58].

94. Several types of operators also offer facilitator services on behalf of their customers. These operators include digital services companies (DSCs) such as Accenture, Atos and Capgemini, along with start-ups such as Dust and AleIA in France.

---

[57] FirstMark, Machine Learning, AI and Data landscape, April 2024.

[58] Wavestone, 2023 Radar of French "GenAI" Startups, January 2024.

D. A GROWING PRIORITY FOR PUBLIC AUTHORITIES

95. Although the generative AI sector has developed relatively recently, the governments of France, Europe and the rest of the world have been quick to mobilise.

1. FRANCE'S AI STRATEGY

96. Following the report "*Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne*"[59] (Giving meaning to artificial intelligence: towards a national and European strategy), in 2018 the French government launched a national AI strategy aimed at equipping France with competitive research capabilities and deploying AI technologies throughout the economy. The launch of the "France 2030" plan in October 2021 also aimed at developing industrial competitiveness and the technologies of the future.

97. The first phase of the national strategy (2018-2022) involved strengthening France's research capabilities by promoting the creation and development of a network of interdisciplinary AI institutes, support for chairs of excellence in AI, funding for doctoral programmes, and investment in public research computing capabilities (Jean Zay supercomputer).

98. The aim of the second phase, launched in 2022, is to deploy AI in the economy with three main levers: training and research, support for a state-of-the-art offering, and matching AI supply and demand. As part of the second phase, the French President announced nine winners of the "AI-clusters" call for expressions of interest, and new investment support schemes[60].

99. A first call for projects, "*Communs numériques*[61] *pour l'intelligence artificielle generative*" (Digital commons for generative artificial intelligence), launched in 2023, aimed to accelerate the creation and accessibility of digital commons across the entire generative AI value chain, provided the incentive nature of public support was justified and proven, and to develop innovative products and services.

100. A second call for projects, "*Accélérer l'usage de l'IA générative dans l'économie*" (Accelerating the use of generative AI in the economy), open until 2 July 2024, focuses on the downstream part of the value chain, encouraging the development of integrated generative AI solutions with an advanced level of functionality and a short-term adoption horizon. Through the second call for projects, the French government aims to support 500 small and medium-sized enterprises and mid-sized establishments in the adoption and use of AI solutions by 2025.

---

[59] Report by C. Villani, Donner un sens à l'intelligence artificielle: pour une stratégie nationale et européenne, submitted to the Prime Minister on 28 March 2018.

[60] Directorate General for Enterprise (DGE), Annonce de 9 nouveaux lauréats pour l'appel à manifestation d'intérêt « IA-clusters », 22 May 2024.

[61] According to the "Communs numériques pour l'intelligence artificielle générative" call for projects, a "digital commons" is a resource produced or maintained collectively by a community of players, and governed by rules that ensure its collective and shared nature. Digital commons can, for example, include learning and test databases that enhance national data.

101. Furthermore, the French government set up the aforementioned **French AI Commission** on 19 September 2023. Chaired by Mr Y... and Ms Z..., the AI Commission includes representatives from various sectors (cultural, economic, technological, research) to "*help make France a country at the forefront of the AI revolution*". The report, published on 14 March 2024, made 25 recommendations with a total estimated cost of €27 billion.

102. In particular, the report recommends:

- the launch of an awareness-raising and training plan to meet current and future needs;
- massive investment in digital companies and corporate transformation to support the French AI ecosystem and make France a world leader;
- the creation of a major computing power cluster in France;
- easier access to quality data (personal data and data protected by literary or artistic property rights);
- improved conditions for high-level public AI research in France;
- the establishment of global AI governance, including monitoring of the development of market concentrations and the rapid introduction of the regulations necessary to prevent abuses of dominant position.

103. In April 2024, the Prime Minister also announced the creation of an AI service called "Albert", which will simplify administrative procedures for citizens and automate certain tasks such as complaint transcription[62].

2. AT EUROPEAN LEVEL

a) The EU Artificial Intelligence Act (AI Act)

104. On 21 April 2021, the Commission published a proposal for a regulation on AI[63]. Following approval by the European Parliament and then the Council of the European Union on 21 May 2024, the AI Act is due to be published soon.

105. According to the latest publicly-available versions[64], the Act will apply to both public and private sector entities, based inside and outside the European Union, when the "*AI system*"[65] is placed on the EU market or its use has an impact on people located in the European Union. The Act establishes obligations for AI systems based on their potential risks. Systems that

---

[62] Les Échos, Qu'est-ce qu'Albert, l'intelligence artificielle française déployée par le gouvernement ?, 24 April 2024.

[63] European Commission proposal for a regulation establishing harmonised rules on artificial intelligence, 21 April 2021.

[64] Text version dated 13 June 2024 available at the following link: https://data.consilium.europa.eu/doc/document/PE-24-2024-REV-1/en/pdf.

[65] An AI system is defined in Article 3 as follows: "*a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*".

pose unacceptable risks are banned if they threaten citizens' rights (such as exploiting people's vulnerability), while "*high-risk*" systems are subject to strict obligations[66].

106. Specific obligations are also imposed on providers of "*general-purpose AI models*", including large-scale generative AI models (Article 53). In particular, they must prepare and keep up-to-date technical documentation on the model, including its training and testing process and the results of its evaluation, which must be provided on request to the AI Office (the new centre of expertise for AI, tasked with implementing the powers given to the Commission by the Act, promoting the European AI ecosystem, and collaborating with the competent authorities of the Member States as part of the governance provided for by the Act) and to the competent national authorities. They must make information and documentation available to AI system suppliers that intend to integrate their general-purpose AI models into their AI systems. They must also put in place a policy to comply with EU copyright legislation and publish detailed summaries of the content used for their training. These obligations do not apply to models published under a free and open-source licence and whose parameters are made publicly available, unless they present a systemic risk[67].

107. Several governance bodies have been set up at European level, such as an AI Office, which will develop the European Union's expertise and capabilities in AI, and a Committee on Artificial Intelligence, made up of one representative per Member State, for coordination purposes in particular. Each Member State will establish or designate as competent national authorities at least one notifying authority[68] and at least one market surveillance authority.

108. The Act will come into force 20 days after its publication in the Official Journal of the European Union and will be fully applicable 24 months after its entry into force, i.e. *a priori* during 2026, with the exception of certain provisions such as the classification rules for high-risk systems (which will be applicable 36 months after its entry into force).

---

[66] This classification depends on the function performed by the AI system, as well as the specific purpose for and way in which the system is used. Annex III of the Act refers, for example, to systems used in the field of "*law enforcement, in so far as their use is permitted under relevant Union or national law*", such as profiling.

[67] According to Article 3(65) of the draft Act, systemic risk is defined as "*a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain*". According to Article 51 of the draft Act, a general-purpose AI model presents a systemic risk if the model meets two conditions: having high-impact capabilities according to a technical methodology ("*shall be presumed to have high impact capabilities* [...] *when the cumulative amount of computation used for its training measured in floating point operations is greater than* 10^{25}") or a Commission decision.

[68] According to Article 28 of the draft Act, "*each Member State shall designate or establish at least one notifying authority responsible for setting up and carrying out the necessary procedures for the assessment, designation and notification of conformity assessment bodies and for their monitoring*".

b) Other EU regulations likely to have an impact on AI

The Digital Markets Act (DMA)

109. The **Digital Markets Act (DMA)**[69] was adopted on 14 September 2022 to regulate the practices of digital giants. The Act sets out certain obligations that could, subject to an assessment by the Commission, apply in the AI sector if the core platform services designated by the Commission, such as search engines, social networks and voice assistants, incorporate AI services[70]. The obligations include the following[71]:

- a ban on combining or cross-using personal data from a core platform service with data from any further service, unless users give their consent (Article 5(2));
- a ban on using non-publicly-available data, including data generated by business users (Article 6(2));
- an obligation for gatekeepers to ensure the effective and free portability of the data provided by the end user or generated through the activity of the end user in the context of the use of the relevant core platform service (Article 6(9));
- an obligation to provide business users, free of charge, with continuous and real-time access to the data provided for or generated in the context of the use of the relevant core platform service, including personal data (Article 6(10));
- an obligation to share ranking, query, click and view data with third-party search engines (Article 6(11)).

The EU Data Act

110. The EU Data Act[72] aims to remove barriers to data access for both private and public sector bodies, while preserving incentives to invest in data production by ensuring the balanced control of data for its creators.

111. The Act establishes rules for making data from connected products more accessible, in order to foster the development of a truly competitive and fair data economy that benefits both users and European companies. It also aims to prevent abusive exploitation of contractual imbalances with regard to data access and use, such as terms enabling a party that has unilaterally imposed a term to access and use the data of the other contracting party in a manner that is significantly detrimental to its intellectual property rights in particular (Article 13(5)(b)). It also gives public authorities the right to access data held by companies in exceptional situations in the public interest. Lastly, it aims to remove the main barriers to the use of competing cloud services by, for example, phasing out switching charges (Article 29) and introducing measures to make switching easier from a technical point of view (Article 30).

---

[69] 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 (published on 12 October 2022).

[70] The obligations of the DMA can only be applied to the core platform services mentioned in Article 2 of the DMA, which do not include MaaS. See developments below and *Autorité* proposal 1.

[71] The obligation to inform the Commission of any proposed merger (Article 14) is discussed below in paragraphs 394 *et seq*.

[72] Regulation (EU) 2023/2854 of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

112. In France, Law 2024-449 of 21 May 2024 to secure and regulate the digital space (hereafter the "SREN Law") is intended to anticipate certain provisions of the Act relating to the cloud sector (see paragraph 252 below). It was adopted on 21 May 2024.

3. RULES IMPLEMENTED IN THE REST OF THE WORLD

113. Outside France and Europe, initiatives are also being implemented in the rest of the world.

114. Several international organisations have adopted common principles. On 30 October 2023, G7 leaders adopted an agreement on international AI guiding principles and a voluntary code of conduct, as part of the Hiroshima Process[73]. The Italian presidency of the G7, whose priorities include AI, adopted a ministerial declaration on 15 March 2024 aimed at advancing work towards safe and trustworthy AI[74]. Other international initiatives have emerged, such as the Framework Convention on AI recently adopted by the Council of Europe[75].

115. In November 2023, the first AI Safety Summit, hosted by the United Kingdom, adopted the "Bletchley Declaration"[76], signed by 28 countries and the European Union, to foster a common understanding of the technological risks posed by AI and develop international cooperation on the safety of these systems. After Korea, **the next summit will be held in France on 10 and 11 February 2025**[77].

116. At national level, a series of proactive measures are being implemented.

117. After promoting an approach based on voluntary commitments by AI operators[78], the Biden administration issued an executive order on AI regulation on 30 October 2023[79]. On 27 April 2024, the US government established a Federal AI Council to provide recommendations to ensure the safe adoption of AI in the United States[80]. The Council is made up of the heads of the biggest companies in the sector (such as OpenAI, Microsoft, Google and Nvidia), government representatives and researchers.

---

[73] G7, Hiroshima Process International Code of Conduct for Advanced AI Systems, 30 October 2023. The Code of Conduct contains 11 non-binding recommendations to promote "*safe, secure and trustworthy AI*", including the most advanced foundation models and generative AI systems.

[74] G7 Ministerial Declaration of 14-15 March 2024.

[75] On 17 May 2024, the Council of Europe adopted the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law. The Convention, signed by 46 countries including the EU Member States, as well as others such as the United States, Canada and Japan, aims to establish rules for respecting fundamental rights, in the face of risks of discrimination or invasion of privacy when using AI technologies.

[76] The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023.

[77] Le Monde, Emmanuel Macron veut faire de la France « un des pays champions de l'IA », 22 May 2024.

[78] Contexte, À Washington, le gratin de l'IA promet de s'autoréguler, 24 July 2023.

[79] The White House, FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, 30 October 2023. It aims in particular to (i) put in place a series of standards to ensure the creation of safe and secure AI tools before they are released to the public (ii) call on Congress to pass legislation on the protection of personal data, (iii) tackle issues of algorithmic discrimination, (iv) promote innovation and competition, under the authority of the Federal Trade Commission in particular, and (v) strengthen cooperation on AI at international level.

[80] Le Monde, Intelligence artificielle : création d'un conseil fédéral pour aider le gouvernement américain, 27 April 2024.

118. In February 2024, the United Kingdom clarified its approach to AI regulation, entrusting the regulation of AI systems to sector-specific regulators, which will rely in particular on the five non-statutory principles laid down by the UK government to guide their action: system safety, system transparency, system compatibility with existing laws, system accountability, and system contestability. It also pledged to develop binding rules in the longer term for "*the small number of companies developing general-purpose AI systems*", although no such measures have yet been announced[81].

119. In August 2023, the Cyberspace Administration of China issued "provisional measures" designed to impose obligations on providers of generative AI, among others. Providers are obliged to perform safety assessments and submit reports on their tools to the authorities, particularly if these tools are likely to influence public opinion[82]. In addition, several measures on artificial intelligence are currently being drawn up, primarily to promote industrial development[83]. During the visit of the President of the People's Republic of China, it was announced that France will participate in the World AI Conference and the High-Level Meeting on Global AI Governance to be held in China in 2024[84].

120. As far as competition authorities are concerned, a great deal of work has been conducted or is underway on the competition issues raised by the generative AI sector. In addition to the above-mentioned authorities (see paragraph 8 above), the Canadian[85], Indian[86] and Hungarian[87] authorities have also announced the launch of studies into the sector.

[81] French Treasury, Brèves numériques Royaume-Uni, 14 December 2023 to 7 February 2024.

[82] La Tribune, IA générative : la Chine instaure de nouvelles réglementations, 18 August 2023.

[83] Forbes, China's New Draft AI Law Prioritizes Industry Development, 22 March 2024.

[84] Déclaration conjointe entre la République française et la République populaire de Chine sur l'intelligence artificielle et la gouvernance des enjeux globaux, 6 May 2024.

[85] Competition Bureau Canada, Competition Bureau seeks feedback on artificial intelligence and competition, 20 March 2024.

[86] The Telegraph, Competition Commission of India to undertake market study on AI, 6 March 2024.

[87] Hungarian Competition Authority launches study on impact of AI on competition and consumers, 17 January 2024.

II. Competitive analysis

121. The generative AI sector is characterised by high barriers to entry (A), which are likely to favour major digital companies that also enjoy advantages linked to their activities in other digital markets (B). Despite the sector's recent development, competition risks may arise, particularly upstream in the value chain (C).

A. A SECTOR CHARACTERISED BY HIGH BARRIERS TO ENTRY

122. As indicated above, generative AI requires three key inputs, the access to or ownership of which constitute barriers to entry (1). They also require significant investment (2). However, certain innovations could limit these barriers to entry (3).

1. THE INPUTS REQUIRED TO DEVELOP FOUNDATION MODELS CAN ACT AS BARRIERS TO ENTRY

123. The training and inference of generative AI models require essential inputs, such as computing power, data and a skilled workforce.

a) The need to use GPUs or other AI-specialised processors

124. As mentioned above, the training of foundation models for generative AI requires specific hardware. The chips used must be capable of performing a large number of operations in parallel, and require a high degree of precision to accurately determine several billion parameters. The chips most commonly used are GPUs produced primarily by Nvidia, although several major digital companies are developing their own AI accelerators.

125. Through its public consultation, the *Autorité* learned that GPUs and other AI accelerators cannot easily be replaced by CPUs, given their performance. As several stakeholders explained, GPUs provide greater computing power and capacity than CPUs. The AI sector is characterised by a race to train ever more powerful models, with only GPUs capable of enabling sufficiently rapid training that is compatible with the speed at which the market is developing. Another stakeholder considered that "*the learning time required to develop generative AI services would be multiplied by 1,000 without access to GPUs*".

126. Given the explosion in demand for AI-specific computing over the past two years, GPU users are experiencing supply difficulties. Several operators confirmed that creating an infrastructure with sufficient computing power is particularly costly and difficult due to the shortages caused by high demand and a limited supply of semiconductors.

127. In addition to hardware, the creation of foundation models also requires a software layer to run the IT code directly on the graphics card, in order to distribute the computing as evenly as possible on the graphics card and thus optimise its performance.

128. The proprietary CUDA software environment developed by Nvidia, which is exclusive to its own chips, is the most widely used by operators in the sector. One stakeholder confirmed that "*CUDA, a low-level matrix computing framework used almost systematically by the above-mentioned higher-level deep learning libraries, is predominant in LLM training. It*

was developed by Nvidia and is exclusively associated with its GPUs. Its GPU competitor AMD has developed a similar framework (ROCm), but its use appears to be very marginal. More broadly, there is fundamental interplay between the hardware [...] *and the off-the-shelf software used to perform AI(G)* [generative AI]*: they form ecosystems whose maturity and community size are likely to lead to mergers and inertia*".

b) The cloud: the key to computing power

Costly on-site infrastructure

129. At present, most companies in the sector do not manage their own on-site infrastructure due to the high costs involved. According to one stakeholder, "[t]*he cost of acquiring this type of hardware is not affordable for all companies. Nvidia's H100 GPUs, currently the benchmark, cost between €30,000 and €40,000 each. While one GPU may suffice for inference, training a foundation model requires several thousand of these processors (with the largest models using several tens of thousands of GPUs)*".

130. As a result, only a few companies, such as Meta or Samsung, can achieve the computing power required for the various stages in the development of generative AI models with on-site infrastructure. In addition to the initial investment costs, operators with their own infrastructure are responsible for server operation, maintenance and upgrades. According to one stakeholder, "[i]*n addition to the cost of acquiring GPUs, there are significant operating (electricity, cooling) and implementation (in particular, the space required) costs. The investments required to enter and then develop on this market are therefore considerable. And it's highly uncertain that a company would be able to recoup its investment simply by designing a foundation model for its own needs*".

131. These constraints limit the scope for a new operator in generative AI to develop its own in-house infrastructure to train its own models. According to one stakeholder, "[i]*n practice, only a small number of companies have sufficient infrastructure to achieve the computing power required for AI development*". Another said that "[a] *company specialising in LLM development will not be developing its own infrastructure*".

132. According to one stakeholder, "*as GPU research is extremely dynamic, this hardware may be affected by 'obsolescence' in the near future. In these conditions, it may make sense to use computing power 'as a service'*". Rapid hardware obsolescence is characteristic of the AI market, particularly for generative AI, and reinforces the constraints associated with the use of on-site infrastructure.

133. However, the computing power required for fine-tuning is much lower than that required for training, due to the much smaller volume of data required for this phase.

134. Similarly, for the inference phase, while the computing power needs depend on the demands placed on the model (e.g. the number of users), they do not require GPUs that are as powerful (and therefore as expensive) as those used for training.

135. Accordingly, some major companies that already have on-site computing infrastructure as part of their business could upgrade this infrastructure to enable the fine-tuning or inference of generative AI models in-house.

The cloud is the preferred solution for training or fine-tuning models and also facilitates downstream deployment

136. The cloud gives companies access to on-demand IT infrastructure that is scalable and can be adapted to their needs. In addition to infrastructure, it also facilitates user access to a wide range of managed services (PaaS and SaaS).

137. According to one stakeholder, "*from a developer's point of view, there is no fundamental technical difference between the computing resources provided by cloud providers and by on-premises infrastructure. FM* [foundation model] *developers are often 'digital natives' that prefer to use scalable, cost-effective cloud infrastructure rather than incur significant costs by investing in on-premises infrastructure*".

138. In aforementioned Opinion 23-A-08[88], the *Autorité* identified the advantages and disadvantages of using cloud infrastructure. The same advantages and disadvantages apply to generative AI model training, as confirmed by stakeholders. The cloud therefore offers the advantage of avoiding initial investment and maintenance costs and allows for pay-as-you-go pricing based solely on business needs. It also offers flexibility and rapid access to the most advanced technologies.

139. In the AI sector, cloud service providers (CSPs) play a dual role, both upstream, for the training or fine-tuning of models, and downstream, where they are the preferred platform for the deployment of foundation models to companies. For example, one stakeholder said that "[f]*or more traditional companies wanting to fine-tune pre-trained models for their business, the 'off-the-shelf' services of CSPs are very useful, as companies can start with pre-trained models (a sort of marketplace for AI models) and then use the computing power needed for fine-tuning as and when required, at reasonable rates".*

140. MaaS services such as Model Garden on Google Cloud, Amazon Bedrock on AWS and Azure AI on Microsoft Azure appear to be **major points of contact between model developers and user companies**, which are more often than not already customers of the cloud service provider. These services simplify access to AI models for companies, often via an API, and make it easier for them to deploy generative AI-based applications.

141. It is in the interests of model developers to make their models available on as many cloud service providers as possible in order to maximise the number of potential customers, thereby reinforcing the role of cloud service providers as essential gateways for both training and inference.

142. Accordingly, generative AI is present at every layer of the cloud. IaaS (computing, storage, etc.) and PaaS (vector databases, AI tools, etc.) services enable model training and fine-tuning, while more and more SaaS services include generative AI tools. Developers can also offer their generative AI models or services via cloud marketplaces.

c) Model training requires large datasets

143. In the current state of generative AI technology based on LLMs, data is essential for the training and fine-tuning of models, as well as for inference when techniques such as RAG or grounding are used. Based on the data used as input, the model can learn to generate content.

[88] See *Autorité* Opinion 23-A-08, paragraphs 18 to 22, pages 21 and 22.

Large amounts of data of sufficient quality are needed to train generative AI models

144. Data volume is crucial for generative AI model training. In 2020, a research paper published by engineers at OpenAI estimated that LLM performance increased with the volume of data provided to train the model, triggering a race for model size[89]. Stakeholders confirmed the importance of "*large quantities of heterogeneous data such as text, image, audio and video, depending on the type of content the model is intended to generate*".

145. Model developers provide little information about the data used for training. As an example, Meta reported that over 15,000 billion tokens (see glossary) were used to train its Llama 3 model, seven times more than its predecessor Llama 2, released less than a year earlier for the same number of parameters[90].

146. The contributions to the public consultation also stressed data quality, particularly for future models. For example, one stakeholder said that it "*expects the success of future open-source and proprietary models to depend more on the quality or relevance of the data to the task in hand, or the greater performance of the algorithms, than on the use of larger volumes of data*".

147. Data quality is mainly the result of data cleansing and processing, in particular to exclude poor-quality data. These processing operations are a necessary pre-training step, as well as a differentiating factor between operators. For example, one stakeholder stated that "[it] *is rare for foundation model designers to extensively disclose the detailed composition of their databases, since this is a major competitive advantage*", while another confirmed that this initial stage of data cleansing and scoring can take several months.

148. Stakeholders differentiate between the data used for training and the data used for fine-tuning. One stated that "*the data used for pre-training is open-source, general and extensive data* [...]. *Its purpose is to train the language model in the acquisition of general knowledge.* [...] *the data used for fine-tuning is generally either open-source, recent data that has not been used during pre-training, or internal company data (proprietary data that is more precise, specialised and technical but often of lesser quantity)*". During inference, techniques such as RAG may require the use of data relevant to the use case, which may be internal company data or news data. The use of other types of data, such as synthetic data, is discussed below.

Generative AI models are mainly trained using public data

149. According to the stakeholder contributions to the public consultation, the majority of generative AI models are mainly trained using publicly-accessible databases. This data, often described by operators as "*public*", contains public datasets or content accessible on the Internet, even though some may be protected, for example by copyright. In this respect, one stakeholder commented that "*data from these publicly-available sources typically makes up the majority of the data used to train FMs. It's often supplemented by a smaller amount of proprietary and third-party data*".

[89] Kaplan et al, Scaling Laws for Neural Language Models, January 2020.

[90] Meta press release, Introducing Meta Llama 3: The most capable openly available to date, 18 April 2024.

150. Many datasets are freely available on the Internet and are frequently used by developers to train foundation models. The best-known and most widely-used include:

- Common Crawl, a non-profit organisation founded in the United States in 2007. Its mission is to provide free Internet archives. Since 2008, numerous internet crawls (see glossary) have been performed, and the latest, dated May 2024, included almost 3 billion Internet pages, or just under 400 terabytes of data;
- C4 (Colossal Cleaned Crawled Corpus), a filtered version of Common Crawl, in English only (mC4 is the multilingual version) and published by Google;
- LAION-5B (Large-scale Artificial Intelligence Open Network), a dataset published in 2022 and containing almost 6 billion image-text pairs with descriptions in English and other languages. LAION is a non-profit organisation that provides data and models to encourage AI research;
- The Pile, an open-source language dataset consisting of 22 data subsets, such as books, scientific publications, blogs, etc., published by EleutherAI, a non-profit AI research group founded in 2020.

151. In addition to the above historical datasets, numerous other datasets are frequently published, which can help to reduce some of the barriers to accessing data for training generative AI models. For example:

- Common Corpus, an open-source multilingual dataset containing over 500 billion words[91];
- FineWeb, a dataset of 15,000 billion tokens from Common Crawl published by Hugging Face in May 2024;
- YouTube Commons, a dataset published in April 2024 containing transcriptions of over two million YouTube videos.

152. In addition to the above individual datasets, several platforms, such as **GitHub** and **Hugging Face**, provide access to a large number of datasets. Hugging Face offers over 158,000 datasets of all types on its platform, including some of the datasets presented above.

153. At European level, a number of **sector-specific data** space initiatives have been launched, such as CatenaX for the automotive sector, AgDataHub for agriculture and agrifood, and EonaX for air transport. These data spaces can be used, for example, to adapt a generative AI model to a specific sector, rather than relying on **proprietary** or **third-party** data.

154. For fine-tuning, while stakeholder responses confirmed that non-public proprietary data is primarily used, there are also examples of public fine-tuning datasets such as **OpenOrca**, a dataset published in 2023 and containing instructions to enable models to answer user questions.

[91] Common Corpus and YouTube Commons are published by Pleias, a French start-up supported by the French Ministry of Culture and the French Interministerial Digital Department (DINUM).

Access to public data faces uncertainties

155. The lack of publicly-accessible data of sufficient quality could force operators to use a greater volume of proprietary data to train generative AI models.

156. First, model developers are concerned about the legal issues involved in using databases such as Common Crawl. Many content providers are now asserting their rights, which is having the consequence of reducing access and/or preventing data collection by indexing robots. Some press publishers, such as the New York Times, have taken **legal action** against Microsoft and OpenAI on copyright grounds[92], while in France, the Society of Authors in the Graphic and Plastic Arts (*Société des auteurs dans les arts graphiques et plastiques* – ADAGP) has put together a practical guide for publishers on how to opt out of the use of content for AI search purposes[93]. According to one stakeholder, "[i]*t's no longer possible to do a 'complete' scraping of the Internet as OpenAI did to train GPT-3. It's impossible for copyright reasons in particular, with all the lawsuits surrounding OpenAI and Google on the issue*".

157. Second, some community platforms such as Reddit and Twitter increased the price of their APIs, often used by developers to collect data, in 2023, to better value their proprietary data, used in particular to train generative AI models[94].

158. In view of these risks, some model developers are forging **partnerships** with publishers and rights holders. For example, Google has signed agreements with Reddit[95] and StackExchange[96]. OpenAI has signed agreements with a number of content providers and press publishers in several countries, including Associated Press in the United States[97] and Le Monde in France[98]. The table below lists all the data use agreements announced by OpenAI to date. Observers reported that OpenAI has approached several media groups with offers ranging from $1 million to $5 million a year[99], but the deal with News Corp would be for a much higher amount, close to $250 million (over €230 million) over five years (i.e. $50 million a year)[100].

---

[92] Le Monde, Le « New York Times » poursuit en justice Microsoft et OpenAI, créateur de ChatGPT, pour violation de droits d'auteur, 27 December 2023.

[93] ADAGP, L'ADAGP prend des mesures pour protéger ses membres face à la menace des intelligences artificielles génératives, 23 February2024.

[94] Forbes, Death By API: Reddit Joins Twitter In Pricing Out Apps, 1 June 2023.

[95] Le Figaro, IA : Reddit noue un accord de licence inédit avec Google pour 60 millions de dollars, 22 February 2024.

[96] Stack Overflow blog, Stack Overflow and Google Cloud Announce Strategic Partnership to Bring Generative AI to Millions of Developers, 29 February 2024.

[97] Associated Press, AP, Open AI agree to share select news content and technology in new collaboration, 13 July 2023.

[98] Le Monde, Intelligence artificielle : un accord de partenariat entre « Le Monde » et OpenAI, 13 March 2024.

[99] The Verge, OpenAI's news publisher deals reportedly top out at $5 million a year, 4 January 2024.

[100] Les Echos, « Méga accord » entre OpenAI et News Corp, 23 May 2024.

Table: Agreements between OpenAI and content providers (as of 20 June 2024)

Content provider	Country	Agreement date
The Atlantic	United States	29/05/2024
Vox Media Inc.	United States	29/05/2024
News Corp	United Kingdom	22/05/2024
Reddit	United States	16/05/2024
Dotdach Meredith	United States	07/05/2024
Financial Times	United Kingdom	29/04/2024
Le Monde	France	13/03/2024
Prisma Media	Spain	13/03/2024
Axel Springer	Germany	13/12/2023
Associated Press (AP)	United States	13/07/2023

Source: Autorité de la concurrence *and OpenAI announcements*

d) Rare, highly sought-after technical skills

159. Technical skills are another resource needed to design foundation models. Model developers must have advanced skills in data science, machine learning and deep learning, skills in natural language processing (NLP, see glossary) or computer vision, and knowledge of engineering and development and operations (DevOps), in order to be able to both develop the code and set up the architecture to run this code efficiently. In addition to their theoretical training, engineers must have worked with neural networks and, more specifically, transformers.

160. The training of a foundation model relies on a number of hyperparameters (see glossary), which can only be mastered and optimised through empirical expertise, with each having a cross-impact on model performance.

161. For example, one stakeholder mentioned "*the ability to remain at the cutting edge of fundamental research (new model architectures, new breakthrough innovations in learning paradigms) and more applied research (minor optimisations)*" as an essential skill to excel in this sector at the forefront of innovation. Another, however, said that "[a]*n increasing number of data scientists and engineers have been attracted to this dynamic sector*" given the high level of interest in the sector.

162. On the other hand, the development of language models does not require very large teams. As a result, many start-ups have been built around a small number of highly-skilled employees and have subsequently developed models, like Mistral AI, which announced its first model in September, just a few months after it was founded, even though it had just 22 employees in December 2023[101].

[101] Le Monde Informatique, Mistral lève 385 M€ et devient une licorne française, 11 December 2023.

2. SIGNIFICANT INVESTMENT REQUIRES AGREEMENTS BETWEEN MAJOR COMPANIES AND FOUNDATION MODEL DEVELOPERS

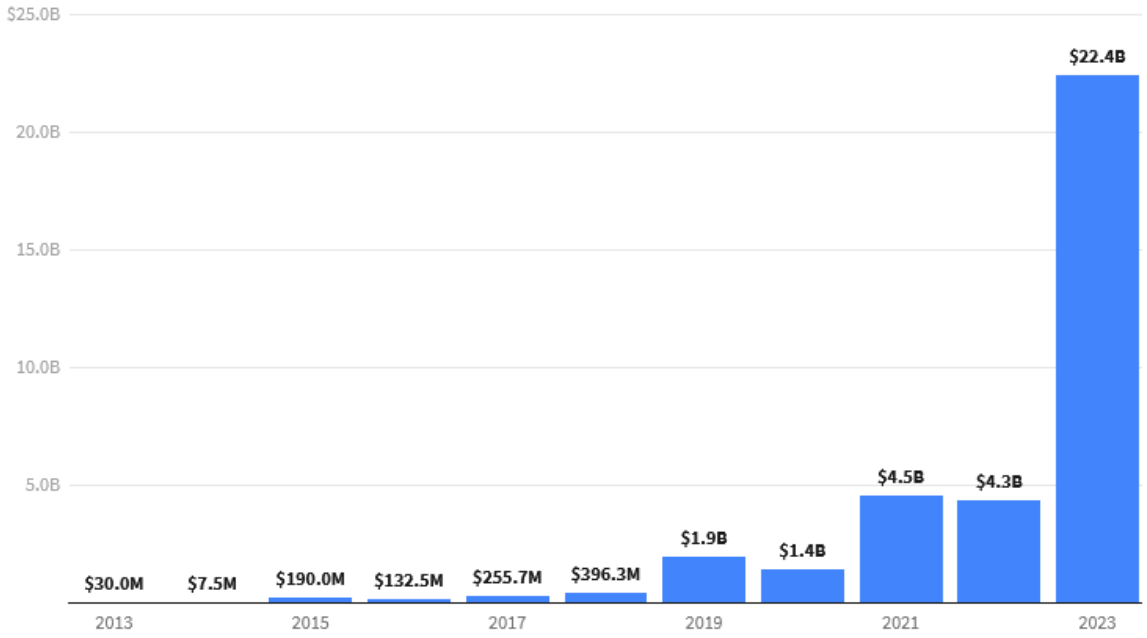
163. As described above, the development of foundation models requires substantial investment in computing power, very large datasets and access to highly-specialised skills.

164. The scale of the investment required creates significant barriers to entry. According to estimates, training OpenAI's GPT-3 foundation model alone cost over $4 million (around €3.7 million), while the GPT-4 model that followed cost over $78 million (around €72 million)[102].

165. The challenge is even greater because repeated investment is required. According to several stakeholders, foundation model developers need to invest continuously to improve their models and bring improved versions to market.

166. Consequently, **investments in the sector increased almost six-fold between 2022 and 2023**. Companies in the sector raised over $22 billion in 2023 (around **€20 billion**), compared with around $4 billion in 2022 (around €3.7 billion). Over 70% of the funds raised go to foundation model developers. However, the European Court of Auditors has highlighted the low level of private investment at European level (compared with other players, the United States and China) and has criticised the lack of governance and coordination of public investment in AI[103].

Figure 4: Investment in the generative AI sector



Source: Dealroom, Generative AI, January 2024.

[102] Stanford University, Artificial Intelligence Index Report 2024, page 64.

[103] European Court of Auditors, Special Report 08/2024: EU Artificial intelligence ambition – Stronger governance and increased, more focused investment essential going forward, 29 May 2024.

167. The increase in AI investment in 2023 is all the more remarkable given that global start-up investment fell over the same period by 38% compared with 2022, a decline observed at all funding levels. In addition to AI, the semiconductor and battery sectors also saw increased investment in 2023[104].

168. In conjunction with these investments, the sector is seeing **a number of partnership agreements between major digital companies and generative AI model developers** (see paragraphs 71 *et seq.* above). As indicated by Amazon when announcing its collaboration with Hugging Face, "*building, training, and deploying large language and vision models is an expensive and time-consuming process that requires deep expertise in machine learning (ML). Since the models are very complex and can contain hundreds of billions of parameters, generative AI is largely out of reach for many developers*"[105].

3. DEVELOPMENTS THAT COULD LIMIT BARRIERS TO ENTRY

a) Public supercomputers, an alternative for model training

169. Traditionally focused on high-performance computing (HPC), public supercomputers have started to transition in recent years to host more AI research projects. As a result of the transition, hardware has evolved to accommodate GPUs in addition to the CPUs historically used for HPC. For example, the Jean Zay supercomputer, launched in France in 2020, has since undergone several hardware upgrades and extensions to add GPUs and enable more specialised AI projects. A new extension, scheduled for June 2024, will boost its computing power to over 125 petaFLOPS (125 million billion computing operations per second), thanks to the addition of 1,456 Nvidia H100 GPUs[106].

170. However, the computing power of public supercomputers remains well below that of the supercomputers used to train the largest models, such as GPT-4 and Llama 3, which have tens of thousands of GPUs[107]. The computing power of public supercomputers must also be shared between a large number of AI projects. However, public supercomputers offer users the advantage of benefitting from technical support, for example to optimise IT code.

171. In return for contributing to open science (for example, publishing work in an academic journal), access to public supercomputers is free, which can help to reduce the barriers to entry for operators, in particular in the research world. For example, a team of researchers from the CentraleSupélec university has trained a bilingual French-English language model called "**CroissantLLM**" on the Jean Zay supercomputer[108]. At European level, a call for applications was launched in March 2024 to give public and private operators access to the computing power of EuroHPC supercomputers (MareNostrum, Leonardo and others)[109].

---

[104] Crunchbase, Global Startup Funding In 2023 Clocks In At Lowest Level In 5 Years, 4 January 2024.

[105] Amazon, AWS and Hugging Face collaborate to make generative AI more accessible and cost efficient, 21 February 2023.

[106] IDRIS press release, Jean Zay Extension: 1456 H100 GPUs to attain 125 PFlop/s, 28 March 2024.

[107] See, for example, the blog post announcing the release of Llama 3: "*We performed training runs on two custom-built 24K GPU clusters*".

[108] L'Usine Digitale, CroissantLLM : Des chercheurs de CentraleSupélec lancent un modèle d'IA open source et bilingue, 4 March 2024.

[109] EuroHPC, EuroHPC JU Access Call for AI and Data-Intensive Applications, 5 March 2024.

172. Public resources such as supercomputers can be used to train and fine-tune models. However, supercomputers are not a solution for inference, due in particular to the constraints associated with access to computing power, which is often granted for a limited time window (weeks or months). For example, one stakeholder said that "*many French start-ups that use* [a supercomputer] *for learning are turning to private operators (including AWS, which benefits greatly) to offer inference and commercial services, in the absence of an integrated sovereign offer*".

b) Technological innovations that reduce computing power and data requirements

173. In a market as dynamic as that of generative AI, **technological innovations** are continually emerging, enabling the development of simpler models that use less data and therefore have more limited computing power. These innovations lower the barriers to entry, by reducing model training costs, dependence on particularly large volumes of data, and inference costs during use.

174. In terms of architecture, several state-of-the-art models such as Mixtral 8x22B are based on **Mixture of Experts** (MoE, see glossary). This architecture is divided into several subsets of sparse neural networks called experts, specialised in a specific task, and uses a router to determine which expert should be used to answer a query[110]. This specific type of architecture **improves training efficiency**, as it will be more efficient than a conventional architecture for a given computing budget. It also significantly reduces inference costs. Instead of using all the weights in the model, only some of the experts and therefore weights (e.g. two experts out of eight) are used to answer a query. For example, Mistral AI highlights the efficiency of its Mixtral 8x22B model, "*that uses only 39B active parameters out of 141B, offering unparalleled cost efficiency for its size*"[111].

175. As with model architecture, innovations may emerge to **improve the efficiency of fine-tuning techniques** and reduce their cost. For example, in 2021 several Microsoft researchers presented a technique called "Low Rank Adaptation (LoRA)"[112], which reduces computing power needs for fine-tuning.

176. Computing power needs can also be reduced by developing smaller models. These can provide answers to more specialised queries that require less computing power.

177. In the short term, however, the possibility of these more economical models replacing the current large-scale models should be put into perspective. First, while these models are less expensive, they are currently less efficient than larger models, leading some operators in the sector to reserve them for specific uses (such as document retrieval) and to question their economic profitability. According to the Artificial Intelligence Index Report published by Stanford University, the use of computing power for the main AI models has increased exponentially, especially over the last five years. As an example, the report compares Google's "Transformer" model, released in 2017, which required 7,400 petaFLOPS, with Google's "Gemini Ultra" model, released at the end of 2023, which required 50 billion petaFLOPS[113].

---

[110] Hugging Face, Mixture of Experts explained, 11 December 2023.

[111] Mistral, Cheaper, Better, Faster, Stronger, 17 April 2024.

[112] Hu, Shen et al., LoRA: Low-Rank Adaptation of Large Language Models, June 2021.

[113] Stanford University, Artificial Intelligence Index Report 2024, page 51.

178. Many stakeholders also mentioned the possibility of using **synthetic data** (see glossary), for example data generated by another foundation model, to train generative AI models. New types of neural networks, such as GANs and variational autoencoders (VAEs), can be used to generate content that resembles the input data. Synthetic data can be less costly to acquire. For example, Stanford University developed a specialised version of the Llama model, called "Alpaca-7B", in March 2023, using synthetic data from ChatGPT, for a data generation cost of less than $600 (€555)[114]. The use of synthetic data also reduces the constraints associated with personal data. However, it entails certain risks, such as the propagation of bias and higher error rates[115].

c) Open-source models help reduce barriers to entry

179. In the IT sector, the *Autorité* defines open-source software as "*software in which the source code is available to the general public. The development of this 'free software' involves a collaborative effort in which programmers improve the source code together and share changes within a community*"[116]. Modifications can also be made to existing software.

180. In the context of AI, the Open Source Initiative (OSI) – a non-profit organisation dedicated to defending the principles of open source – has launched a consultation on a specific definition of open source principles as applied to AI[117]. The consultation focuses on the transparency and availability criteria to be met by an AI model in order to meet open source criteria.

181. In practice, in the generative AI sector, open source covers a wide range of scenarios, from open-weights models where only the weights are made public to fully-open models where all the code, architecture, training data, weights and learning process are made available. The lack of an objective definition of open source in the context of generative AI creates a risk of confusion for users, and even misleading communication on the part of developers (sometimes referred to as "open-washing")[118].

[114] Stanford University, Alpaca: A Strong, Replicable Instruction-Following Model, 13 March 2023.

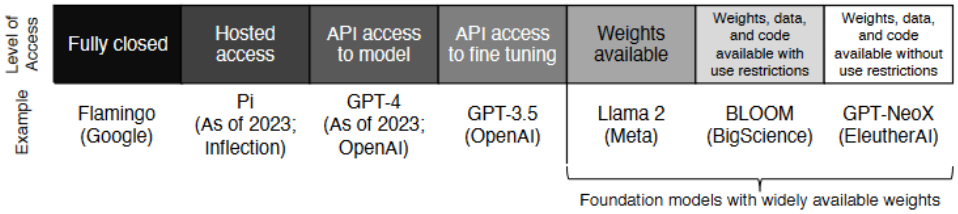
[115] United Nations University, The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development, 4 September 2023.

[116] See *Autorité* Opinion 14-A-18, paragraph 19, page 9.

[117] Open Source Initiative, The Open Source AI Definition – draft v. 0.0.8, accessed on 18 June 2024.

[118] Liesenfeld, A., & Dingemanse, M. Rethinking open source generative AI: open-washing and the EU AI Act. In the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). ACM.

Figure 5: Publication modes for generative AI models

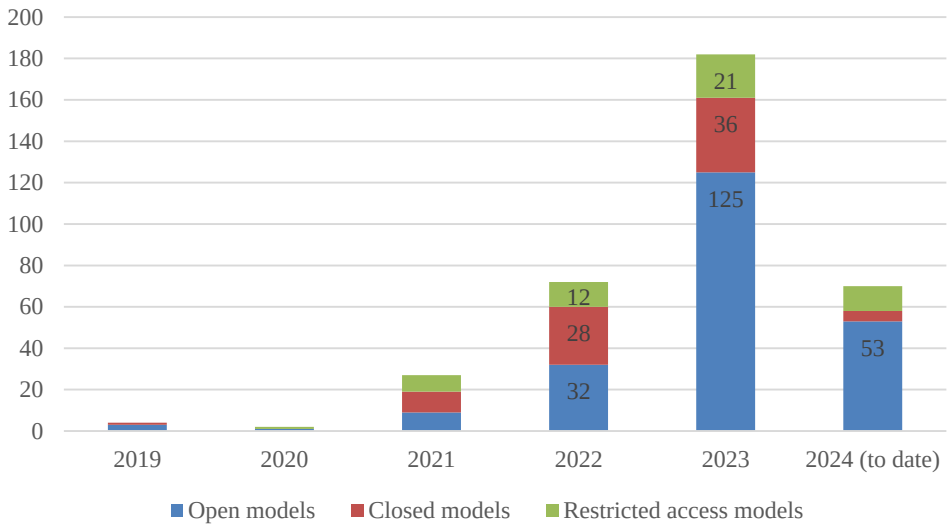


Source: Graph published by the Stanford Institute in December 2023, itself taken from Irene Solaiman's paper, The Gradient of Generative AI Release: Methods and Considerations, 2023.

182. The *Autorité* observes that among model developers, non-commercial operators generally publish all their work in open source, while commercial operators publish some of their models as open-weights but keep their most successful models as proprietary modes. For example, Mistral AI has released several of its models as open-weights, but not Mistral Large, its most powerful model. Meta, for its part, released its Llama range of models as open-weights, but with licensing restrictions on commercial use in applications with more than 700 million users.

183. The big vertically integrated companies like Google and Microsoft are also tending to make some of their smaller language models available as open-weights, with Gamma (Google) and Phi (Microsoft). Figure 6 shows the sharp increase in the number of generative AI models released by operators, as well as their propensity to turn increasingly towards the open-source deployment of their models, mainly via an open-weights strategy.

Figure 6: Types of generative AI foundation models released

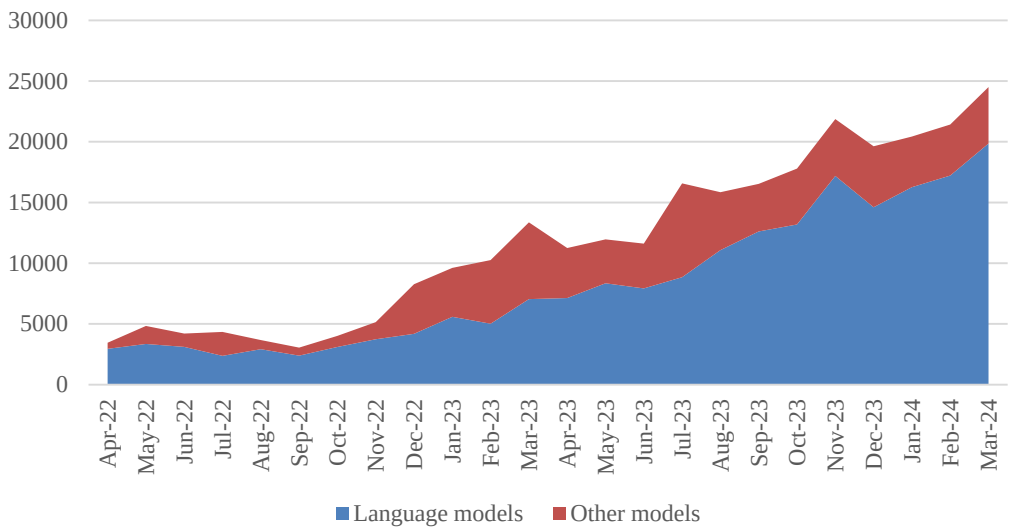


Source: Ecosystem Graphs: The Social Footprint of Foundation Models, Stanford CRFM, data consulted on GitHub on 7 June 2024.

184. The existence of so many open-source models means that a greater number of operators can enter the downstream part of the generative AI value chain, in particular the many operators (companies, administrations and researchers, for example) that **do not have the resources to develop their own generative AI foundation models and for which fine-tuning existing models is therefore easier**.

185. Thus, the fine-tuning of open-source foundation models contributes to the development and furthering of research while reducing barriers to entry downstream. While foundation models only number in the hundreds, **the number of new fine-tuned models published on the Hugging Face collaboration platform each month is in the tens of thousands**. As of 5 June, the total number of models (of all types) on the platform exceeded 700,000[119].

Figure 7: Number of new AI models published each month on Hugging Face



Data obtained from the OECD.AI (2024) website based on Hugging Face data and the OECD counting methodology.

186. However, while the deployment of open-source models and technologies can reduce barriers to entry by giving a greater number of operators access to the technologies, it does not remove the barriers for an operator wanting to develop its own foundation model or obtain sufficient computing power to fine-tune a generative AI model. As one stakeholder said, while the strategy applied by the majority of generative AI operators, consisting of making the weights of a model available, "*enables widespread reuse, it does not, on the other hand, reduce the barrier to entry for new operators wanting to train foundation models. Knowledge of these weights is of only marginal use for training foundation models, the purpose of which is to create new models and their own weights*". In order to reproduce an AI model, transparency about other elements would be necessary, such as the code and data for training or the data used.

187. In addition, the publication of high-performance generative AI models can raise security issues. Unlike models that are accessible via applications or APIs, which generally come with security filters, open-source models can be reused by malicious players to produce problematic content (such as child pornography or disinformation, for example).

[119] Hugging Face website, consulted on 5 June 2024.

B. SOME COMPANIES MAY BENEFIT FROM ADVANTAGES LINKED TO THEIR ACTIVITIES IN OTHER DIGITAL MARKETS

188. The position of certain operators in other markets linked to generative AI may give rise to a range of competitive advantages, which are not easily replicable by foundation model developers.

1. PREFERENTIAL ACCESS TO THE INPUTS NEEDED TO TRAIN AND DEVELOP FOUNDATION MODELS

a) Easier access to computing power

189. As discussed above, creating a sufficient computing infrastructure is particularly costly and difficult for a new entrant. As discussed in Opinion 23-A-08[120], **AWS, GCP and Microsoft Azure are the three leading cloud service providers in France, and among the top in the world**. Accordingly, they already have the financial and technical capacity and expertise required to create and manage such an infrastructure.

190. Furthermore, **major digital companies' ability to buy in large quantities and negotiate preferential agreements with suppliers such as Nvidia** can give them access to resources even in times of high demand and scarcity. For example, Meta is developing its AI infrastructure to integrate 350,000 Nvidia H100 GPUs by the end of 2024[121], a purchase estimated to have cost around $9 billion (around €8.3 billion)[122]. **Access to a large user base** also enables these companies to better optimise their computing infrastructure.

191. Even for major digital companies, however, the GPU shortage can pose challenges. In a market where demand outstrips supply, they may also find it difficult to obtain enough GPUs to meet their needs and those of their customers, which can lead to delays in the deployment of new products or services.

192. Consequently, **several digital giants have started to develop their own AI accelerators specifically tailored to their ecosystems**, such as Google's TPUs, AWS' Trainium and Microsoft's Maia. Supplying the chips in-house has the advantage of not being dependent on the product development cycles of external GPU suppliers and of reacting more quickly to market developments. Hyperscalers can also design chips specifically to meet their needs and those of their customers, which may include optimisations for certain AI tasks or specific functionalities not found in external GPUs. One stakeholder confirmed that "*optimising your own chips, for your own data centres, for your own models, can be a significant competitive advantage*". Furthermore, although developing AI accelerators in-house requires a significant initial investment given the need to acquire the necessary production capabilities externally or to hire new staff, it may be more cost-effective in the long term to develop your own chips rather than to source them from a single player, especially if cloud service providers can produce these chips on a large scale. These advantages could gradually

[120] Hyperscaler cloud services are discussed in detail in above-cited Opinion 23-A-08.

[121] Meta, Building Meta's GenAI Infrastructure, 12 March 2024.

[122] Le Monde Informatique, Pour l'IA, Meta va acquérir 350 000 accélérateurs Nvidia H100, 19 January 2024.

position these chips favourably in relation to the competition[123], although major companies currently seem reluctant to market their own chips to third parties outside their cloud.

193. In addition to chip development, several major companies are working on alternatives to Nvidia's CUDA software. Google's TPUs are designed for use with Google's TensorFlow software. OpenAI has developed Triton, whose "*aim* […] *is to provide an open-source environment to write fast code at higher productivity than CUDA*"[124]. The UXL Foundation, which brings together suppliers such as Google and Intel, also plans to create a suite of open-source software and tools capable of powering several types of accelerator[125].

194. Companies therefore find themselves in an ambiguous position, as they are both **partners and competitors** of suppliers of microprocessors for generative AI. As France Digitale reported, "[m]*ajor Nvidia customers include vertically integrated cloud companies: AWS, Alibaba, Google and Microsoft. These operators occupy a specific position, combining the roles of Nvidia's partners and competitors. They are investing in the design of their own specific chips to reduce their dependence on Nvidia, and also forging strategic partnerships and announcing major investments in the purchase of Nvidia GPUs*"[126].

b) Preferential access to data

195. A number of key digital companies enjoy significant advantages as regards the collection of the data required for training and fine-tuning models.

196. First, they have access to the volume of data needed to train foundation models.

197. For example:

- in addition to being one of the only fully vertically integrated companies in the sector, with access to both an infrastructure developed in-house (TPUs) and a self-developed LLM (Gemini), Alphabet has access to a considerable wealth of data, thanks in particular to data from its Google Search index, the use of Google Chrome, Google Ads and Google Maps, as well as YouTube and Google Books. For example, YouTube is said to host 10 billion videos, offering Alphabet a major source of training data for AI models (videos or language with text transcriptions of videos);
- Meta benefits from large datasets thanks to its Facebook, Instagram and WhatsApp platforms. Mark Zuckerberg recently stated that Facebook and Instagram have "*hundreds of billions of publicly shared images and tens of billions of public videos, which we estimate is greater than the Common Crawl dataset* [...]"[127];
- Microsoft owns the search index that powers the Bing search engine, as well as GitHub, the leading code sharing platform for developers.

---

[123] In its quarterly results presentation, Amazon confirmed that "[w]*e have the broadest selection of NVIDIA compute instances around, but demand for our custom silicon, Trainium and Inferentia, is quite high given its favorable price performance benefits relative to available alternatives*" (page 5).

[124] Triton GitHub project.

[125] Le Monde informatique, La Fondation UXL peaufine une alternative au Cuda de Nvidia, 26 March 2024.

[126] France Digitale, Des puces aux applications, l'Europe peut-elle être une puissance de l'IA générative ?, April 2024, page 13. (English translation also available: From chips to apps, can Europe compete in generative AI?, page 12).

[127] Meta Platforms, Inc. (META), Fourth Quarter 2023 Results Conference Call, 1 February 2024 (page 3).

198. **Their position also gives them preferential access to a wealth of metadata and data associated with the use of their services**. Accordingly, they benefit from access to indirect data to which smaller operators do not have access. When using a model in the inference phase, a company that collects data, notably on user satisfaction, can optimise the model and future models. The more a model is used, the more feedback the developer will have to improve its performance. Given this positive feedback loop, a strong competitive position is likely to consolidate rapidly, or even become dominant.

199. In addition to the large volume of data that these operators have access to, they also enter into a number of agreements with third-party data owners (see paragraph 158 above). For example, Alphabet agreed to pay $60 million (around €55 million) a year for access to data from Reddit, a US social news aggregation and forum social network[128].

200. **Access to this data is not the same for model developers and major digital companies.** First, because they are not able to enter into agreements with content providers on the same financial terms as major digital companies. Second, because they cannot easily access the data of the major operators, given the conditions set by them.

201. First, while major digital companies have access to a large volume of data within their ecosystems, this data is not necessarily free of copyright. According to one stakeholder, "*we need to make a clear distinction between the notion of 'proprietary' or 'first-party' data, as used by technology conglomerates – corresponding in essence to any data passing through them and remaining within their ecosystem with operators with which they have agreements –, and 'third-party' data, which, according to the same technology conglomerates, corresponds to data shared with operators other than those operating in closed ecosystems, with data or content protected by copyright or related rights, or by database rights under Directive 96/9/EC, or even personal or non-personal data within the meaning of the GDPR and the e-Privacy Directive, or commercially sensitive data that may be protected under business secrecy*".

202. Consequently, emblematic content producer groups have begun discussions with the major generative AI operators, either by reaching agreements, as Axel Springer and Le Monde did recently (see above), or by filing legal proceedings, as the New York Times has done vis-à-vis OpenAI. Currently, however, the sums involved in accessing this content limit the ability of smaller operators to enter into similar agreements, as demonstrated by the agreement between OpenAI and NewsCorp for almost $250 million (over €230 million) over five years. According to the information available to the *Autorité*, only OpenAI has been able to enter into such agreements with press publishers to date.

203. Furthermore, data derived from the services of major digital companies is not easily accessible to developers, unless they violate the rules of use applicable to these services. Following rumours of Sora (the OpenAI model capable of producing videos following text requests) being trained on YouTube videos, YouTube's CEO indicated that if this were the

[128] Le Figaro, IA : Reddit noue un accord de licence inédit avec Google pour 60 millions de dollars, 22 February 2024. In its filing with the US Security and Exchange Commission, Reddit said: "[w]*e are also in the early stages of monetizing our emerging opportunity in data licensing by allowing third parties to access, search, and analyze data on our platform. In January 2024, we entered into certain data licensing arrangements with an aggregate contract value of $203.0 million and terms ranging from two to three years.* [...] *Reddit data constantly grows and regenerates as users come and interact with their communities and each other. We believe our growing platform data will be a key element in the training of leading large language models (LLMs) and serve as an additional monetization channel for Reddit*".

case, "[it would be] *a clear violation of our ToS* [terms of service]"[129]. Microsoft has also reportedly threatened to deny access to data from its search engine (Bing) to its licensed competitors if they use this data to train their generative AI tools[130].

204. However, most digital giants consider that the privileged nature of their data access can be nuanced, insofar as developers have access to large quantities of data on the Internet and in publicly-accessible datasets.

205. While there are many public datasets available (see paragraphs 150 *et seq.* above), major digital companies still have an undeniable advantage in terms of accessing data. Their foundation models have been trained not only on public datasets but also on proprietary data that is not accessible to third parties under the same conditions. Their access to proprietary data is all the more important because access to public data is becoming progressively more difficult. Agreements between model developers and content providers also show that public data is not necessarily sufficient. Furthermore, even when data access is comparable, major digital companies also control the computing power required, enjoy better access to specialised expertise and have acquired extensive experience in data collection, labelling (see glossary) and analysis.

206. Some smaller companies may have an advantage as regards the fine-tuning of a pre-trained model for a specific sector, and have access to sector-specific data. For example, a pharmaceutical company might have access to proprietary clinical trial data, which it could use to train or fine-tune a foundation model used in the medical sector. However, major digital companies also enjoy preferential access to customised and specialised datasets in many fields, such as healthcare, finance and transport. This is the case, for example, with Google for mapping and health data (thanks to Fitbit) and with Amazon, which has access to sensitive health data thanks to its recent acquisition of One Medical[131]. In the financial sector, the *Autorité* found in its "FinTech" opinion[132] that Amazon, Apple, Google and Meta were likely to have access to financial data as part of the development of payment methods.

207. In addition to companies' access to sector-specific data, it is also worth noting that different activities produce data to a greater or lesser extent. Some companies may be required to process very large quantities of data due to the nature of their business (e.g. banks), or because their very purpose is data intermediation.

---

[129] The Verge, OpenAI training Sora on YouTube videos would violate the platform's rules, 4 April 2024.

[130] Reuters, Microsoft threatens to restrict data from rival AI search tools – Bloomberg News, 27 March 2023.

[131] Open Markets Institute, AI in the Public Interest: Confronting the Monopoly Threat, November 2023.

[132] *Autorité* Opinion 21-A-05 of 29 April 2021 on the sector of new technologies applied to payments, paragraphs 356 to 358.

c) The ability to attract highly-skilled employees

208. The *Autorité* recalled above the need for advanced technical skills to develop high-performance foundation models. Major digital companies have a number of advantages to attract the best talent.

209. Thanks to their financial capabilities, they can offer attractive salaries, shares and stock options, as well as significant fringe benefits. Furthermore, they can offer attractive job prospects, given their reputation for innovation, their global positioning and their wide catalogue of services. For example, as mentioned above, most of the major companies have high-performance in-house research laboratories, offering very comfortable working conditions. Google's new AI research centre in Paris has a budget of €300 million and is expected to bring together over 300 researchers with access to advanced AI tools[133]. Lastly, while their reputations may have suffered from scandals such as Cambridge Analytica[134], digital giants like Google, Microsoft and Apple featured in the rankings of the most attractive companies for young graduates in 2023, despite a decline in their position since 2022[135].

210. In addition to attracting the best talent, major digital companies can also forge partnerships with highly innovative start-ups, giving them access to the best possible expertise (see above).

211. The above is confirmed by stakeholders. According to one, "[t]*he major operators in the digital sector, mainly the Americans, enjoy a number of structural advantages to attract the best talent: the reputation of innovative companies, stimulating projects, virtually unlimited resources and tools, and very high levels of remuneration. They therefore offer a general working environment that is particularly attractive to talent*".

2. THE ADVANTAGES OF VERTICAL AND CONGLOMERATE INTEGRATION FOR MAJOR DIGITAL COMPANIES

a) Economies of scale, scope and network effects

212. Large, vertically integrated or conglomerate digital companies have access to funding, talent, data and computing power due to their activities in distinct but related markets in the generative AI sector.

213. The sector is characterised by the high fixed costs involved in the initial training of a foundation model (in particular, the acquisition of computing resources and any data acquisition agreements), which gives rise to **economies of scale** as operators seek to spread costs over as many users as possible. As a result, an established operator with significant production capacity and a user base will have an advantage over smaller operators. Furthermore, at the level of data centres themselves, increased activity can lead to an increase in the size of the data centre, which in turn increases fixed costs, but also reduces unit costs thanks to gains in energy (lower cooling costs in particular), labour or security, for example.

---

[133] La Tribune, IA : avec son nouveau centre de recherche à Paris, Google entend former 100.000 professionnels, 15 February 2024.

[134] CNBC, Facebook has struggled to hire talent since the Cambridge Analytica scandal, according to recruiters who worked there, 16 May 2019.

[135] Les Echos Start, Classement des boîtes préférées des jeunes cadres : l'industrie progresse, les GAFAM en baisse, 23 June 2023.

214. Potential economies of scale (due to high fixed costs) appear to be less significant downstream, with new entrants able to use open-source models and the computing power required for inference depending on the number of users.

215. Generative AI products can also be characterised by **economies of scope**. Economies of scope appear when a company can increase its output by producing different goods from the same factors. Once developed, a foundation model can be used for a variety of applications, with fine-tuning costs modest compared to the initial development. For example, Google's general-purpose model Gemini has been used to train several specialised models, such as MedGemini in the healthcare sector.

216. Furthermore, major digital companies are particularly well placed to leverage their access to computing infrastructure and data, as well as their technical knowledge, to launch and develop models, as confirmed by the Organisation for Economic Co-operation and Development (OECD) in a recent note: "*data could be an economy of scope if operating in adjacent markets allows firms to capture data that improves its ability to develop better generative AI models. There may also be synergies if staff that work in related fields can work on AI development*"[136].

217. As a result, many stakeholders believe that model developers must achieve critical mass in model deployment, downstream of the market, to recoup the very high initial investment costs.

218. The generative AI sector is also characterised by **network effects**. The performance of a generative AI service can be improved as and when the service is used, since user feedback data can be used to refine the model. Several stakeholders indicated that "*access to a large user base can create a virtuous circle, as the larger the customer base, the greater the scope for improving the model and thus for attracting new users. Companies or end users could be encouraged to choose a foundation model that already has a significant presence in the market, on the assumption that it's the most efficient*". Digital giants with large user bases are in a better position to take advantage of network effects, which could constitute a barrier to entry in the future.

219. Lastly, the major digital companies' existing activities can help to **finance new AI initiatives**, if they are also present in high-margin business sectors, such as the cloud. At the same time, model developers must continue to raise capital and generate revenues to continue funding their R&D activities and model deployments.

[136] OECD, Artificial intelligence, data and competition – Background Note, 6 May 2024, page 27.

220. Downstream, by being directly involved in every stage of the value chain, major digital companies can also **better understand market needs** and adjust their products more quickly to meet customer requirements, as confirmed by one stakeholder: "*by developing generative AI tools integrated into company cloud services, hyperscalers are able to provide their generative AI services with highly-qualified information that can be used, for example, for fine-tuning models, based on a very advanced understanding of their customers' behaviour in terms of digital usage and their needs in terms of IT services and data management, often on a global scale for multinational customers. This highly-qualified information is clearly a differentiating factor that is difficult to reproduce and perfectly suited to the business of analysing needs that can be met by AI, creating a significant barrier to entry over time*". It is possible, for example, to imagine a hypothetical situation in which a major digital company, leveraging the power of an AI model and the data and metadata collected through the operation of a social network, could design specialised AI-enabled applications in the human resources sector.

221. In the not-too-distant future, companies in the sector could be encouraged to enter into agreements to facilitate their access to energy, which could lower their costs, especially for the major operators. Microsoft's partnership with the nuclear fusion company Hélion may be an indicator of this trend[137].

b) The gradual creation of ecosystems

222. The *Autorité* also notes that major digital companies are starting to integrate generative AI tools into **their product and service ecosystems**.

223. In Opinion 23-A-08, the *Autorité* found that some service providers were creating cloud ecosystems: "*an analysis of the way the sector operates, and the positioning of the different players, tends to show that some providers are building cloud ecosystems, i.e. a set of integrated services that customers can access, including the provider's proprietary services, but also, generally through marketplaces, a set of services from third-party developers, designed to operate within this ecosystem. The industry could therefore be structured around competition between cloud ecosystems*" (paragraph 252).

224. The same logic seems to be at work in the generative AI sector.

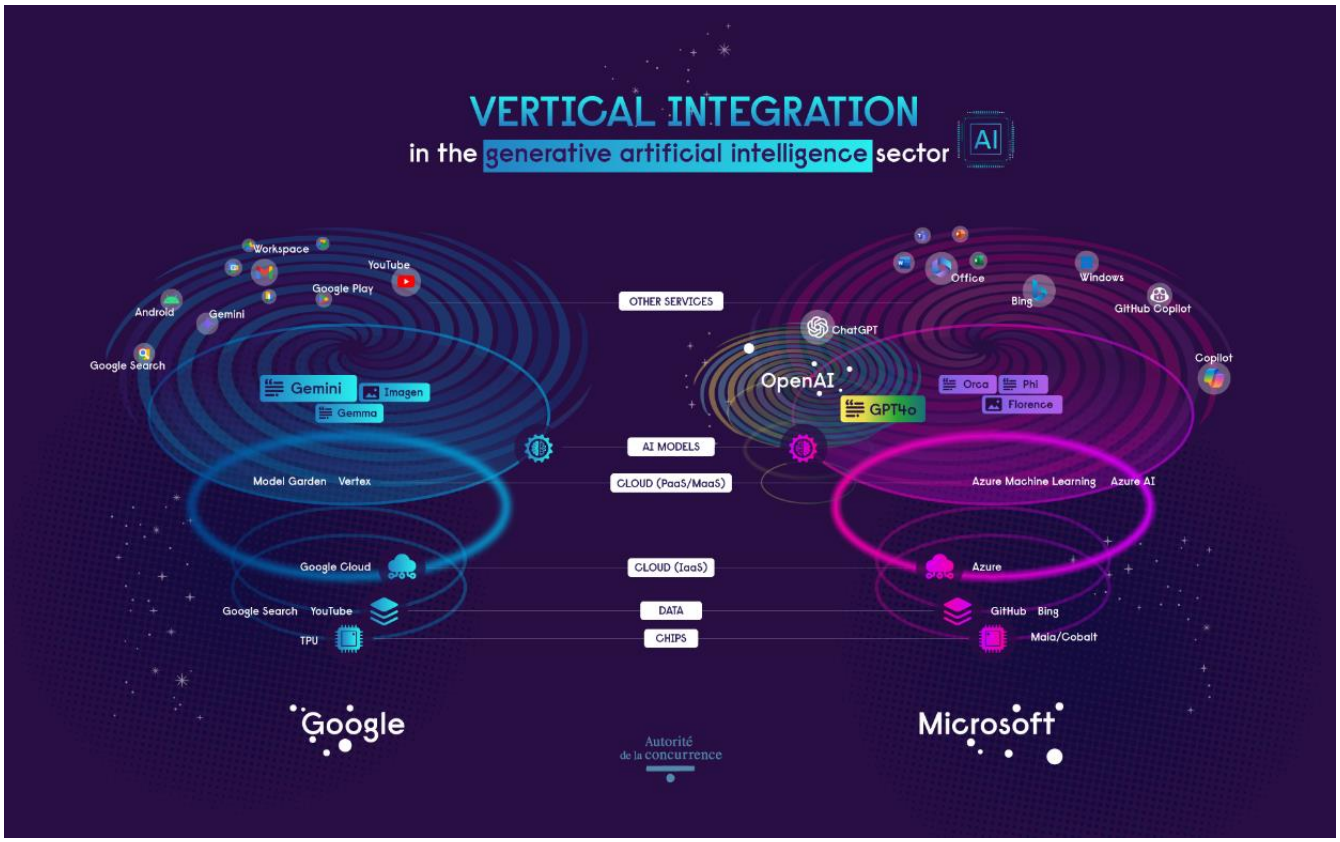
225. **Generative AI services are increasingly integrated with services located in distinct but related markets, in which major digital companies have significant market power**. For example, Microsoft deploys its own models and those of its partner OpenAI in the "Copilot" function to enhance Microsoft Bing's search functionality and offers "Copilot for Microsoft 365", an AI assistant designed to work with the Microsoft 365 suite (including Word, Excel, Outlook, Teams and PowerPoint). Similarly, Google uses Gemini to improve its search engine (AI Overview) and is starting to offer the "Gemini for Workspace" service to facilitate writing in Gmail and Docs, as well as a creative image generator in Slides.

226. In addition, **major digital companies' MaaS marketplaces provide access to proprietary and third-party generative AI models designed to run in their ecosystems**. For example, Google's Model Garden offers over 130 foundation models, including proprietary models such as Gemini, open-source models such as Meta's Llama 2, and third-party models such as Anthropic's Claude 3.

[137] Hélion Energy, Announcing Helion's fusion power purchase agreement with Microsoft.

227. It appears that Microsoft and Google in particular enjoy significant advantages in the generative AI sector. They control most of the access to the inputs needed to develop foundation models and are vertically integrated along the value chain, enabling them to control the development of their own and third-party foundation models. Thanks to their activities in other digital markets, they benefit from the necessary investments and are gradually creating ecosystems. Figure 8 shows Google's and Microsoft's ecosystems throughout the generative AI value chain.

Figure 8: Vertical integration of Google and Microsoft in the generative AI value chain



Source: Autorité de la concurrence.
OpenAI is included in Microsoft's value chain, as most of Microsoft's downstream services (Copilot, Bing, etc.) are based on OpenAI's GPT-4 or GPT-4o models.

228. Major digital companies therefore have preferential access to all the inputs needed to develop foundation models. In addition to their financial power, which they can use to enter into a number of partnerships, they also have easier access to the necessary computing power, preferential access to data, and the ability to attract highly-skilled employees. All these factors combine to form a considerable barrier to the entry and expansion of competitors in the market, and the advantages enjoyed by the major companies can entail risks to competition.

C. COMPETITION RISKS UPSTREAM IN THE VALUE CHAIN

229. During the preparation of this opinion, a number of practices implemented or likely to be implemented in the sector were identified, which could restrict competition.

1. PRACTICES THAT COULD BE SANCTIONED UNDER COMPETITION LAW

a) Introductory remarks

230. Before discussing potential abuses of dominant position, there are a number of observations on the question of the market power of certain companies.

231. At this early stage in the development and implementation of generative AI, we do not currently have sufficient information to precisely determine, as part of this opinion, the relevant markets and assess the market shares of the different operators involved. The characteristics of foundation models vary widely, and downstream market access routes for these models and the way in which they are marketed are still under development. Some contributions nevertheless identified "*two main markets, the primary market for the design and pre-training of primary fundamental models, and the secondary market for the development of models that are specialised or fine-tuned to meet predetermined objectives*". Others questioned the impact of integrating generative AI into the definition of relevant markets: "*With regard to the analysis of relevant markets, should the integration of generative AI in the marketing of a product or service be assessed as an innovation that adds a new functionality to existing offers, or does it position them on a different market compared to the same products or services not equipped with generative AI? Similarly, should a generative AI system integrated into the core functionalities of a company's product or service be considered as an additional component belonging to a market distinct from that of this product or service?*".

232. Apart from market shares, competition authorities take into account other criteria when assessing the market power that a company might have in *a priori* dynamic sectors, such as the existence of barriers to entry or expansion.

233. While it seems premature to draw definitive conclusions on the definition of relevant markets and the market power of certain operators, vigilance is nevertheless required because certain companies' access to key inputs and the advantages linked to their vertical and conglomerate integration create the conditions for strong concentration, to their benefit, and reinforce their power on distinct but linked or related markets, such as productivity software, search engines and online advertising. **In certain cases, it may therefore be useful to perform the competitive analysis in terms of ecosystems being either created or reinforced, rather than market by market.**

234. In addition, **the generative AI sector does not appear to call into question the relevance of traditional competition law tools and concepts.**

235. Accordingly, the use of the traditional tools of competition law, **such as antitrust law** (see below) **and, above all, abuse of dominant position**, remains fully effective and can be justified. According to decision-making practice and case law, a company with a dominant position in a given market may be accused of an abuse whose effects are felt in other markets, if the market in which the company holds a dominant position and those in which the abuse has an effect are sufficiently related, and special circumstances justify the application of the rules prohibiting abuse of a dominant position[138].

236. Furthermore, within these ecosystems, an abusive practice may aim to drive out certain competitors of the dominant operator, but it may also exploit the fragility of other operators in the ecosystem by imposing higher prices, contractual restrictions (such as mono-hosting constraints or exclusivity clauses), or by directly or indirectly restricting their use of services (such as data to which access is asymmetrically restricted or excessive extraction of data).

237. Antitrust law also plays its part, as discussed below.

238. Lastly, other legal tools could be used, such as **abuse of economic dependence**. An abuse of economic dependence is not assessed based on a company's position in a given market, but on the specific nature of its commercial relationships with upstream or downstream partners, and can be used where no position of dominance exists. As the *Autorité* has already indicated in aforementioned Opinion 23-A-08 and in its June 2020 study on "E-commerce", abuse of economic dependence can be used to address abusive contractual practices by digital operators and marketplaces[139].

239. The **law on restrictive competition practices**, the implementation of which falls mainly within the remit of the Directorate General for Competition Policy, Consumer Affairs and Fraud Control (DGCCRF) and the commercial courts, also provides an effective weapon for sanctioning unfair practices. These tools are described in detail in aforementioned Opinion 23-A-08 (paragraphs 562 *et seq.*), to which this opinion refers.

240. The *Autorité* will therefore focus on the competition risks present at each layer of the upstream value chain. It will also take a broader look at the risks arising from the presence of major operators in several linked or related markets.

---

[138] CJEU judgement, 14 November 1996, Tetra Pak International SA v. Commission, Case C-333/94 P, ECR 1996 I-05951, paragraph 27; CJEU judgement, 17 February 2011, Konkurrensverket v. TeliaSonera Sverige AB, cited above, paragraph 86; *Autorité* Decision 22-D-20 of 15 November 2022 regarding practices implemented in the sector of payroll management solutions for entertainment workers, paragraphs 89 *et seq*.

[139] Aware of the profound imbalances characterising digital markets, national competition authorities, such as the Italian authority, have incorporated provisions relating to abuse of economic dependence. For example, under Italian law, since 1 November 2022, there has been a presumption of economic dependence in commercial relationships between digital platforms and the companies that use their intermediation services if the platform plays a decisive role in reaching end users or suppliers, particularly in terms of network effects or data availability.

b) Risks of abuse upstream in the value chain

The risk of abuse by IT component providers

241. It is clear from the above developments (see paragraphs 124 *et seq.*) that a single operator, Nvidia, appears to have a dominant position in the sector for the IT components needed to train foundation models.

242. Given the severe constraints on the supply of GPUs and the concentration of the sector, several operators fear that the current climate is conducive to potentially anticompetitive practices:

- a recent study by France Digitale[140], based on interviews with some 40 companies in the sector, pointed to potential risks such as **price fixing, supply restrictions, unfair contractual conditions and discriminatory behaviour**;
- concern has also been expressed regarding the sector's dependence on Nvidia's **CUDA** chip programming software, the only one that is 100% compatible with the GPUs that have become essential for accelerated computing;
- recent announcements of Nvidia's investments in AI-focused cloud service providers, such as CoreWeave, are also raising concerns among general-purpose cloud providers. According to France Digitale, "[s]*uch specialized cloud providers also benefit from a Preferred Partnership with Nvidia, which allows them to offer GPU access at 80% more cost-effective rates than general purpose cloud competitors. This could lead to unfair pricing competition with general purpose cloud providers, especially mid-tier actors that lack the financial firepower of hyperscalers*".

243. **Another concern is the risk that computing power will eventually be concentrated in the hands of the major digital companies**. As explained in paragraphs 189 *et seq.* above, their preferential access to Nvidia GPUs, the development of their own AI accelerators, and their investments in innovative companies in the sector could gradually reduce competition for access to computing power[141].

244. The graphics card sector, **which was the target of a dawn raid in September 2023**, is being closely scrutinised by the *Autorité*'s Investigation Services[142].

The risk of lock-in by major cloud service providers

245. Incumbent cloud service providers play an important role in the development of new AI technologies, as they provide large amounts of the computing resources needed primarily by language model developers. However, their position as suppliers of an essential input for AI technologies creates the risk of them abusing that position to extend their market power and reduce competition.

---

[140] France Digitale, Des puces aux applications, l'Europe peut-elle être une puissance de l'IA générative ?, April 2024, page 13. (English translation also available: From chips to apps, can Europe compete in generative AI?, page 12).

[141] Maurice E. Stucke andAriel Ezrachi, Antitrust & AI Supply Chains, 11 March 2024, page 18.

[142] *Autorité* press release, The General Rapporteur of the *Autorité de la concurrence* indicates that an unannounced inspection was carried out in the graphics cards sector, 27 September 2023.

246. **The *Autorité* notes that several financial and technical lock-in practices, already identified in Opinion 23-A-08 on competition in the cloud sector, appear to remain and even to be intensifying to attract the largest possible number of start-ups active in the generative AI sector.**

247. First, particularly high levels of **cloud credits** are being offered, especially to innovative companies in the sector.

> As a reminder, cloud credits are trial offers in the form of service allowances offered by a provider and granting free access to a customer for a defined period. In its aforementioned opinion on the cloud sector, the *Autorité* considered that credit programmes aimed exclusively at certain target groups, in particular those with a high potential for innovation, such as start-ups, developers, researchers and students, merit particular attention. The *Autorité* indicated that "[t]*he high amounts sometimes proposed, the vast ecosystem of companies they cover, their validity periods and the lock-in risks described above set them apart significantly from the free trials that can traditionally be seen in other industries, and raise doubts about the ability of all cloud service providers to respond*"[143].

248. A similar strategy is being developed here, notably by hyperscalers, for start-ups active in the generative AI sector. As an example, Google Cloud offers "*AI start-ups*" eligible for the "Google for Startups Cloud Program"[144] up to $350,000 (around €325,000) over two years for the use of Google Cloud and Firebase (a Google platform for rapid application development), which is $150,000 (around €140,000) more than start-ups active in other sectors[145]. Similar programmes with bonuses for AI start-ups are also offered by Amazon[146] and Microsoft[147].

249. The amount of credits can also be increased on condition that the beneficiary start-ups use the suppliers' new AI products. For example, companies selected for "AWS Activate"[148], a programme reserved for start-ups, can receive up to $100,000 (around €92,000) in AWS promotional credits to start creating models. In addition, start-ups using AWS' AI accelerators (AWS Trainium and AWS Inferentia) may be eligible for up to $300,000 (around €280,000) in additional credits.

---

[143] See above-cited *Autorité* Opinion 23-A-08, paragraph 423.

[144] Google Cloud offers its customers discounts and credits for computing and other products (including support for AI-focused start-ups) via the Google for Startups Cloud Program.

[145] Google website states: "*Get access to startup experts, your Google Cloud and Firebase costs covered up to $200,000 USD (up to $350,000 USD for AI startups) over two years, technical training, business support, and Google-wide offers*".

[146] In 2023, the "AWS Generative AI Accelerator" offered a global programme designed to help 10 start-ups in the field of generative AI realise their potential. Eligible participants could receive up to $300,000 (around €280,000 euros) in AWS credits. Amazon recently announced that it was extending its credits offer to cover the costs of start-ups using leading AI models (such as Anthropic, Mistral and Cohere): "*This is another gift that we're making back to the startup ecosystem, in exchange for what we hope is startups continue to choose AWS as their first stop*".

[147] Microsoft supports start-ups active in the generative AI sector through its "creators' hub" programme, which offers up to $150,000 (around €140,000) in Azure credits to eligible companies.

[148] https://aws.amazon.com/startups/generative-ai/.

250. These offers are particularly attractive for start-ups, as they give them free access to the cloud services needed to train, fine-tune and deploy their solutions. Nevertheless, given the scale of the costs involved in training and fine-tuning AI models, this practice has the effect of encouraging users to choose the services of hyperscalers due to the amount of cloud credits offered, and not just because they would best meet their long-term needs. Cloud credits could therefore have the effect of locking the companies concerned into hyperscaler ecosystems, against a backdrop of technical and price barriers to migration.

251. In addition to cloud credits, other practices have been identified as **technical lock-in practices**, such as hyperscalers offering proprietary solutions (e.g. automated machine learning services, see glossary) for companies wanting to create or fine-tune their models more easily. However, users do not have access to the model itself after the final model is created and can only use or deploy the model from the cloud service provider's infrastructure. This type of practice would lock in users that, if they wanted to change cloud service provider, would have to recreate their AI model from scratch, as the model could not be transferred to another provider.

252. In addition to being qualified under competition law, notably as abuses of dominant position, some of these practices are also governed by the SREN Law or the EU Data Act.

The SREN Law

The SREN Law is based on three main principles: protection of (i) citizens, (ii) young people, and (iii) businesses and local and regional public authorities. The aim of the third principle is to anticipate the implementation of the EU Data Act and limit the use of cloud credits over time.

In Opinion 23-A-05 on the draft SREN Law[149], the *Autorité* emphasised that, given the European regulatory context in which the draft law was being introduced, it was important to ensure that the planned measures were properly coordinated with the future European framework, so as not to penalise stakeholders operating in the French market.

Article 26 of the SREN Law stipulates that "*cloud credits*" are limited to one year and cannot include an exclusivity clause. Said Article prohibits the unfair trade practice of making the sale of a product or service conditional on the simultaneous conclusion of a contract for the supply of cloud services. Lastly, the law provides for a report by the *Autorité* on the practice of "self-preferencing", as well as any necessary procedural or legislative improvements.

Article 27 prohibits egress fees in excess of the costs charged when a customer wants to transfer its data to its own infrastructure or to the infrastructure of another provider within a multi-cloud architecture. Article 28 sets out requirements for the interoperability of cloud services and data portability. Both articles include a sunset clause of 12 January 2027 to ensure compatibility with the provisions of the Data Act.

[149] *Autorité* Opinion 23-A-05 of 20 April 2023 on the draft law to secure and regulate the digital space.

The risks associated with data access

253. A number of competition concerns are likely to be raised in both the upstream and downstream parts of the value chain, particularly with regard to access to data.

254. Developers need huge amounts of general data during the training phase of foundation models and specialised data during the fine-tuning phase. At the inference stage, AI chatbots need access to data to answer user queries.

255. However, innovative companies in the sector may be confronted with **practices of refusal of (or discriminatory) access by companies with significant access to data**. For example, France Digitale indicated: "*companies with significant access to data (e.g. web indexes or search engines) could deny or restrict access to data under their control. Such players could provide more favorable treatment to developers with whom they have a partnership (e.g. for the provision of cloud or platform services), or to their in-house services. Also, companies with a dominant position could compel their contracting parties not to provide their data to rival AI developers. For example, they could impose anti-web scraping measures or exclusive data use rights in exchange for advertising, web referencing, or cloud services. Finally, big players could offer services or technology (e.g. inference rights) in exchange for data, making the discussions with other players less appealing for data providers*"[150].

256. With this in mind, **the *Autorité* will be vigilant regarding the data made accessible to major digital companies as part of their partnerships with companies**. A partnership between a major digital company and a French industrial company for the fine-tuning of its models on its proprietary data would not give the digital company access to the same volume of data as a partnership with a data intermediation company, for example. The latter partnership would be carefully examined by the *Autorité.*

257. Refusal of access to data could take more subtle forms. For example, powerful operators could seek to acquire or consolidate a dominant position in the generative AI sector by proposing the payment of substantial remuneration to content creators, in particular to **exclude less established competing operators** or potential entrants. According to some stakeholders, the high remuneration could be offset by increased market power due to the marginalisation or exclusion of less established operators[151]. High remuneration for content creators could therefore potentially constitute an abuse of dominant position.

258. However, **digital data is economically a non-rival good**; in other words, selling data to one operator does not *a priori* limit the ability to sell the same data to another operator, possibly at a different price. However, **the question remains as to whether content providers are inclined** to enter into such differentiated agreements, and, to the best of the *Autorité*'s knowledge, no press publisher has yet signed an agreement with several model developers at different prices.

[150] Above-cited France Digitale study, page 32 (page 29 of English translation).

[151] See the economic literature on incentives to raise rivals' costs, including Salop and Scheffman (1983), *"Raising rivals' costs*", American Economic Review; Krattenmaker and Salop (1986), "*Anti-competitive foreclosure: raising rivals' cost to achieve power over price*", the Yale Law Journal; Salop (2017), "*The raising rivals' cost foreclosure paradigm, conditional pricing practices, and the flawed incremental price-cost test*", Antitrust Law Journal.

259. Moreover, **such practices, combined with exclusivity clauses,** could reinforce potential competition concerns. Exclusivity clauses, implemented by powerful operators, would be likely to prevent their competitors from accessing data under the same conditions. This type of agreement would therefore be likely to lock in data providers, thus limiting the opportunities for competitors.

260. Some stakeholders are also concerned about the advantage of being the first to form partnerships, via the inclusion in agreements between model developers and content providers of clauses that exchange data for cloud services. As a result, one stakeholder indicated that "*the big players could offer services or technologies (such as inference rights) in exchange for data, making discussions with other players less attractive for these data providers*".

261. Access to **user data** is also a major challenge, as indicated by the French AI Commission: "*it's clear that a lot of the data of interest for training AI is personal. In healthcare, of course, but not only. Even generative AI,* a priori *more interested in cultural data, may need it to develop a specific interaction capability. In education, training a model capable of interacting in a credible and relevant way with a student will probably require training on student-teacher dialogue data, which is personal data*"[152]. Several stakeholders reported that major companies in the sector continue to use various strategies to restrict third-party access to their users' data, by abusing the legal rules, such as personal data protection, or security concerns.

262. The interplay between competition law and personal data protection is a subject of particular attention for the *Autorité*, as demonstrated by its recent joint declaration with the CNIL on "Competition and personal data: a common ambition". The declaration recalls how the *Autorité* takes account of the competition parameter of "personal data" and also confirms that "*some privacy policies raise the question of the possible use of privacy arguments for anticompetitive purposes*"[153]. The *Autorité* is particularly attentive to ensuring that the implementation of the General Data Protection Regulation (GDPR) by the major digital companies does not create a risk of exclusionary or self-preferential behaviour.

263. The Meta judgement by the Court of Justice of the European Union (CJEU) also confirmed[154] that a national competition authority can find a breach of the GDPR and qualify such breach as an abuse of dominant position, specifying the terms of cooperation with the GDPR supervisory authorities.

264. Lastly, publishers are very concerned about the use of their content by foundation model providers **without the authorisation of rights holders**. The recent decision by the *Autorité* in the "related rights" case established that Google had used content from press agencies and publishers to train its foundation model Gemini (formerly Bard), without notifying them and without giving them an effective possibility to opt out[155]. While this question raises issues

---

[152] Above-cited French AI Commission report, page 100 (page 97 of English translation).

[153] Joint declaration by the *Autorité de la concurrence* and the CNIL, Competition and personal data: a common ambition, 12 December 2023, page 7.

[154] CJEU judgement, 4 July 2023, Meta Platforms Inc. and others v. Bundeskartellamt, C-252/21.

[155] *Autorité* Decision 24-D-03 of 15 March 2024 regarding compliance with the commitments in *Autorité* Decision 22-D-13 of 21 June 2022 regarding practices implemented by Google in the press sector. In particular, the *Autorité* found that Google's failure to inform press agencies and publishers of the use of their content by its Bard service constituted a breach of the transparency obligation under Commitment 1. Google also breached

relating to the enforcement of intellectual property rights that go beyond the scope of this opinion[156], competition law could, in principle, address these issues based on an infringement of fair trading, for example, and therefore, exploitative abuse. In this respect, the *Autorité* reiterates that it sanctions behaviour which, under the guise of protecting intellectual property rights, actually constitutes anticompetitive practices, as it goes beyond what is necessary for this legitimate protection[157].

The risks associated with access to a skilled workforce

265. While several companies have developed innovative generative AI solutions and therefore have significant expertise in the sector, major digital companies could implement practices that could limit or prevent the free movement of skilled employees, and therefore the associated competition.

266. For generative AI, as for the rest of the digital sector, human resources are a particularly scarce commodity for which companies are competing, as they constantly seek to attract and retain talent[158].

267. In France, and more generally in Europe, the legal means used by companies are first and foremost strictly governed by civil law. At national level, while Article L. 1121-1 of the French Labour Code (*Code du travail*) affirms the principle of freedom to work ("[n]*o restrictions may be placed on the rights of individuals and on individual and collective freedoms which are not justified by the nature of the task to be performed or proportionate to the aim pursued*"), companies may insert certain clauses in their employees' employment contracts, on the one hand, and in contracts between companies, on the other, to restrict or prevent the mobility of their employees. These types of clauses are closely monitored by the courts.

Commitment 6 by linking the use of press agencies' and publishers' content by its AI service to the display of protected content on services such as Search, Discover and News. The question of whether the use of press publications as part of an AI service qualifies for protection under related rights regulations has not been settled. Google did not contest the alleged practices and requested the benefit of the settlement procedure. The *Autorité* imposed a total fine of €250 million on Alphabet Inc, Google LLC, Google Ireland Ltd and Google France, and presented a series of corrective measures to address the concerns identified.

[156] Several content publishers indicated that they have never been informed of or authorised the use of their content for this purpose, nor have they received any remuneration, while some model developers are said to be taking advantage of Article L. 122-5-3, III of the French Intellectual Property Code (*Code de la propriété intellectuelle*), which stipulates that digital copies or reproductions of works that have been lawfully accessed may be made for the purpose of text and data searches carried out by any person, whatever the purpose of the search, unless the author has objected in an appropriate manner, in particular using machine-readable processes for content made available to the public online. The French Higher Council for Literary and Artistic Property (*Conseil supérieur de la propriété littéraire et artistique*) has been commissioned to examine possible legal mechanisms to guarantee fair remuneration for rights holders, and to analyse the economic issues underlying access to protected data when used by AIs.

[157] *Autorité* Decision 23-D-14 of 20 December 2023 regarding practices implemented in the sectors for eighth-generation static video game consoles and control accessories compatible with the PlayStation 4 console.

[158] See, for example, https://www.rhmatin.com/formation/digital-learning/enjeux-rh-et-formation-aux-ia-generatives-quelle-echelle-faut-il-atteindre-en-france.html and, more broadly, the 2022 review and 2023 outlook by Numeum, a professional organisation representing the digital ecosystem in France, reporting a "*shortage of talent trained in all the skills needed to deploy the latest technological innovations*".

268. The following are examples of these types of clauses:

- **a non-competition clause** is a clause in an employment contract whereby the employee undertakes not to engage in any activity competing with that of their employer, on their own behalf or that of another employer, for a specified period after the termination of the employment relationship[159]. To be valid, a non-competition clause must meet several cumulative criteria, such as being justified by the legitimate interests of the company and being limited in time and space. It must also target a specific activity and provide for financial compensation for the employee subject to the clause[160];
- **an employee non-solicitation clause** is an inter-company clause under which the beneficiary prohibits the other party from soliciting and/or hiring its employees, on pain of compensation generally based on the monthly salaries of the employees concerned. These clauses are often used when a service provider makes employees available to a company. They are used in particular in the digital sector, for the provision of digital products or solutions deployed in customer companies, and are designed to enable providers to retain their technicians, experts and consultants. According to the French Supreme Court (*Cour de cassation*), a non-solicitation clause is neither a variant nor a clarification of a non-competition clause[161].

269. In competition law, supervisory authorities pay particular attention to practices in the labour markets. In addition to wage-fixing agreements, no-poach agreements may also constitute prohibited anticompetitive practices.

270. In 2010, for example, the US Department of Justice brought an action against Adobe, Apple, Google, Intel, Intuit and Pixar for prohibiting each other from soliciting the most qualified employees[162].

271. In France, in a case concerning floor coverings[163], the *Autorité* also examined and sanctioned, among other anticompetitive practices, tacit non-aggression agreements or "gentleman's agreements" between competitors, aimed in particular at prohibiting the canvassing of their respective employees. In November 2023, its Investigation Services also announced that several companies from the engineering, technology consulting and IT services sectors had been notified of statements of objections regarding anticompetitive agreements in job markets[164].

---

[159] https://www.dalloz.fr/documentation/Document?id=DZ%2FOASIS%2F000185.

[160] See, in particular, Cassation, soc. 10 July 2002, 00-45.135.

[161] See, in particular, Cass. com., 31 Jan. 2012, no. 11-11.071, P+B, SAS Capp invest immo v. Sté Socorpi.

[162] Antitrust Division U.S. V. Adobe Systems, Inc, Apple Inc, Google Inc, Intel Corporation, Intuit, Inc, And Pixar. The agreements between Apple and Google, Apple and Adobe, Apple and Pixar and Google and Intel prevented the companies from directly soliciting each other's employees ("no cold call agreements"). An agreement between Google and Intuit prevented Google from directly soliciting Intuit employees. The agreements were seen as eliminating an important form of competition for attracting highly-qualified employees and reducing overall competition to the detriment of the employees concerned. The action, brought under Section 1 of the Sherman Act, ended in a settlement agreement which put an end to the proceedings.

[163] *Autorité* Decision 17-D-20 of 18 October 2017 regarding practices implemented in the hard-wearing floor coverings sector.

[164] *Autorité* press release, The General Rapporteur of the *Autorité de la concurrence* announces that several companies from the engineering, technology consulting and IT services sectors were notified of statements of objections regarding anticompetitive agreements in job markets, 23 November 2023.

272. According to the Commission[165], non-waiver agreements, such as wage-fixing agreements, are likely to be classified as restrictive of competition by object, and are prohibited by Article 101(1) TFEU. It specifies that while these types of agreements may have procompetitive effects, these effects are uncertain: the effects must be demonstrated and significant, while there may be less restrictive means or means more respectful of employees' rights and freedoms to achieve the same result, taking non-competition clauses as a reference, if they comply with national legislation.

273. With regard to the aforementioned clauses, the United States seems to be taking a different approach since the FTC has just banned the majority of non-competition clauses[166]. Existing non-competition clauses for senior executives remain in force.

274. An additional area of concern is the **recruitment by digital giants of entire teams** (such as Microsoft's hiring of most of start-up Inflection's 70-person staff) or **strategic employees** of model developers (such as Microsoft's brief recruitment of OpenAI's founder after his dismissal, before he was eventually hired back by OpenAI). While this type of practice may be examined under merger control rules (see paragraphs 295 *et seq.* below), it can also be analysed as an attempt to exclude competitors from the sector. The scarcity of talent could prevent foundation model developers from training high-performance models capable of competing with those of the industry giants. These practices could be addressed under competition law by prohibiting abuse of dominant position.

275. A specific, tried-and-tested legal remedy is also available under national common law, namely an action for unfair competition, whereby acts of disorganisation, parasitism[167] and disparagement resulting from mass recruitment campaigns or actions targeting key employees, may be sanctioned, based on common extra-contractual civil liability law, in order to obtain compensation for the damage suffered. This type of practice can be considered wrongful if it is implemented under unfair conditions and lead to the disorganisation of the targeted company[168].

276. Given the extremely rapid pace of technological change in the field of generative AI, it is vital that companies, and in particular model developers, are able to recruit and retain highly-skilled employees who are newly trained and/or have mastered the very latest technologies. While the adoption of the aforementioned clauses may have a legitimate objective (protection of investment in employee training, protection against abusive unfair competition, protection of skills and industrial secrets), legal remedies are available, and major digital companies must refrain from practices that are unfair, anticompetitive or prejudicial to employees, so as not to dissuade the entry of new innovative companies or unduly or abusively limit the mobility of employees to their competitors. These types of practices and restrictions could result in a concentration of talent within a limited number of companies, leading to significant distortions of competition in the labour market, to the detriment of the employees concerned and ultimately harming consumers.

277. Thus, while it appears from the preparation of this opinion that such restrictions are not, for the time being, raising any particular concerns for stakeholders, the *Autorité* considers that vigilance is required.

---

[165] https://competition-policy.ec.europa.eu/document/download/adb27d8b-3dd8-4202-958d-198cf0740ce3_en.

[166] FTC, FTC Announces Rule Banning Noncompetes, 23 April 2024.

[167] See, in particular, Cass. com. 5 February 1991, no. 88-16.214.

[168] See, in particular, Cass. com, 8 July 2020, no. 18-17.169 and Cass. com, 9 March 1999 no. 97-12.009.

The risks associated with open-source models

278. While open-source models can help to lower barriers to entry (see paragraphs 179 *et seq.* above), they can also raise competition concerns. In some cases, the conditions of access and reuse of models or some of their components can lead to users being locked-in.

279. The "Google Android" case[169] illustrates anticompetitive restrictions in the free software sector.

> **Judgment of the General Court of the European Union in Google and Alphabet v. Commission (Google Android)**
>
> Google had prevented device manufacturers from using another version of Android (its open-source operating system) not approved by Google (Android "forks", i.e. new software created from the source code of existing software). To pre-install Google's proprietary applications on their devices, including Play Store and Google Search, manufacturers had to undertake not to develop or sell any devices running Android forks.
>
> The General Court of the European Union found that the practice in question had led to a strengthening of Google's dominant position on the market for general search services, while at the same time deterring innovation, insofar as it had limited the diversity of offers available to users. On this basis, the Commission fined Google for abuse of a dominant position.

280. Similar restrictions on competition may also exist in the generative AI sector. The *Autorité* sees two types of risk associated with open-source models.

281. First, there are risks as soon as the models are made available. For example, model developers can prohibit the development of models that compete directly with their own models, and impose limits on their commercial exploitation or the design of competing products or services. This is the case, for example, with Meta's Llama 2 model, which requires an additional licence if usage exceeds 700 million users[170].

282. Second, some companies may initially adopt an open approach to generative AI to extend their market power[171] by locking-in user companies. One stakeholder warned that "*the competitive benefits of open source presuppose that model producers do not subsequently restrict access by exploiting user dependency ('lock in')*".

c) The risks associated with the presence of companies on several markets

283. More globally, the vertical integration of certain digital operators and their service ecosystems may give rise to a number of abusive practices.

---

[169] See European Commission case AT.40099 – Google Android and the judgment of the General Court in case T-604/18, Google and Alphabet v. Commission (Google Android) (appeal pending).

[170] L'Usine Digitale, Meta lance Llama 2, un grand modèle de langage open source et gratuit même pour une utilisation commerciale, 19 July 2023.

[171] Widder, David Gray and West, Sarah and Whittaker, Meredith, Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI, 17 August 2023.

284. At the upstream level, model developers could be **denied or given limited access to the chips or data needed to train competing foundation models.** For example, developers could be harmed by agreements that give a cloud infrastructure provider exclusive access to key data needed to train LLMs or a monopoly over the chips needed to develop and run the models. This type of practice could lead to delays or the introduction of less ambitious models, thereby undermining effective competition in the market.

285. Several stakeholders are also concerned about **exclusivity agreements** between cloud service providers and foundation model developers. In their view, such agreements aim to make the developers exclusively dependent on the cloud service providers for access to the necessary cloud services and for customer deployment, and are therefore likely to have an **impact on innovation** and competition between providers, especially when a particular model occupies a significant position on the market.

286. Such lock-in effects are further exacerbated when combined with other measures that confer influence over the model developer (such as high investments).

287. Other risks arise from the downstream use of generative AI models:

- **tying**: companies holding pre-eminent or dominant positions in related markets could tie the sale of products or services to that of their own AI solutions. This could be the case with practices or agreements linking different products together, for example by integrating a generative AI solution directly into a software offering (such as Microsoft's deployment of Copilot, its AI assistant powered by OpenAI's GPT-4 model) or into search engines (as with Google's recent introduction of AI Overview, which saves users from having to consult source sites to obtain information). **In particular, the integration of generative AI tools on certain devices, such as smartphones, is also raising concerns.** These devices may be part of major companies' ecosystems (Google recently extended the availability of its Gemini application to more Android smartphones) or belong to other manufacturers (Samsung and Google Cloud have announced a long-term collaboration to make Google Cloud's generative AI technology available to Samsung users worldwide[172]). This type of practice could permanently consolidate the generative AI sector around already dominant digital companies;
- downstream competitors could also be harmed by **self-preferencing practices of vertically integrated operators**, affecting the ability of developers of non-vertically integrated models to compete with those operators. This could be the case, for example, with practices that exploit the user data collected via their different products and use this data to improve the performance of their AI models, or with language model lock-in strategies that have the effect of limiting downstream competition.

288. Across the entire value chain, stakeholders observe that they have very little room for manoeuvre when negotiating prices and contractual conditions for using hyperscalers' AI models, as they suffer the same treatment as for other products and services offered by these hyperscalers.

289. Through any of the above behaviours, certain companies could use their market power to the detriment of alternative operators, thereby restricting the choice available to users and the incentive to develop alternative solutions.

[172] Samsung press release, Samsung and Google Cloud Join Forces to Bring Generative AI to Samsung Galaxy S24 series, 17 January 2024.

2. COMPETITION CONCERNS ABOUT MINORITY INVESTMENTS AND PARTNERSHIPS BY DIGITAL GIANTS

290. In a sector such as AI, where investment is very high given the cost of access to inputs, only a few major companies have the financial capacity to enter into agreements with or invest in innovative start-ups (see above). Investments and partnerships between operators in the sector are not problematic *per se*. They can give start-ups the opportunity to benefit from the financial and technological resources of major digital companies, and thus foster innovation. For the buyer, such investments enable diversification or access to innovative technologies to improve the quality of its services. They are therefore essential to the development of the AI sector.

291. Nevertheless, they present competition risks that call for particular vigilance by competition authorities. Since minority investments generally do not confer control, they are rarely examined *ex ante* under merger control rules. However, they may be examined *ex post* through competition law.

a) The need for particular vigilance in the generative AI sector

292. Certain investments require extra vigilance. While the transactions are likely to entail a number of competition risks, competition authorities have no clear information on the conditions of the agreements.

293. Even if they are minority investments, they can have a competitive impact on the sector:

- if the investment holder is a competitor of the target company, it may have rights over the target company's revenues, which could **weaken competition** between the two entities insofar as the target company's revenues also contribute to the buyer's revenues. The partnerships can also eliminate one of the competitors if this competitor simply integrates the partner's foundation models into its products instead of developing its own;
- **vertical effects** can also be found on the markets, since a customer company with a minority investment may have an incentive to buy from the target company, resulting in a competitive disadvantage between the target company and its competitors in the upstream market. Moreover, agreements between cloud service providers and model developers can strengthen the provider's market power, especially when they include an exclusivity clause;
- minority investments can also lead to coordinated effects by **increasing market transparency**, as the buyer is likely to acquire commercially sensitive information (such as business plans, up-to-date pricing data, information on model infrastructure such as hyperparameters, the process applied to clean data, or competitors' use of the developer's language model). One stakeholder expressed concern that partnerships between a cloud infrastructure provider and an LLM developer could give the cloud infrastructure provider access to sensitive information: "[i]*n the absence of robust safeguards, partnerships between a cloud infrastructure provider and an LLM developer could give the cloud infrastructure provider access to competitively sensitive information about competitors using the developer's language models, including, for example, information about its competitors' relationships and use of the model provider, prompts and responses generated by the model, and product roadmaps (including planned innovations related to AI functionalities)*". The buyer would then be able to anticipate the competitive behaviour of the target company and react accordingly;

- the coordinated effects can be greater **when a company invests in several competing companies** (such as Microsoft, with model developers OpenAI and Mistral AI), **or when several powerful companies in the digital sector invest in the same target company** (for example, investments by Amazon and Google in Anthropic). As one stakeholder said, "*when a major player holds stakes in several competing companies, this can, depending on the rights held by that player, create conflicts of interest and potentially reinforce lock-in with certain suppliers. This can limit competition by giving the player an unfair competitive advantage, restricting access to key technologies or resources, or influencing the strategic decisions of the companies in which the player holds stakes. This can also lead to a concentration of power and a reduction in competition, which will be detrimental to innovation and consumers*";
- lastly, Microsoft's **recruitment** of a large number of Inflection employees led to discussions as to whether this type of practice could be considered a merger[173].

294. During the preparation of this opinion, the *Autorité* confirmed that such partnerships are a source of concern for many stakeholders in the sector:

- for one stakeholder: "[t]*he investments made by the latter* [the GAFAMs] – *even if they are minority investments – in emerging technologies lead* de facto *to market foreclosure, with domination in the upstream market largely benefiting the downstream market. Here too, Microsoft's position and its investments in multiple related markets distort the balance of power.* [...] *Microsoft leverages its strong position to protect its position in generative AI, a key field for strengthening its position both in the collaborative solutions market and in the search engine market, thus penalising minority players in both the search and generative AI markets*";
- for another stakeholder: "*while in the short term, the agreements have the advantage of facilitating access to essential resources and deployment channels for start-ups, they could also leave them dependent on the partner company's chips and/or infrastructure. Furthermore, the partner company's cloud platform could try to establish itself as the exclusive deployment channel for certain foundation models. This could happen, for example, if current minority investments in established companies become majority investments or are transformed into fully-fledged acquisitions*".

b) Certain investments may be examined *ex ante* under merger control rules

Transactions are subject to prior authorisation if they give investors de facto *control and exceed EU and national notification thresholds*

295. A minority investment is subject to prior authorisation under merger control rules if the investment confers a lasting change of control within the meaning of Article 3(1) of EC Regulation 139/2004 on merger control, i.e. the ability to exercise **decisive influence** over the target's strategy. French law adopts the same approach in Article L. 430-1 of the French Commercial Code (*Code de commerce*), defining the notion of control. Section III of said Article stipulates that "*control derives from rights, agreements or other means which, alone or in combination and having regard to the circumstances in fact or in law, confer the possibility of exercising decisive influence on the activity of a company*". Once control has been established, the target company's revenues must exceed EU and national notification thresholds.

[173] Mlex, Microsoft's AI hires resemble 2017 case evading merger veto, Germany's Mundt says, 9 April 2024.

296. According to the *Autorité*'s merger control guidelines, a minority investment may enable a shareholder to exercise decisive influence if the investment is accompanied by rights that go beyond what is normally granted to minority shareholders in order to protect their financial interests, or if those rights, examined according to the body of evidence method, are such as to demonstrate the existence of decisive influence[174]. Special rights conferring a decisive share in the company's decisions (such as veto rights[175]), specific shareholders' agreements or the possibility of appointing certain executives to the company's governing bodies can therefore grant control over the target company, within the meaning of merger regulations. In exceptional cases, a company may have a decisive influence without any shareholding[176].

297. **Beyond the question of decisive influence, the *Autorité* can also take into account the economic links between companies and situations of *de facto* control, such as being the main or even the only active shareholder, either in its own sector or in related sectors, while the other shareholders are financial investors, for example, or the existence of privileged commercial relationships such as exclusive commercial contracts** (paragraph 48 of the aforementioned guidelines).

298. In the case of generative AI, over and above the scale of the investment in the target company's capital, particular attention could be paid to the special influence of major digital companies, which distinguishes them from other investor profiles such as venture capital funds or public bodies. In addition, exclusivity agreements relating to the provision of cloud services or to marketing channels for the target company's products and services could also be examined to determine whether the buyer has decisive influence on the target's strategy.

299. Under of Article L. 430-8, I and II of the French Commercial Code (*Code de commerce*), failure to notify, and early completion of a transaction, can each be sanctioned by a fine of up to 5% of the revenues of the company responsible for notification. The companies concerned must therefore be vigilant in the event of changes to their capital or the agreements entered into in the context of investments and partnerships.

---

[174] See paragraphs 35 *et seq.* of the *Autorité*'s guidelines.

[175] *Autorité* Decision 11-DCC-10 of 25 January 2011 regarding the acquisition of sole control of the Parisot group by Windhurst Industries and the acquisition of joint control of Windhurst Industries by the FSI.

[176] For example, in Opinion 91-A-09 of 15 October 1991, the *Conseil de la concurrence* considered that although Gillette's holding of bonds convertible into shares in Eemland, which owned the Wilkinson brand, and the existence of agreements enabling Gillette to influence the commercial policy of the Wilkinson brand, did not confer any voting rights, they gave Gillette decisive influence over Eemland and should be analysed as a merger.

An investment may also be examined by competition authorities below the notification thresholds

300. The new Commission guidance on the referral mechanism under Article 22 of Regulation 139/2004 provides an appropriate response for examining transactions that do not have an EU dimension and escape the control of national competition authorities due to their national law, despite the potential harmful effects on competition. The CJEU has been called upon to rule on the validity of the updated interpretation and the application of Article 22. In his Opinion of 21 March 2024 on the Illumina/Grail case, Advocate General Emiliou took the view that Article 22 does not empower the Commission to accept such referrals[177]. If the CJEU were to follow his conclusions, the issue would have to be reconsidered.

301. In its Towercast judgement of 16 March 2023, the CJEU held that Regulation 139/2004 did not preclude a merger with no EU dimension, below the jurisdictional thresholds for *ex ante* control under national law, and which has not been referred under Article 22 of said Regulation, from being analysed by a competition authority of a Member State "*as constituting an abuse of a dominant position prohibited under Article 102 TFEU, in the light of the structure of competition on a market which is national in scope*"[178].

Investments may also be examined as part of the analysis of a merger

302. In specific cases, even if they do not meet the criterion of decisive influence, investments may be examined as part of the analysis of the effects of a merger.

303. In 2022, the *Autorité* ruled for the first time that a non-controlling minority stake acquired at the same time as the acquisition of sole control was likely to harm competition. As part of the acquisition of Bio Pôle Antilles by the Inovie group[179], the latter had informed the *Autorité* of its intention to acquire a minority stake in Synergibio, Bio Pôle Antilles' only private competitor in Guadeloupe and Saint-Martin. The *Autorité* concluded that the acquisition would not give Inovie any controlling rights enabling the group to exercise decisive influence over Synergibio. However, the *Autorité* considered that, given its sufficiently certain nature, the proposed acquisition of a non-controlling minority stake should be taken into account when analysing the transaction's effects at local level. Under these conditions, and to obtain merger clearance for its first transaction with Bio Pôle, Inovie committed to refrain from acquiring any stake in the capital of Synergibio for a period of 10 years.

304. In addition, minority investments made prior to a merger may also be examined. In the Carrefour/Promodes case[180], the Commission noted that Carrefour held a 42% stake in GMB, which controlled Cora. In response to the Commission's competition concerns regarding the ability of Cora and Casino to counterbalance Carrefour/Promodes, Carrefour committed to divest its stake in GMB.

---

[177] Opinion of the Advocate General in Joined Cases C-611/22 P | Illumina v. Commission and C-625/22 P | Grail v. Commission and Illumina, paragraph 265.

[178] CJEU judgement, 16 March 2023, Towercast, C-449/21, paragraph 53.

[179] See *Autorité* Decision 22-DCC-35 of 27 April 2022 regarding the acquisition of exclusive control of Bio Pôle Antilles by the Inovie group.

[180] Case no. COMP/M. 1684 – Carrefour/Promodes.

A lack of transparency regarding investments and partnerships

305. Despite the numerous competitive risks highlighted above, **the authorities do not always have the information** needed to determine whether agreements are likely to harm competition and hence consumers.

306. In 2014[181], the Commission proposed the introduction a mandatory information system in cases where the transaction would create a "*competitively significant link*". To guarantee legal certainty for the parties, only transactions meeting the following cumulative criteria would be concerned: "*acquisitions of a minority shareholding in a competitor or vertically related company* [...] *and the competitive link would be considered significant if the acquired shareholding is (1) around 20% or (2) between 5% and around 20%, but accompanied by additional factors such as rights which give the acquirer a 'de-facto' blocking minority, a seat on the board of directors, or access to commercially sensitive information of the target*"[182]. The initiative was eventually shelved during the following Commission term.

307. In the generative AI sector, some minority investors can play a more decisive role than the title of certain partnerships suggests, as demonstrated by Microsoft's involvement in OpenAI's governance changes in November 2023, despite the absence of formal decision-making power within the company. Recent statements by Microsoft's Chairman and CEO seem to confirm that the partnership gives important rights to Microsoft[183]. In response to such developments, the President of the German *Bundeskartellamt* questioned whether some of the agreements are not in fact "*mergers in all but name*"[184].

[181] European Commission, White Paper – Towards more effective EU merger control, 9 July 2014.

[182] Above-cited White Paper, paragraph 47.

[183] Discussing Microsoft's investment in OpenAI in a press article, Mr A..., Microsoft's Chairman and CEO, said: "[i]*f OpenAI disappeared tomorrow, I don't want any customer of ours to be worried about it quite honestly, because we have all of the rights to continue the innovation. Not just to serve the product, but we can go and just do what we were doing in partnership ourselves. We have the people, we have the compute, we have the data, we have everything. But at the same time, I'm committed to the OpenAI partnership and that's what I expressed to them* [...]. *And also this thing, it's not hands off, right? We are in there. We are below them, above them, around them*".

[184] According to Andreas Mundt: "*Big Tech partnerships with startup AI developers such as OpenAI should put competition authorities on alert to the idea that cooperation agreements could be mergers in all but name*" (Mlex, Watch out for AI cooperation agreements that are really mergers, Germany's Mundt warns, 21 September 2023).

308. These concerns are shared by competition authorities around the world, as evidenced by the ongoing investigations by the Commission[185] and the CMA in the United Kingdom[186] into Microsoft's investments in OpenAI, the investigations launched in early 2024 by the FTC in the United States into Alphabet, Amazon, Anthropic, Microsoft and OpenAI[187], and the call for submissions launched by the CMA in April 2024[188]. The *Bundeskartellamt* also examined the cooperation between Microsoft and OpenAI in 2023. While the German authority concluded that the cooperation in question did not fall within the scope of national merger control, given the absence of a sufficient link between the transaction and German territory, it nevertheless confirmed that Microsoft's influence over OpenAI constituted a merger within the meaning of the law.

c) Investments may be covered by competition law

309. If minority investments do not give the buyer control over the target company (and therefore do not meet merger law criteria), they can be assessed *ex post* through competition law, notably on the basis of antitrust law or abuse of dominant position.

310. This possibility was confirmed by the CJEU in an old case. In the Philip Morris case, the CJEU considered[189] whether, and where applicable under what conditions, the acquisition of a minority stake in the capital of a competing company could constitute an infringement of Articles 85 and 86 of the Treaty (now Articles 101 and 102 TFEU). After recalling that a company acquiring a stake in the capital of a competing company does not in itself constitute anticompetitive behaviour under competition law, the CJEU stated that "*such an acquisition may nevertheless serve as an instrument for influencing the commercial conduct of the companies in question so as to restrict or distort competition on the market on which they carry on business*" (paragraph 37). According to the CJEU, that would in particular be the case "*where* […] *the investing company obtains legal or* de facto *control of the commercial conduct of the other company or where the agreement provides for commercial cooperation between the companies* [...]. *That may also be the case where the agreement gives the investing company the possibility of reinforcing its position at a later stage and taking effective control of the other company*" (paragraphs 38-39). The CJEU then indicated the need for particular vigilance, by examining in particular "*whether an agreement which at first sight provides only for a passive investment in a competitor is not in fact intended to result in a take-over of that company, perhaps at a later stage, or to establish cooperation between the companies with a view to sharing the market*" (paragraph 45). In the end, the CJEU rejected the appeal in its entirety.

---

[185] European Commission press release of 9 January 2024: "*the European Commission is checking whether Microsoft's investment in OpenAI might be reviewable under the EU Merger Regulation*".

[186] On 8 December 2023, the CMA opened a merger control investigation into the partnership between Microsoft and OpenAI.

[187] FTC Launches Inquiry into Generative AI Investments and Partnerships, 25 January 2024.

[188] CMA seeks views on AI partnerships and other arrangements, 24 April 2024.

[189] CJEU judgement, 17 November 1987, British-American Tobacco Company Ltd and R.J. Reynolds Industries Inc. v. Commission, Joined Cases 142 and 156/84, paragraph 37 *et seq*.

311. Minority investments could therefore potentially be examined from the angle of abuse of dominant position, including on a collective basis.

312. To demonstrate the existence of collective dominance, it must be established that the companies "*together, in particular because of factors giving rise to a connection between them, are able to adopt a common policy on the market and act to a considerable extent independently of their competitors, their customers, and also of consumers*"[190], which may be demonstrated by examining the legal links or factors of connection between the companies, or by examining the structure of the market in accordance with the criteria set out by the Court in the Airtours judgement[191]. Therefore, the existence of structural links between companies, such as capital ties or formal agreements between them, on the one hand, and the adoption of a common line of action in the market, on the other, are sufficient to demonstrate the existence of collective dominance. Investments by several companies in the same target could therefore be examined on this basis. In this respect, the *Autorité* recalls that collective dominance is not in itself problematic, only abuse of that position is.

313. Inter-company agreements could also fall within the scope of competition law, for example if they are aimed at sharing markets or promoting market transparency.

314. The *Autorité* recently applied the above-cited Towercast case law, examining whether a merger below review thresholds constituted an anticompetitive practice contrary to the TFEU, in this case an agreement contrary to Article 101[192] (see box below).

[190] See, for example, *Autorité* Decision 20-D-11 of 9 September 2020 regarding practices implemented in the sector for the treatment of age-related macular degeneration (AMD), and *Autorité* Decision 12-D-06 of 26 January 2012 regarding practices implemented in the aggregates sector and downstream markets in Saint-Pierre-et-Miquelon.

[191] General Court of the European Union, T-342/99, Airtours v. Commission, 6 June 2002, point 62.

[192] *Autorité* Decision 24-D-05 of 2 May 2024 regarding practices implemented in the rendering sector.

Decision 24-D-05 of 2 May 2024 regarding practices implemented in the meat-cutting sector

In the notification of objections, the Investigation Services accused Akiolis, Saria and Verdannet of creating and implementing a geographic market allocation agreement, ultimately achieved through the cross-divestiture of business assets.

In this case, the *Autorité* analysed whether the mergers, which had not been notified *ex ante* under EU or national merger control, were likely, on their own, to constitute an anticompetitive agreement contrary to Article 101 TFEU and Article L. 420-1 of the French Commercial Code (*Code de commerce*). The respondents argued that the Towercast case law only concerned the applicability of Article 102 TFEU and could therefore not be transposed to Article 101 TFEU. They also argued that applying antitrust law to a merger required an anticompetitive practice that could be separated from the merger to be identified.

The *Autorité* considered that, "*in accordance with CJEU case law, a merger which 'has no Community dimension within the meaning of Article 1 hereof* [the Merger Regulation]*, is below the thresholds for mandatory* ex ante *control laid down by national law, and has not been referred to the Commission under Article 22 of that regulation' may be subject to* ex post *control based on Article 101 TFEU and Article L. 420-1 of the French Commercial Code (*Code de commerce*)*". It was the first time that the *Autorité* had examined, under antitrust law, mergers below national notification thresholds. The case was eventually dismissed.

3. THE RISK OF COLLUSION BETWEEN COMPANIES IN THE SECTOR

315. The use of generative AI could have consequences for the potential implementation of concerted practices[193].

316. Almost all the stakeholders consulted during the public consultation did not express any specific concerns about the risk of collusion. Moreover, the majority of concerns relate to the downstream part of the value chain, and therefore fall outside the scope of this opinion. Nevertheless, it is worth citing the example of Samsung, which recently banned its employees from using generative AI tools such as ChatGPT after discovering that staff members had uploaded sensitive code to the platform, which could lead to the disclosure of the information to other users[194]. This example shows why competition authorities should pay attention to collusion.

[193] According to case law, "concerted practices" can be distinguished from "agreements between undertakings": "*the object is to bring within the prohibition of that article* [Article 101 TFEU] *a form of coordination between undertakings which, without having reached the stage where an agreement properly so-called has been concluded, knowingly substitutes practical cooperation between them for the risks of competition*" **(**CJEU judgement, 14 July 1972, Imperial Chemical Industries Ltd. v. Commission of the European Communities, Case 48-69, paragraph 64).

[194] Bloomberg, Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak, 2 May 2023, quoted in C. Carugati, "Competition in generative artificial intelligence foundation models", Working Paper 14/2023, Bruegel, 2023.

317. The joint study by the *Autorité* and the German *Bundeskartellamt* on "Algorithms and competition", to which this opinion refers[195], deals specifically with the issues of collusion between algorithms and the legal framework likely to apply to these issues. As a reminder, the majority of current language models are developed using the same deep learning algorithm, Transformer. The findings of the joint study can therefore be extended to generative AI algorithms, and several situations could lead to competition risks, in particular:

- generative AI algorithms can be a means of supporting or facilitating pre-existing anticompetitive practices (such as a cartel);
- collusion can be based on an algorithm between competitors involving a third-party hub and spoke, where a third party, such as an external consultant or software developer, provides competitors with the same algorithm or coordinated algorithms, without any direct communication between the different competitors;
- collusion can be induced by the parallel use of separate individual algorithms, or by the use of machine learning algorithms. In the latter situation, the algorithms can teach themselves to converge towards a collusive equilibrium.

318. One stakeholder summed up the questions that are likely to arise, particularly with regard to liability: "*How can we assess the risks of collusion induced by the cross-use of generative AI by companies in the same market, particularly in terms of transparency and the exchange of sensitive information? How can we define the associated chain of responsibility for collusion when human decision-making is gradually being replaced by AI*?".

[195] *Autorité* and *Bundeskartellamt* study, Algorithms and competition, 6 November 2019.

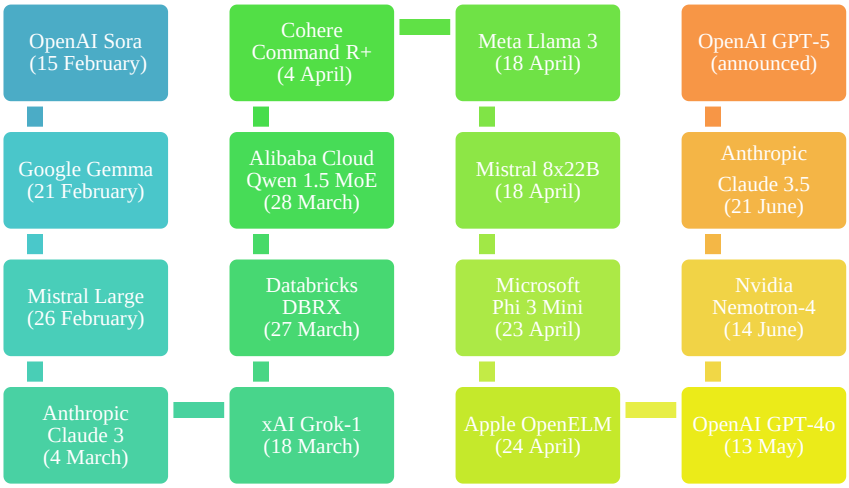
III. Outlook and recommendations

A. THE GENERATIVE AI SECTOR IS FAR FROM HAVING REACHED ITS POTENTIAL

319. The generative AI sector is booming. Less than two years after the launch of ChatGPT, many established operators have invested in this field and a multitude of start-ups have emerged to accelerate research and enable the use of this innovative technology by the majority of companies and consumers. At the beginning of 2024, OpenAI passed the $2 billion mark in revenues (€1.85 billion), most of which has been achieved since December 2022.

320. Many generative AI models were announced in the first half of 2024 (including several during the preparation of this opinion – Mistral Large by Mistral AI, Claude 3 by Anthropic and Llama 3 by Meta), underlining the dynamism and volatility of the market. **The race to innovate** and develop new generative AI models is likely to continue along at least two aspects: model size (the larger the model, the better it performs) and model optimisation at constant size.

Figure 9: Main models published between 8 February 2024 and 24 June 2024



Source: Autorité de la concurrence.

321. Although the larger models seem to have the best performance, they are not suitable for all applications. Many operators therefore prefer smaller, less expensive models. The question of the portability of generative AI models on less powerful media and without GPUs is also arising and should lead to competition between operators. Apple and Samsung have already announced the future integration of generative AI tools on their mobile phones.

322. Model size is also a key factor in the **environmental impact of generative AI**. Although it is difficult to estimate the additional cost of using generative AI, it is nonetheless certain that this technology will lead to an increase in energy consumption, at least in the short term. Estimates indicate, for example, that a search engine using AI would consume 10 times more energy than one without. Reducing energy impact is therefore another possible area for innovation, given that some operators in the sector have announced their goal of achieving

a carbon-neutral environmental footprint by 2030[196], which will encourage them to innovate to cut costs. This innovation could be technological, at the different stages of model creation, from training to inference; it could concern the ways in which generative AI is used by end users, for example to reduce the frequency of queries; or it could concern access to energy for sector operators. As mentioned above (paragraph 221), it is important to ensure that the growing role of energy does not create new barriers to entry for reasons of cost, or in the event of the vertical integration of certain operators.

323. The development of new forms of computing such as quantum or edge computing could accelerate the development of the generative AI sector and its adoption by users; however, it could also reinforce the competition risks identified in this opinion if these new forms of computing are controlled by major operators, either directly or through partnerships.

324. The *Autorité* has also observed a trend towards "platformisation" in the generative AI sector. For example, OpenAI offers the possibility of adding plugins to ChatGPT via the GPTStore, a marketplace where all developers and companies can offer their specific plugins. On the other hand, Hugging Face is establishing itself as the reference marketplace for the publication of open data and generative AI models. Cloud service providers, and hyperscalers in particular, are providing their customers with MaaS marketplaces to facilitate access to the main generative AI models. These cloud marketplaces are emerging as the only way for model developers to reach consumers or AI-using companies.

325. One of the main challenges for the development of competition in the generative AI sector lies in the deployment of open-source resources. However, as mentioned above, open source in the generative AI sector represents various scenarios, with each operator having its own characteristics and needs. If the sector had more precise criteria for qualifying the degree of openness of a model, operators who so wished could use model openness as a competitive advantage.

326. Companies in the sector, especially start-ups, must also establish themselves on a long-term basis, which requires a remuneration model that enables them to recoup their substantial initial investments while continuing to develop. This issue is particularly relevant for companies turning to open source, as underlined by the PEReN: "*Open source software is by its very nature freely distributable, and therefore* a priori *free of charge, which makes business models based solely on the sale of software a complex proposition*"[197].

327. The analyses and risks identified in this opinion concern the upstream generative AI value chain at constant technological model, i.e. based on LLMs. A new technological step beyond LLMs would require a new competitive analysis.

[196] Les Échos, Comment l'IA plombe le bilan carbone de Microsoft, 16 May 2024.

[197] PEReN, Éclairage sur…n°7 – Open source et IA : des synergies à repenser ?, 3 April 2024. (Summary also available in English: Shedding light on…n°7 – Open source and AI: rethinking synergies?).

B. RECOMMENDATIONS

328. The potential consequences of generative AI for the productivity of public services and companies, for the organisation of work and beyond, and for key public policy priorities such as education and health, call for a regulatory framework that fosters the adoption of AI by households and companies[198] and allows for a diversity of usages and models, while ensuring vigilant control of risks in areas such as national security, privacy and intellectual and artistic property. In view of the above developments, the *Autorité* considers that the following recommendations would strengthen competition in the sector.

329. For the most part, these recommendations do not require any legislative initiative at French or European level.

1. PROPOSALS THAT AIM TO MAKE THE REGULATORY FRAMEWORK APPLICABLE TO THE SECTOR MORE EFFECTIVE, WITH NO CHANGE TO EXISTING REGULATION

330. As explained above (see paragraphs 104 *et seq.*), a number of regulations have been introduced at global, European and French level in recent years. Not all have been implemented yet[199], while others are still in the draft phase[200]. Given the speed of change in the sector, these regulations must be fully implemented and their impact evaluated, in order to avoid negative impacts on innovation and competition.

331. However, improvements could be made with no change to existing legislation. Some regulations were introduced before the emergence of generative AI, so its effects cannot be fully captured by the legislation.

332. For example, the obligations of the DMA can only be applied to the core platform services of gatekeepers mentioned in Article 2 of the DMA, which do not include MaaS.

333. However, it cannot be excluded that, by virtue of the characteristics described in paragraphs 140 *et seq.*, certain MaaS services may fall into one of the categories listed in Article 2, in particular that of "*cloud computing services*"[201], subject to the Commission's interpretation.

---

[198] Above-cited French AI Commission report.

[199] For example, most of the provisions of the AI Act (which is pending publication in the EU Official Journal) will only apply from 2026, with the exception of certain specific provisions. In addition, the abolition of cloud service provider switching fees (known as "egress fees"), provided for under the EU Data Act, will be effective from 12 January 2027.

[200] For example, the Commission published a proposal on adapting non-contractual civil liability rules to AI on 28 September 2022, which is still under discussion. The aim of the proposal is to "[ensure] *victims of damage caused by AI obtain equivalent protection to victims of damage caused by products in general. It also reduces legal uncertainty of businesses developing or using AI regarding their possible exposure to liability and prevents the emergence of fragmented AI-specific adaptations of national civil liability rules".*

[201] To date, the Commission has not designated any company providing cloud computing services as a gatekeeper for those services.

334. In this case, companies providing MaaS services could be designated as gatekeepers for those services, either through designation under Article 3.4 of the DMA if quantitative thresholds are exceeded or, if those thresholds are not exceeded, through designation based on qualitative criteria following a market investigation under Article 3.8 and Article 17 of said Act. In the course of this analysis, it would nevertheless be necessary to verify that the MaaS services really do act as an interface between the companies using their services and the end users[202].

Proposal no. 1: the Commission should pay particular attention to the development of MaaS services to assess the possibility of designating companies providing such services as gatekeepers under the DMA.

335. As indicated by the *Autorité* in aforementioned Opinion 23-A-08 of 29 June 2023, the provisions of the EU Data Act, particularly in terms of data interoperability, will have an overall positive effect on competition in the cloud sector. Nevertheless, other competition risks remain, such as those associated with cloud credits, which in France are covered by Article L. 442-12, II to V of the French Commercial Code (*Code de commerce*) (created by Article 26 of the SREN Law), but not at European level. Article L. 442-12, II limits the duration of cloud credits and prohibits any exclusivity conditions, on pain of a fine imposed by the Minister of the Economy. Article L. 442-12, IV also prohibits making the sale of a product or service conditional on the simultaneous conclusion of a contract for the provision of cloud computing services, where this practice constitutes an unfair trade practice, with the prohibition also carrying an administrative fine. Lastly, according to Article L. 442-12, V, "[the] Autorité de la concurrence *may, either on its own initiative or at the request of the Minister for Digital Affairs or any other legal person concerned, process any reports regarding self-preferencing practices. The* Autorité de la concurrence *sanctions those practices or adopts any necessary measures, where applicable, on the basis of Titles II and VI of this Book. To implement these provisions, the* Autorité de la concurrence *has the powers granted to it under Title V of this Book*".

336. The *Autorité* recommends that, in applying these new provisions, the DGCCRF pay particular attention to practices more specifically concerning the AI field. For its part, the *Autorité* is committed to vigilance with regard to the self-preferencing practices referred to in section V above.

Proposal no. 2: at the French level, in implementing the provisions of the SREN Law on cloud credits, the DGCCRF should pay particular attention to the use of such credits in AI.

337. Lastly, close attention should be paid to the effects of the **EU AI Act** (see paragraphs 104 *et seq.*) on competition the sector. The Act subjects suppliers of generative AI systems to a number of regulatory obligations, which require the mobilisation of substantial financial, human and technical resources likely to hinder the emergence or expansion of smaller operators. For example, it is essential to ensure that major operators do not use certain provisions of the AI Act to further consolidate their market power. In addition, vigilance is required as regards the exemptions from the Act for open-source models, insofar as these models can vary significantly (see paragraphs 179 *et seq.*).

[202] If, on the contrary, it were considered that MaaS and other services upstream in the AI value chain were not covered by Article 2, the Commission could launch a market investigation under Article 19, in order to expand the list in Article 2, which would then require a revision of the Act.

Proposal no. 3: the future AI Office, established under Article 64 of the AI Act, and the competent national authority in France, which will be designated in accordance with Article 70 of said Act, should ensure, on the one hand, that the implementation of the Act does not hinder the emergence or expansion of smaller operators and, on the other hand, that the largest operators in the sector do not misuse the text to their advantage.

338. Lastly, international coordination is necessary, given the various initiatives underway in France, Europe and the rest of the world, to ensure that such initiatives do not create distortions and additional costs for companies. The AI Summit to be hosted by France in February 2025 will be an opportunity to strengthen global AI governance.

2. USING THE TOOLS OF COMPETITION LAW AND THE LAW ON RESTRICTIVE COMPETITION PRACTICES

339. In view of the risks identified above, the use of competition tools will play an essential role in preventing the emergence or consolidation of dominant positions or agreements that would affect competition in the sector.

340. Competition authorities must remain committed and attentive to ensuring that no operator is able to foreclose access to essential inputs for the development of generative AI, while giving markets the opportunity and incentive to continue to develop and innovate. To this end, the *Autorité* has already launched several initiatives, such as issuing an opinion on the competitive functioning of the cloud sector and analysing the role of press publishers' data at the grounding stage as part of the "Google related rights" case.

341. The *Autorité* will remain vigilant to developments in the sector, including the situation of graphics cards (following the dawn raid conducted in the sector in 2023), the agreements between digital giants and content providers (including sensitive data such as financial data), and the competition risks associated with the deployment of models on separate markets.

342. The *Autorité* has the tools required to act quickly and effectively. The same instruments are available **at European level**, although the conditions of their application may differ[203].

343. In the event of a situation requiring immediate action, the *Autorité* may order **interim measures** pending a decision on the merits of the case, if there is serious and immediate harm to the interests of an economic sector or company. The practices of a company in a dominant position may give rise to interim measures because the company must be prevented, during the structuring phase of new markets, from taking too great a technological lead over its competitors[204] or reinforcing the oligopolistic structure of the market[205]. In recent years, the interim measures procedure has enabled decisions to be made within short timeframes, often less than six months. Pending the latest regulatory texts becoming fully effective, using interim measures to safeguard the conditions of competition in the sector may seem particularly relevant.

[203] At European level, the standard of evidence for interim measures is stricter, which has led to a more limited use of interim measures by the Commission.

[204] *Conseil de la concurrence* Decision 00-MC-01 of 18 February 2000 regarding a request for interim measures submitted by 9 Télécom Réseau.

[205] *Autorité* Decision 23-MC-01 of 4 May 2023 regarding the request by Adloox for interim measures.

344. The *Autorité* may also decide to **reach a settlement** with the company or companies in question, in order to speed up the processing of a case. The settlement procedure has been used in particular in the market for advertising servers for publishers of online sites and mobile applications[206].

345. The *Autorité* may also impose a **fine** and/or **behavioural** or **structural injunctions** aimed at putting an end to the practices in question or forcing the company concerned to modify its behaviour. These injunctions can take a variety of forms, ranging from an injunction to negotiate in good faith, as in the aforementioned "related rights" case, to the modification of a platform's operating rules[207] or the cessation of all discrimination[208].

346. In appropriate cases, the *Autorité* may choose a negotiated solution whereby the company proposes structural and/or behavioural **commitments** that are made binding, provided the commitments remedy the competition concerns[209]. The remedies procedure enables certain situations to be resolved quickly and early. It also spares the *Autorité* of conducting a time-consuming investigation, freeing up resources for other cases. For example, the *Autorité* accepted Google's commitments to create a framework for negotiating and sharing the information needed for a transparent assessment of the remuneration of related rights[210], and Meta's commitments to put an end to practices likely to raise competition concerns in the French market for online non-search advertising[211].

347. The *Autorité* therefore already has a toolbox for taking effective action against the practices in question, if necessary using the instruments in isolation, simultaneously or sequentially, provided its resources are sufficient.

348. **Restrictive competition practices**, which are primarily the responsibility of the DGCCRF and the commercial courts, can also be an appropriate response to the risks observed in the sector. The provisions relating to restrictive competition practices (Title IV of Book IV of the French Commercial Code [*Code de commerce*]) have been applied to the digital economy, with certain contractual practices by digital platforms sanctioned in recent years (see paragraphs 612 *et seq*. of aforementioned Opinion 23-A-08). To implement these provisions, the authorities and the courts must have the necessary resources.

Proposal no. 4: the authorities responsible for enforcing competition in the markets must remain vigilant in the generative AI sector and, if necessary, use all the tools at their disposal to act swiftly and effectively.

---

[206] *Autorité* Decision 21-D-11 of 7 June 2021 regarding practices implemented in the online advertising sector.

[207] *Autorité* Decision 19-D-26 of 19 December 2019 regarding practices implemented in the online search advertising sector.

[208] *Autorité* Decision 14-D-06 of 8 July 2014 regarding practices implemented by Cegedim in the medical information databases sector.

[209] According to the *Autorité*'s "Behavioural remedies" study of 17 January 2020, "*this dichotomy is traditionally based on the effects generated by the commitments, the first directly modifying the structure of the markets (the number, quality or scope of operators active on a market) by themselves, and the second limited to regulating the behaviour of the parties undertaking them. As a result, when commitments impose the divestiture of assets or a breaking of contractual ties in order to maintain an independent offer on the market, they are considered to be "structural". If they restrict the commercial or strategic behaviour of a company, on the other hand, they are qualified as "behavioural*" (page 262).

[210] *Autorité* Decision 22-D-13 of 21 June 2022 regarding practices implemented in the press sector.

[211] *Autorité* Decision 22-D-12 of 16 June 2022 regarding practices implemented in the online advertising sector.

3. ENSURING ACCESS TO COMPUTING POWER TO ENCOURAGE INNOVATION

349. Access to computing power is essential for the development of research and the emergence of new companies in the generative AI sector.

350. A number of initiatives are underway in the IT components market, notably in Europe with SiPearl, a start-up that aims to equip European supercomputers with its CPUs. Although it does not provide a direct solution for the specific sector of generative AI, which requires more specialised chips than CPUs, the example of SiPearl demonstrates the ability of European initiatives to meet industrial needs.

351. The stakeholders consulted considered access to computing power to be a competition issue, in terms of both time (in a highly dynamic market) and cost, but one that is likely to become less acute with the emergence of greater competition in the market for IT components for AI.

352. The *Autorité* stresses the importance of the availability of public computing resources, via supercomputers, accessible free of charge to operators in return for a contribution to open science. Several public players have recently called for European computing power to be strengthened[212].

> **Proposal no. 5**: continue to invest in the development of supercomputers at European level, to give as many parties as possible access to computing power.

353. Given the dynamism of the market and the need for the latest-generation chips for model training, supercomputers require ongoing investment to remain a viable alternative for training generative AI models and/or fine-tuning pre-trained models. These investments could be financed, at least in part, by private operators using computing resources.

> **Proposal no. 6**: the government and/or companies responsible for managing supercomputers could look into how to propose an open, non-discriminatory framework that would enable companies to use public supercomputer resources for a fee, while maintaining priority for research, particularly academic research.

354. In view of the growing demand for supercomputer resources, emphasis must be placed on the open nature of AI models trained on public supercomputers, and priority given to the projects with the most open strategy, particularly with regard to the open source criteria set out in the EU AI Act[213].

> **Proposal no. 7**: in conjunction with the AI Act in particular, set criteria for the openness of generative AI models trained on public supercomputers.

---

[212] See the above-cited French AI Commission report, Emmanuel Macron's 22 May 2024 speech on the fringes of the opening of the VivaTech trade show, and European Internal Market Commissioner Thierry Breton's blog post on AI in Europe in September 2023.

[213] Article 53(2) of the AI Act states, "*the obligations set out in paragraph 1, points (a) and (b), shall not apply to providers of AI models that are released under a free and open-source licence that allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available* [...]".

4. IN THE DATA MARKET, ENSURE A BALANCE BETWEEN FAIR REMUNERATION FOR RIGHTS HOLDERS AND ACCESS FOR MODEL DEVELOPERS TO THE DATA NEEDED TO INNOVATE, TAKING INTO ACCOUNT THE DIVERSITY OF DATA USAGE CASES

355. The *Autorité* has noted the concerns both of AI model developers regarding access to the data needed to train and use their models, and of rights holders, such as press publishers and agencies, regarding respect for their rights. It is important to strike a balance between these two considerations to ensure the sustainability of a model based, at constant technology, on the use of an ever-increasing quantity of data.

356. Data uses differ depending on the stage at which the data is used within the value chain, whether upstream of model training or downstream, with techniques such as grounding where the content created by the model is enhanced with external knowledge, such as press articles. Agreements between rights holders and developers should therefore reflect the relative importance of the data for the developers according to the use case, and specify in which circumstances the data may be used.

357. For example, press publishers' data is essential for operators that deploy chatbots for users with grounding and therefore has a high economic value in this specific case. However, for the training of generative AI models, its marginal importance is relative in the considerable volume of data required, which also places great emphasis on descriptions of facts and logical reasoning, for example from encyclopaedias and scientific articles.

358. It follows from the *Autorité*'s analysis that the value of data seems to be linked, in the current state of technology:

- at the training stage, to the volume and descriptive value of the data;
- at the fine-tuning stage, to the specificity of the data. Sector-specific data will therefore be more valuable for a model wanting to specialise in the sector in question (e.g. healthcare data);
- at the inference stage, particularly for RAG or grounding, to the relevance and timeliness of the data, i.e. the ability to provide missing information that a model may not have integrated during its training phase.

359. Transaction costs are also an important consideration with regard to training data. While transaction costs are usually negligible compared with acquisition prices, they could prove prohibitive in the case of training data if a model developer has to contract individually with each operator whose data it wants to use. In this respect, the *Autorité* notes that proposals are emerging from some analysts, such as the introduction of collective licences or the granting of a "safe harbour" that would protect certain model providers from legal liability, provided they meet certain transparency and ethics standards[214].

[214] TechCrunch, OpenAI's deals with publishers could spell trouble for rivals, 13 March 2024.

Proposal no. 8: public authorities, in particular as part of the mission entrusted by the French Ministry of Culture to the French Higher Council for Literary and Artistic Property (*Conseil supérieur de la propriété littéraire et artistique*), could encourage rights holders to take account of the economic value of data according to the use case (for example, by introducing differentiated pricing), and to propose bundled offers to reduce transaction costs, in order to safeguard the innovation capacities of model developers.

360. The *Autorité* also believes that data openness plays an effective role in stimulating competition in the sector, by lowering barriers to entry and reducing uncertainty regarding access to data (see above). Where possible, the public sector should play a leading role in making public authorities' data more open via the data.gouv.fr portal and in the above-mentioned call for digital commons project. For example, the French National Audiovisual Institute (*Institut national de l'audiovisuel* – INA) and the French National Library (*Bibliothèque nationale de France* – BNF) have massive datasets that could be made available to model developers, under conditions to be defined. The private sector can also make a contribution, notably by providing specific data, for example sector-specific data, which can be used for model fine-tuning.

361. Such initiatives can also help to ensure better representation of French (and European) language and culture among generative AI models, where English currently predominates. The use of non-English language data will improve model performance in those languages, while taking better account of cultural diversity, benefiting innovation and end users.

Proposal no. 9: make public and private data available for the training or fine-tuning of generative AI models, and encourage public and private initiatives to distribute French-language data, whether text, image or video.

5. GREATER TRANSPARENCY ON INVESTMENTS BY DIGITAL GIANTS IN INNOVATIVE COMPANIES IN THE SECTOR SEEMS JUSTIFIED

362. The *Autorité* considers that, pending a decision by the CJEU on Article 22, the existing legal framework can be used to address most competition concerns regarding agreements between companies, whether through merger control rules or through competition law. Nevertheless, the *Autorité* considers that, even with no change to existing legislation, there should be greater transparency in non-controlling minority investments in the sector.

363. Although the DMA does not directly refer to generative AI services as core platform services, Article 14 of the DMA has a broad scope, as it applies to any proposed merger where the merging entities or the target of the merger provide core platform services or any other service in the digital sector or enabling the collection of data. In line with current merger control practice, the information document provided for under Article 14[215] could include an obligation to provide information on minority investments in the same sector as the target. A gatekeeper informing the Commission of a merger in the generative AI sector would therefore also inform the Commission of any minority investments in the same sector as its target.

364. This proposal would be without prejudice to the *ex post* control of non-controlling minority investments under competition law rules on abuse of dominant position and anticompetitive agreements.

[215] The Article 14 template dated 27 October 2023 is available here.

Proposal no. 10: the Commission could request further information on minority investments in the same sector as the target, in the template relating to the obligation to inform about a concentration pursuant to Article 14 of the DMA.

365. Another possibility, which is stricter and goes beyond just the generative AI sector, would be to modify Article 14 of the DMA so that the Commission is systematically informed of such minority investments.

366. This information obligation could include conditions similar to those envisaged by the Commission in 2014 (see paragraph 306 above), which would limit information obligations to potentially problematic transactions and avoid placing a disproportionate administrative burden on companies and the Commission. It could also include requests for additional information, such as any exclusivity agreements between the parties. In addition, information received under Article 14 could be useful for the application of the rules on abuse of dominant position and anticompetitive agreements.

Conclusion

367. The *Autorité* has closely examined the generative AI sector, focusing in particular on the upstream part of the value chain, which is likely to pose more competition risks.

368. In addition to an in-depth analysis of the sector and the identification of the inputs needed to develop foundation models, the *Autorité* has taken a position on issues of particular importance for the future, such as practices likely to be implemented in labour markets, the challenge of content remuneration, and the competitive assessment of minority investments by major companies in innovative start-ups.

369. The *Autorité*, which acts in an advisory and non-legal capacity in this opinion, does not rule on the lawfulness of the aforementioned practices. Nevertheless, the competition risks set out in this opinion will be carefully monitored by its departments, particularly with regard to practices that unreasonably restrict access to essential inputs, partnerships by already dominant digital companies, with or without exclusivity clauses, and tied or bundled selling practices likely to consolidate the generative AI sector around such companies in the long term, without prejudice to practices downstream in the value chain, which are not the subject of this opinion. Such vigilance is essential for contributing to the development of an open, rights-respecting AI, in which smaller operators have a chance to succeed and companies and users have access to varied, innovative models.

Deliberated on the oral report by Elodie Vandenhende and Quentin Deltour, case officers (*rapporteurs*), and the contribution of Yann Guthmann, Head of the Digital Economy Unit, by Benoît Cœuré, President, Fabienne Siredey-Garnier, Irène Luc and Thibaud Vergé, Vice-Presidents, and Valérie Bros, Julie Burguburu, Catherine Prieto and Jérôme Pouyet, members.

Hearing secretary,	The President,
Caroline Orsel	Benoît Cœuré

Glossary

- **AI accelerator:** integrated circuit designed and optimised for AI workloads.
- **Application programming interface (API):** software interface that enables one software programme or service to "communicate" with another software programme or service, in order to exchange data and functionalities (definition from CNIL glossary).
- **Artificial intelligence:** any tool used by a machine "*to display human-like capabilities such as reasoning, learning, planning and creativity*" (definition from European Parliament).
- **Automated machine learning (autoML):** automatic generation of optimised learning models, which can be used by non-expert users.
- **Central processing unit (CPU):** central processor.
- **Common Crawl:** a non-profit organisation founded in the United States in 2007, whose mission is to provide free web archives. Some 100 Internet crawls have been carried out since 2008.
- **Compute Unified Design Architecture (CUDA):** proprietary software developed by Nvidia to enable programming on its own GPUs.
- **Computer vision:** field of AI whose main aim is to enable machines to analyse and process one or more images or videos captured by an acquisition system (definition from CNIL glossary).
- **Crawl:** automatic collection of web page content by a robot.
- **Data labelling:** process of identifying raw data (images, text files, videos, etc.) and adding one or more informative and relevant labels to provide context for the learning model.
- **Deep learning:** an automatic learning process that uses neural networks with multiple layers of hidden neurons. As the algorithms have a large number of parameters, a very large amount of data is required for training (definition from CNIL glossary).
- **Fine-tuning:** technique that consists of adapting a pre-trained AI model to a specific task, which generally involves training the model as a whole, or just certain layers of a neural network, for a small number of iterations on a specific dataset corresponding to the target task (definition from CNIL glossary).
- **Floating-point operations per second (FLOPS):** measure of computing power.
- **Foundation model:** large-scale AI model, trained on large datasets and adaptable for different downstream tasks.
- **Framework:** coherent set of software components used to create the foundations and architecture of a software product.
- **French Centre of Expertise for Digital Platform Regulation (*Pôle d'expertise et de régulation du numérique* – PEReN):** a national department under the joint authority of the French Ministers of Economy, Culture and Digital Affairs.
- **Generative artificial intelligence:** type of AI capable of generating new content (text, image, sound, video, etc.).

- **Generative pre-trained transformer (GPT):** pre-trained neural network based on the Transformer architecture.
- **Graphics processing unit (GPU):** processor composed of numerous specialised cores. GPUs enable image computation functions to be performed in parallel and are generally found on graphics cards (definition from *Autorité* Opinion 23-A-08).
- **Hyperparameter:** variable governing the training process itself and set by the developer, such as: the number of nodes in each layer, the number of hidden layers in a neural network, weight initialisation, the learning rate, activation functions, the number of times each piece of data will be used during training, etc.
- **Inference:** process by which a trained model is used to make predictions on new data, after its learning phase. In the context of generative AI, the predictions correspond to content production.
- **Infrastructure-as-a-Service (IaaS):** cloud computing service consisting of the provision of processing, storage, networks and other fundamental computing resources where the consumer is able to deploy and run arbitrary software (definition from *Autorité* Opinion 23-A-08, page 183).
- **Large language model (LLM):** text generative AI model with a large number of parameters.
- **Low Rank Adaptation (LoRA):** fine-tuning technique introduced by a team of Microsoft researchers in 2021, requiring less computing power.
- **Machine learning (ML):** field of study in AI that aims to give machines the ability to "learn" from data, using mathematical models (definition from CNIL glossary).
- **Mixture of Experts (MoE):** AI model architecture divided into several subsets of neural networks called experts, specialised in a specific task, and a router that determines which expert should be used to answer a query.
- **Model:** an AI model is a mathematical construct that generates a deduction or prediction from input data. The model is estimated based on data during the training phase of the AI system (definition from CNIL glossary).
- **Model-as-a-Service (MaaS):** cloud platform or marketplace where developers can access multiple foundation models via a common programming interface.
- **Neural network:** in AI, a neural network is an organised set of interconnected artificial neurons used to solve complex problems such as computer vision or natural language processing (definition from CNIL glossary).
- **Natural language processing (NLP):** a multidisciplinary field involving linguistics, computer science and AI (definition from CNIL glossary).
- **Open source:** software in which the source code is available to the general public. The development of this "free software" involves a collaborative effort in which programmers improve the source code together and share changes within a community (definition from *Autorité* Opinion 14-A-18).
- **Open-weights:** foundation model whose weights are publicly accessible.
- **Platform-as-a-Service (PaaS):** cloud computing service consisting of deploying onto a cloud infrastructure consumer-created or acquired applications created using

programming languages, libraries, services and tools supported by the provider (definition from *Autorité* Opinion 23-A-08, page 183).

- **Reinforcement learning from human feedback (RLHF):** a reinforcement learning approach that uses feedback and evaluations from human users to guide the learning of an AI model. RLHF is used in text generators based on LLMs (definition from CNIL glossary).
- **Retrieval-augmented generation (RAG):** technique for improving the accuracy and reliability of generative AI models with facts fetched from external sources.
- **Software-as-a-Service (SaaS):** cloud computing service consisting of the provision to the consumer of the capability to use the provider's applications running on a cloud infrastructure (definition from *Autorité* Opinion 23-A-08, page 184).
- **Supercomputer:** a very large computer, combining several tens of thousands of processors and capable of performing a very large number of simultaneous computing or data processing operations (definition from the CEA).
- **Synthetic data:** artificial data generated from original data and a model trained to reproduce the characteristics and structure of the original data.
- **Tensor processing unit (TPU):** tensor processor.
- **Token:** sequence of a few letters that do not always form complete words.
- **Training (or learning):** machine learning process during which the AI system builds a model from data (definition from CNIL glossary).
- **Training data:** set of data (text, sounds, images, lists, etc.) used during the training/learning phase. The system trains itself on the data to perform the expected task (definition from CNIL glossary).
- **Weight/Parameter:** in a neural network, a weight is a power coefficient of the connection between two neurons, which is adjusted throughout the training phase (definition from safety recommendations for generative AI published by the French Cybersecurity Agency [*Agence nationale de la sécurité des systèmes d'information* – ANSSI], 29 April 2024).