COMPETITIVE BOTTLENECKS IN THE GENERATIVE AI VALUE CHAIN

A STARTUP AND INVESTOR PERSPECTIVE

March 11th, 2024

Table of contents

Executive summary

Over the past five years, digital markets have been in turmoil. New technologies and applications have opened tremendous opportunities for businesses and consumers but also led to the emergence of new barriers to fair competition. Faced with practices that challenged the effectiveness of traditional antitrust enforcement, the European Union (EU) adopted the Digital Markets Act (DMA), the world's first rulebook to prevent anticompetitive practices online. Now, **the mainstreaming of *generative* artificial intelligence (gen AI) raises the question as to the relevance of existing instruments in addressing competitive bottlenecks in emerging, high-tech markets**.

Gen AI ended in the hands of the average user almost overnight when OpenAI released its ChatGPT chatbot for free in late 2022. The massive adoption of the app immediately raised fundamental questions: How will society use this technology? What impact will it have on people and the environment? And how will businesses extract value from this innovation?

2023 saw an almost uninterrupted sequence of unprecedented business announcements: record company valuations, strategic partnerships, massive investments into new companies…Most of these announcements came from well-established, dominant US-based technology companies. In this context, **will incumbents capture all the value in the emerging generative AI market or is there room for new competitors to emerge?**

This will depend on several factors: the barriers to entry for new companies, the potential abuse of dominant position by the incumbents but also the degree of dependence of new entrants from incumbents for strategic inputs and infrastructure. Ultimately, **is a fully European generative AI supply chain possible?**

To answer these questions, we turned to our members, European startups, and venture capital funds, to get their feelings from the ground. We also decided to look beyond the most popular companies and services to have a more comprehensive **overview of what is going on along the whole gen AI supply chain.**

Here are our **main findings**:

- The gen AI supply chain is long and complex and features both hardware and software components. **Each layer — and sublayer— of the supply chain is a market on its own but is also interconnected to the next, making interdependencies among companies strong.**
- The four main components of the gen AI supply chain are **chips** (design software, raw materials, high-precision machinery, manufacturing), **infrastructure** (chips, data centers, networks, software, and services, including distribution services), **foundation models** (open source, open weights, closed source or proprietary) and **applications**.
- While there is a lot of hype around gen AI foundation models**, the majority of the economic value is currently concentrated in chips and infrastructure**. It is also expected that significant value will also come from **specialized applications (such as health, finance, etc.)** addressing concrete enterprise and consumer use cases.

- **Vertically integrated companies (such as Amazon, Google, Microsoft, and, to a certain extent, Nvidia) are the best placed to capture value** all along the value chain; all established players are venturing into new markets to secure their presence at each stage of the supply chain, either directly or indirectly via strategic partnerships and investments, while also securing their presence both in the consumer and enterprise market.

- **Partnerships are becoming the primary way of doing business in the gen AI market**, both between small and large companies but also between large companies, often leading to a dynamic of coopetition (cooperation between competitors).

- While **no overtly abusive market practice has been recorded so far**, the dominant and vertically integrated position of certain companies entails **several risks** (which are detailed in this report).

- Established companies are not only important infrastructure providers for startups but also valuable **entry points to new clients.**

- **European companies excel in highly specialized applications and chips and infrastructure niches** of the supply chain; they are also challenging US actors in the foundation models market.

Based on these, **France Digitale believes that**:

- **The devil is in the details**: anticompetitive behavior will likely result from the accumulation of market practices that, taken alone, could be seen as legitimate, but that become anticompetitive when systematically applied by companies abusing their dominant position and/or the economic dependence of players downstream.

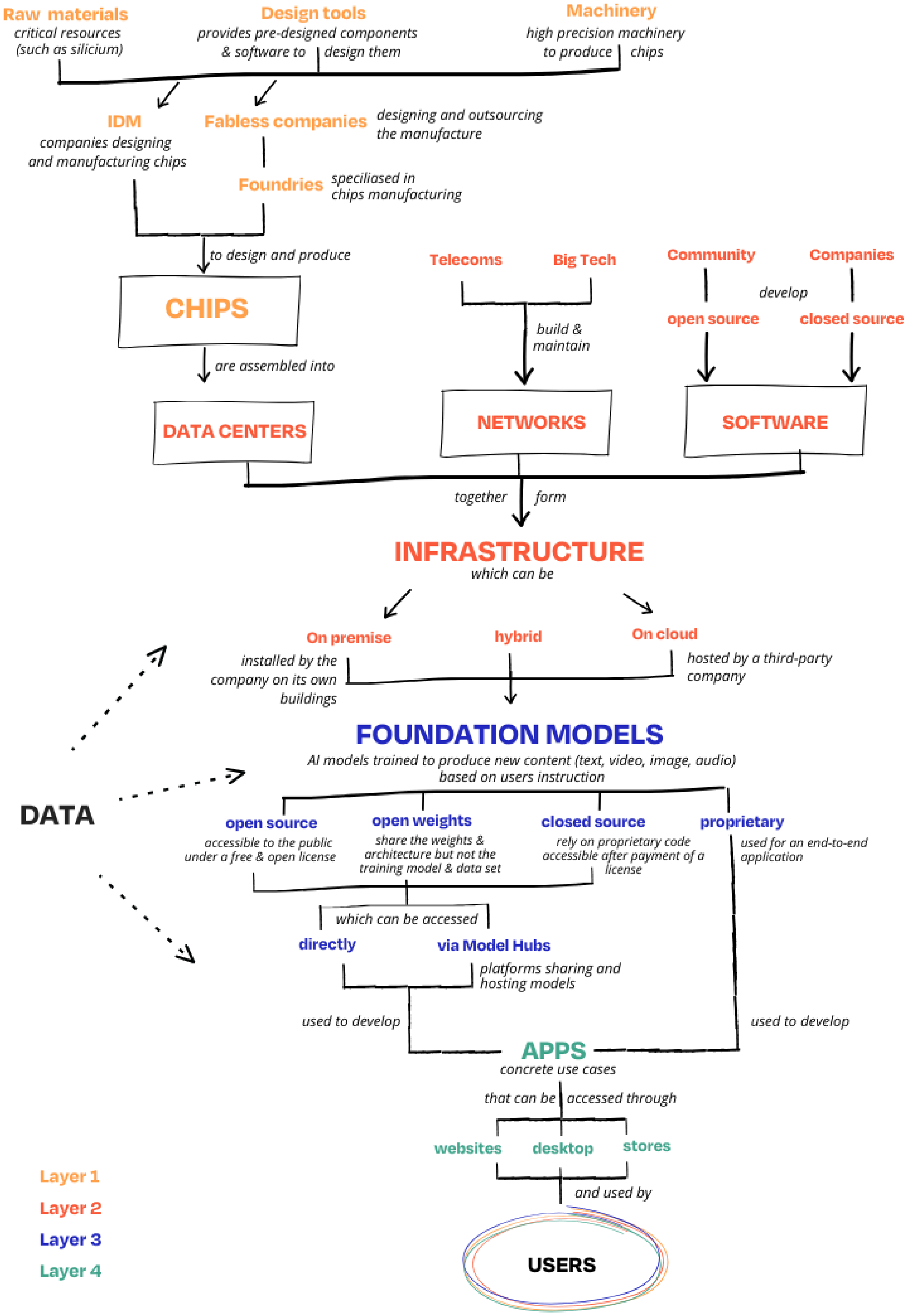
- As a result, **antitrust can't be the only answer to the competitive bottlenecks in the gen AI value chain**: preventive measures like those of the DMA should also be adopted.

- **Open weights and open source** models currently provide valuable alternatives to proprietary gen AI-related models, thus preventing technical lock-in and favoring the takeup of generative AI in existing companies. However, **open source models and software alone are insufficient to address competitive issues in the generative AI supply chain.** Open source is dependent on the technical contributions and the monetary donations of the developer community and of sponsors, which in turn raises questions as to their long-term technical and financial sustainability.

- Any action by authorities should be preceded by a **careful assessment of its side effects on European companies downstream** (startups) **and upstream**, (investors). This assessment should include, among others, considerations on interdependence, cost structures, and compatibility with quick innovation cycles.

- Authorities should not only focus on prohibiting certain practices but also proactively enact **policies, investment, and communication strategies that favor the emergence of strong European alternatives in all layers of the supply chain** to compete with dominant American and Asian companies.

Overview of the gen AI supply chain



Raw materials
critical resources (such as silicium)

Design tools
provides pre-designed components & software to design them

Machinery
high precision machinery to produce chips

IDM
companies designing and manufacturing chips

Fabless companies
designing and outsourcing the manufacture

Foundries
speciliased in chips manufacturing

to design and produce

CHIPS

are assembled into

DATA CENTERS

Telecoms

Big Tech

build & maintain

NETWORKS

Community

Companies

develop

open source

closed source

SOFTWARE

together form

INFRASTRUCTURE
which can be

On premise
installed by the company on its own buildings

hybrid

On cloud
hosted by a third-party company

DATA

FOUNDATION MODELS
AI models trained to produce new content (text, video, image, audio) based on users instruction

open source
accessible to the public under a free & open license

open weights
share the weights & architecture but not the training model & data set

closed source
rely on proprietary code accessible after payment of a license

proprietary
used for an end-to-end application

which can be accessed

directly

via Model Hubs
platforms sharing and hosting models

used to develop

used to develop

APPS
concrete use cases

that can be accessed through

websites

desktop

stores

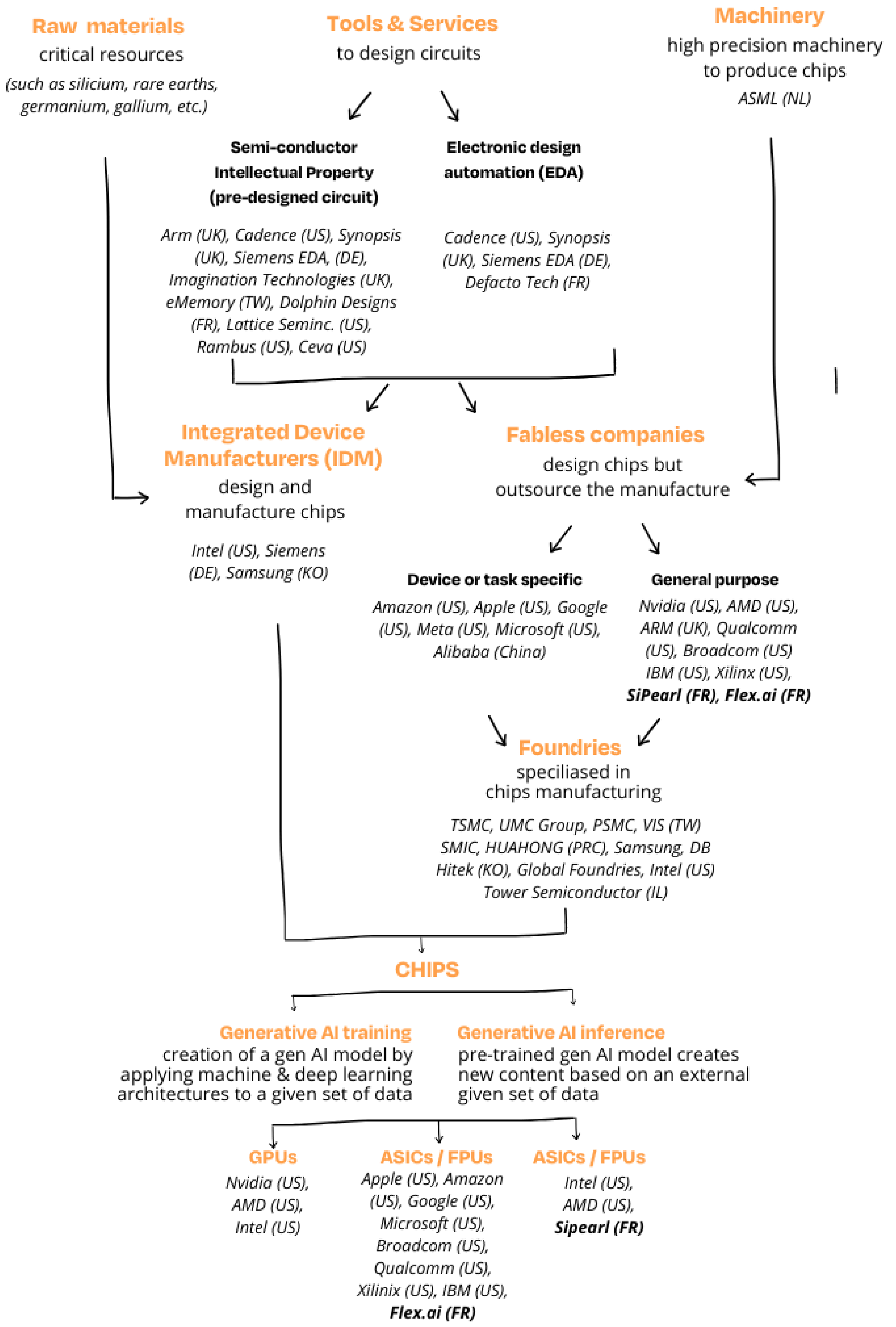
and used by

USERS

Layer 1
Layer 2
Layer 3
Layer 4

Layer 1 – Chips: AI's beating heart

Companies in the "chips" layer of the gen AI supply chain



Raw materials
critical resources
(such as silicium, rare earths, germanium, gallium, etc.)

Tools & Services
to design circuits

Semi-conductor Intellectual Property (pre-designed circuit)
Arm (UK), Cadence (US), Synopsis (UK), Siemens EDA, (DE), Imagination Technologies (UK), eMemory (TW), Dolphin Designs (FR), Lattice Seminc. (US), Rambus (US), Ceva (US)

Electronic design automation (EDA)
Cadence (US), Synopsis (UK), Siemens EDA (DE), Defacto Tech (FR)

Machinery
high precision machinery to produce chips
ASML (NL)

Integrated Device Manufacturers (IDM)
design and manufacture chips
Intel (US), Siemens (DE), Samsung (KO)

Fabless companies
design chips but outsource the manufacture

Device or task specific
Amazon (US), Apple (US), Google (US), Meta (US), Microsoft (US), Alibaba (China)

General purpose
Nvidia (US), AMD (US), ARM (UK), Qualcomm (US), Broadcom (US) IBM (US), Xilinx (US), ***SiPearl (FR), Flex.ai (FR)***

Foundries
speciliased in chips manufacturing
TSMC, UMC Group, PSMC, VIS (TW) SMIC, HUAHONG (PRC), Samsung, DB Hitek (KO), Global Foundries, Intel (US) Tower Semiconductor (IL)

CHIPS

Generative AI training
creation of a gen AI model by applying machine & deep learning architectures to a given set of data

Generative AI inference
pre-trained gen AI model creates new content based on an external given set of data

GPUs
Nvidia (US), AMD (US), Intel (US)

ASICs / FPUs
Apple (US), Amazon (US), Google (US), Microsoft (US), Broadcom (US), Qualcomm (US), Xilinx (US), IBM (US), ***Flex.ai (FR)***

ASICs / FPUs
Intel (US), AMD (US), ***Sipearl (FR)***

How does it work?

Integrated circuits (commonly known as “**chips**”) are the essential electronic components of all computing technology, including generative AI systems. Concretely, a chip is a set of electronic circuits engraved on a small and flat surface of silicon (“wafer”) programmed to execute certain tasks. Concretely, chips **provide the computing power and memory functions** to develop and deploy software systems, including generative AI.

💡 **Focus: What do generative AI systems need computing power for?**

Generative AI systems need computing power to perform two main functions:

- **Training**: it’s the process of creating a generative AI model by applying machine learning and deep learning architectures to a given set of data. *Example: the creation of Open AI’s GPT3 based on a particular deep learning architecture called “Transformer” and the information available on the internet*.

- **Inference**: the process by which a pre-trained generative AI model creates new content based on an external given set of data. *Example: You provide a chatbot a PDF of a contract and ask it to answer your questions about that contract.*

It is possible to **enhance** a model for a specific task or for more accurate responses for a particular use case through two main techniques. One is **fine tuning,** that is, to continue the training process using a particular dataset to improve the model’s performances for specific domains or tasks. The other is **Retrieval Augmented Generation (RAG)**, by which the AI retrieves facts from an external dataset by combining search functions (to find relevant documents) and a language model (to summarize or answer questions based on the given dataset). *Example of RAG: the integration of an LLM in your company information system will enable the chatbot to provide answers about your proprietary information.*

To produce chips, three inputs are needed: **raw materials** (notably silicon, germanium, gallium, phosphorus, indium phosphide, boron, some rare earths, and heavy amount of water), **electronic design tools** (that is, software and services to design chips) and **high-precision machinery** capable of producing highly precise engravings. These inputs are then put together by specialized factories (“**foundries**”), assembled and commercialized.

Not all chips are created equal, and some need more sophisticated production techniques than others. The smaller and more complex the circuit engraving, the more powerful the chip. As a result, **only the most advanced chips are capable of the complex computations required by generative AI systems**.

💡 **Focus: The right chip for the right use case**

While there exist dozens of chip architectures, **not all types of chips are adequate for the training of and inference by generative AI systems**. As of today, **three types of chips can perform these AI functions**:

- **Graphic Processing Units (GPUs)** can perform several computations at the same time ("parallel processing"). Originally used to treat images, especially in video games, GPUs have proven highly effective for the training of generative AI systems. They can also load and run heavy AI models for inference at a fast pace thanks to their heavy VRAM capacities and the use of thousands of cores. *Example: Nvidia's A100 80GB, which costs approximately €21K.*

- **Central Processing Units (CPUs)** provide higher computing power than GPUs but are not as performing in parallel computing due to their use of RAM and their limited number of cores. As a result, CPUs are inefficient for model training, but they can be used alone or with GPUs for specialized inference tasks. They are less expensive and energy consuming than GPUs, but their inference speed is some 3 times lower[1]. *Example: Intel Xeon Scalable processor.*

- **Application-Specific Integrated Circuits** (ASICs) and **Floating Point Units** (FPUs): fast and energy efficient, but only capable of performing specific tasks **(training or inference)**. *Example: Google's Tensor Processing Unit (TPU).*

What are the competitive bottlenecks?

Market concentration and competitive bottlenecks have emerged in most of the markets in this stage of the supply chain.

China's grip on raw materials: a risk for chip production?

Chips' primary material is silicon. While silicon is widely available, most of its extraction has been outsourced to China, which currently controls 70% of the market[2]. China also accounts for 80% of germanium and gallium production[3]. When it comes to rare earths, China benefits from some of the largest mineral deposits on the planet, enabling the country to account for 60% of the production market, compared to 15% by the US[4]. At the moment no rare earths are mined in Europe; while a significant deposit was discovered in Sweden last year, it would still take between 10 and 15 years to get operations starting[5]. When it comes to material processing, China is responsible for 85% of the global market: not only it has spent years developing its mining and processing capacities, but it has also been building remarkable strategic reserves of metals that are crucial to the manufacturing of digital devices[6].

[1] Deci (2024). *CPU vs GPU: How to Narrow the Deep Learning Performance Gap?* Deci

[2] *Mineral Commodity Summaries 2022 - Silicon.* U.S. Geological Survey

[3] Critical Raw Materials Resilience: Charting a Path towards greater Security and Sustainability. European Commission. 2020.

[4] *Mineral Commodity Summaries 2022 - Rare-Earths. U.S.* Geological Survey.

[5] Zimmermann, A. (2023). Mining firm: Europe's largest rare earths deposit found in Sweden. POLITICO

[6] Braw, E. (2023). *Beijing's Grip On Minerals Might Be Shaken By Norwegian Discoveries*. Foreign Policy.

This level of concentration grants China a position of market dominance that exposes companies downstream to several risks, in particular **export restrictions leading to artificial price increases**. Such practices could be imposed **by China individually or collectively through the BRICS** (Brazil, Russia, India, China, South Africa) alliance. The second option would be particularly worrying should BRICS accept all six candidates members, as the group would control 72% of the world's rare earths reserves[7]. As export restrictions are a form of **abuse of the economic dependence of commercial partners** and can lead to major supply chain disruptions, authorities should remain particularly vigilant to commercial and geopolitical developments in this field.

💡 **Focus: China's history with raw materials export restrictions**

In 2023, China restricted its exports of germanium and gallium in retaliation for US protectionist measures on chips[8]. In 2010, China also reduced by 40% the export of some rare earths and imposed an outright export ban to Japan in response to political tensions[9]. The consequent increase in the price of rare earths was condemned by the World Trade Organization (WTO) as an unjustified commercial restriction[10].

A European champion in high-precision machinery

When it comes to this layer, as of now, only one company, the Dutch Advanced Semiconductor Material Lithography (ASML), is capable of providing equipment for the engraving of the most advanced chips (with circuit nodes smaller than 14 nm[11]). While Canon has announced its willingness to enter this market, it is not operational yet[12].

Chip design: an essential but often overlooked market

The market for chip designing tools is divided into two submarkets: That of providers of Intellectual Property (IP), that is, pre-designed circuits that customers can adapt to their needs, and that of EDA systems, that is, software tools to design circuits. In the first submarket (IP) the British company ARM accounts alone for 41% of the market share[13], but plenty of other providers are also active across Europe, the US, and Asia. The EDA submarket is dominated by three companies (Cadence, Siemens EDA, and Synopsis) capturing 75%[14] of the market share. The market is also geographically concentrated, with two US-based suppliers (Cadence and Synopsys) taking 62% of the market share[15]. It should be noted that all three are also **vertically integrated**, as they are also present in the IP submarket (especially Cadence and Synopsys), thus creating a **risk of lock-in** for users.

[7] (2021). Where are the world's rare earth reserves ? Visual Capitalist

[8] *Le Parisien. Ce sont des métaux stratégiques : pourquoi la Chine limite la vente de gallium et de germanium*

[9] Madelin, T. (2010). *La Chine Réduit Ses Exportations de Terres Rares.* Les Echos.

[10] Boisseau, L. (2011). *La Flambée Continue des Prix de Ces Métaux Pourrait Se Détendre À Partir de 2013.* Les Echos.

[11] Terrasson, B. (2024). *ASML, Seul Fournisseur de Systèmes Avancés de Fabrication de Puces, Signe Sa Meilleure Année.* Siècle Digital.

[12] Tazrout, Z. (2023). *Canon Cherche À Concurrencer ASML Sur le Marché des Machines de Fabrication de Semi-conducteurs Avancés.* Siècle Digital.

[13] (2023). *Semiconductor Design IP Revenue by Company.* IPnest.

[14] (2023). *Global EDA Software Market Share.* TrendForce

[15] Ibidem

Cloud giants vs. chip titans: coopetition in the chip market

Regarding the **market for the design of chips**, two main types of actors are present: Integrated Device Manufacturers (IDMs), which are responsible both for the design and the manufacturing of the chips, and so-called *fabless*, that is companies that design chips but outsource their manufacturing to specialized factories, the foundries. Among fabless, two submarkets exist: that of device or task-specific chips and that of general purpose chips.

In the first submarket (device and task-specific chips) we find device producers but also **vertically integrated companies offering cloud and software services**. The main actors are US-based companies, notably, Amazon[16], Apple[17], Google[18], Meta[19], and Microsoft[20], but also Chinese companies like Alibaba[21]. They have entered this market to ensure the autonomy of their chip supply by developing chips that are specific to their devices and/or services (e.g. Google's TPU for cloud computing, developed since 2018).

The second submarket (fabless) is also dominated by US-based companies. It should be noted that not all companies design all types of generative AI compatible chips (GPUs, CPUs, ASICs). The design of CPUs is dominated by Intel, which accounts for 71% of the market, but rising competitors like AMD, AWS, and Ampere are challenging its position[22]. **The GPU market**, instead, **is much more concentrated, with Nvidia controlling 84%[23] of the total market and 92%[24] of the market for GPUs for cloud computing**. Its two main competitors (AMD and Intel) are still struggling to get traction, with AMD being in the best position according to both market observers[25] and startups interviewed as part of our research. Even if AMD emerges as a strong competitor in the cloud GPU market, this will take years. For some time, Nvidia is largely expected to maintain its market dominance. This exposes companies downstream to several risks.

There is a potential for monopolistic practices such as **price fixing**, **output restriction**, imposition of **unfair contractual terms** and **discriminatory behavior against competitors**. None of this, however, has materialized so far to the best of our knowledge. All the startups interviewed that have a direct relationship with Nvidia (both cloud providers and startups) have indicated to have a mutually beneficial - **albeit fragile** - relationship with Nvidia. They all have a "privileged" customer-provider relationship, often complemented by a partnership (for example, a reseller agreement). They are however aware of the imbalance of bargaining power with Nvidia, which could lead to the unilateral interruption of the partnership or the end of the preferred treatment. For example, in case of supply shortages, bigger clients are likely to be privileged.

[16] Novet, J. (2023). *Amazon Announces New AI Chip As It Deepens Nvidia Relationship*. CNBC.

[17] Bajarin, T. (2023). *Apple's AI Prowess Lies In Its Neural Engine*. Forbes.

[18] Gavois, S. (2023). *Tensor Processing Unit (TPU) V5e de Google Cloud : Plus Performant Que les V4 ? Oui… et Non*. Next.

[19] Peters, J. (2023). *Meta Is Working On A New Chip For AI*. The Verge

[20] Warren, T. (2023). *Microsoft Is Finally Making Custom Chips — And They're All About AI*. The Verge.

[21] Pan, C. (2023). *Alibaba's Chip Design Subsidiary Launches RISC-V Chip To Boost Performance Of Its Cloud Data Centres*. South China Morning Post.

[22] Counterpoint (2023). AMD Market Share Surpasses Intel In Share Growth. Counterpoint

[23] Dow, R. (2023). The Graphics Add-in Board Market Continued Its Correction In Q1 2023. Jon Peddie

[24] Wegner, P. (2023). *The Leading Generative AI Companies* . IoT Analytics.

[25] Prickett Morgan, T. (2024). *The tough road still ahead for Intel in the data center*. Next Platform

Among Nvidia's biggest clients are the vertically integrated cloud companies mentioned above -AWS, Alibaba, Google, and Microsoft- which are at the same time partners and competitors to Nvidia. On the one hand, these companies started the production of their own device or task-specific chips to limit their dependence on Nvidia; on the other, they announced grand partnerships[26] and massive purchasing plans[27] of Nvidia's GPUs. This dual relationship may be described as a form of ***coopetition*** (cooperation between competitors). Moreover, as none of these partnerships is exclusive, their effect is maximizing, rather than limiting, the availability of GPUs to users, which could be seen as a strategy to avoid any allegation of monopolistic or oligopolistic behavior.

💡 **Focus: what is coopetition?**

Coopetition is a type of strategic alliance between rival organizations, often in the software and hardware sectors, in the hope of mutual benefits, sometimes on a specific project.[28]

Chip manufacturing: an evolving market exposed to geopolitical risks

The advanced chip manufacturing (3-7 nm) market is also highly concentrated. Excluding IDMs, **TSMC (Taiwan) holds a remarkable 90% market share**[29], followed by Samsung Foundry (South Korea). Other advanced chip production (7-28nm) companies include China's SMIC and the USA's Global Foundries. While there seems to be no exclusivity agreement between fabless and chip manufacturers, the risks related to concentration in this market are multiple. Apart from the abuse of its dominant position, the geographic concentration of chip manufacturing in East Asia raises geopolitical risks[30]. Tensions or conflicts between these countries (China and Taiwan, China and USA, North Korea and South Korea) could disrupt the supply of critical technologies.

To address this issue, **the private sector is taking action, with existing companies expanding into the foundries market**. The American company Intel is entering the foundry industry with the aim of becoming a direct competitor to TSMC by 2030. For decades, Intel has only manufactured chips for its internal consumption. Now, the company intends to separate its activities by creating a division called Intel Foundry Services, dedicated to producing advanced integrated circuits for other fabless providers. This move signifies Intel's openness to suppliers it traditionally viewed as competitors, including Microsoft, Nvidia, Qualcomm, Google, and even AMD. With this new strategy, Intel is adopting a model similar to Samsung, which manufactures chips for its own use as well as for fabless suppliers[31].

[26] Leswing, K. (2023). *Nvidia's stock closes at record after Google AI partnership.* CNBC
[27] Vanian, J (2024). *Mark Zuckerberg indicates Meta is spending billions of dollars on Nvidia AI chips.* CNBC
[28] Oxford English Dictionary (2019)
[29] Macquarie Research, Company Reports, VanEck. Data as of August 31, 2021.
[30] Julied, B. (2023). *Semiconductors Sector : Geopolitical And Climate Change-related Risks At The Heart Of The Semiconductors Sector Outlook.* Credendo.
[31] Knight, W. (2024). *Intel's AI Reboot Is The Future Of US Chipmaking.* WIRED.

Governments are also investing in the development of advanced foundries. In 2022, the USA launched its CHIPS for America Fund, a 53 billion USD grant pool for the construction of foundries on US soil. TSMC, Intel, and others are expected to benefit from the grant[32]. The announcement of the CHIPS for America fund has further led to 166 billion USD in private investment[33].

In Europe, the Chips Act, which was adopted in July 2023, will mobilize more than €43 billion of public and private investments together with Member States and international partners[34]. Following this regulation, TSMC[35], Intel[36], and Global Foundries[37], in partnership with the Franco-Italian company STMicroelectronics, have announced the construction of "megafabs" in France, Germany, and Italy.

[32] Owen, M. (2024). *Massive $ 53B US Chip Fund Grant Announcements Expected Within Weeks*. AppleInsider

[33] Robuck, M. (2023). *US Chip Funding Scheme Attracts Investments Of $ 166B*. Mobile World Live

[34] (2023).European Chips Act . European Commission.

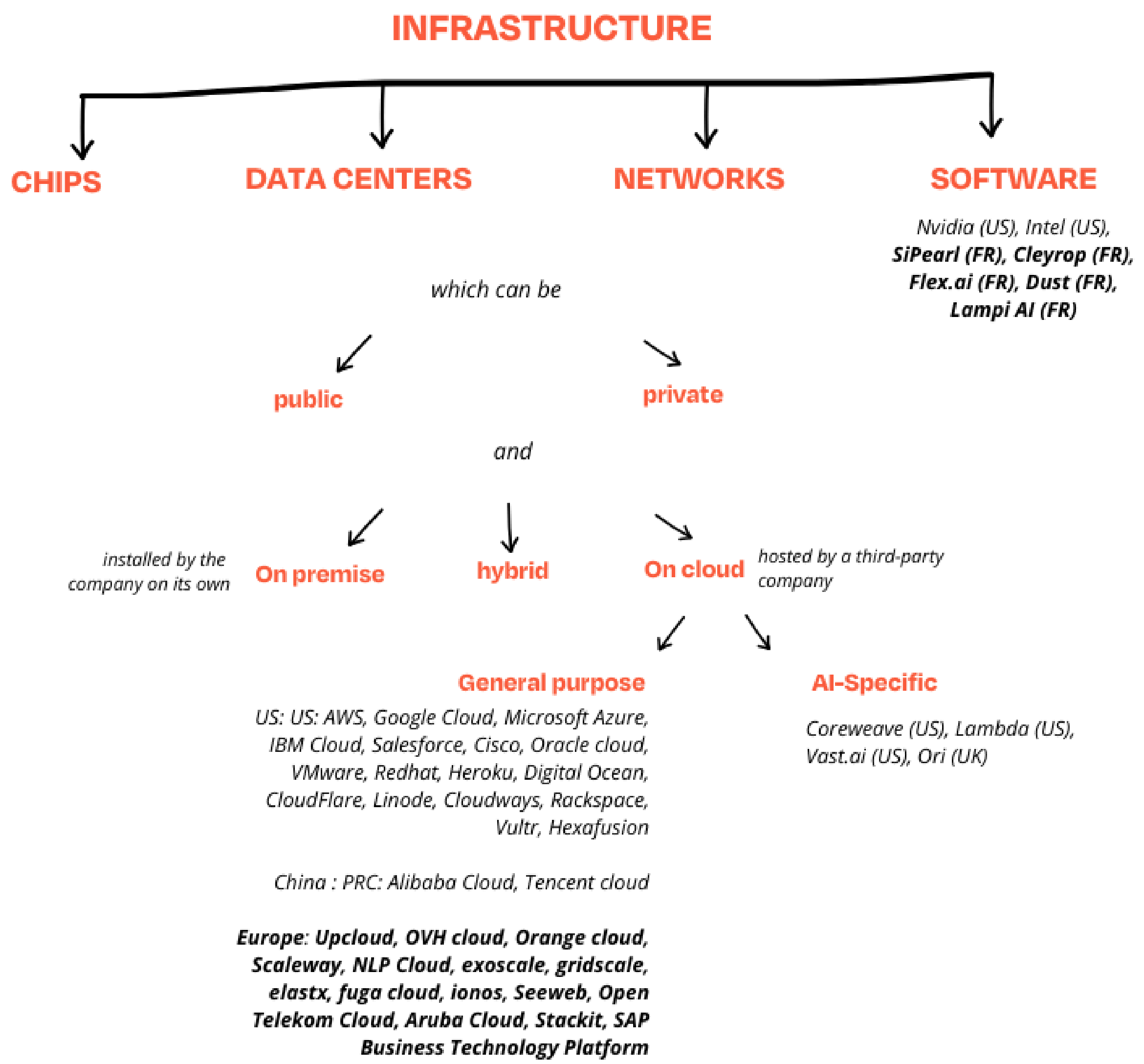
[35] Loukil, R. (2023). *Le Projet de Mégafab de TSMC En Europe Dans les Starting-blocks*. Usine Nouvelle.

[36] Steiwer, N. (2023). *Intel conclut un accord avec L'Allemagne pour une méga-fab à 30 mld d'euros*. Les Echos.

[37] Pollet,M. (2022). Semi-conducteurs : La France Va Accueillir une Nouvelle « Mega-fab ». Euractiv.

Layer 2 - Infrastructure: the omnipresence of hyperscalers

Companies in the "infrastructure" layer of the gen AI supply chain



US: US: AWS, Google Cloud, Microsoft Azure, IBM Cloud, Salesforce, Cisco, Oracle cloud, VMware, Redhat, Heroku, Digital Ocean, CloudFlare, Linode, Cloudways, Rackspace, Vultr, Hexafusion

China : PRC: Alibaba Cloud, Tencent cloud

*Europe: **Upcloud, OVH cloud, Orange cloud, Scaleway, NLP Cloud, exoscale, gridscale, elastx, fuga cloud, ionos, Seeweb, Open Telekom Cloud, Aruba Cloud, Stackit, SAP Business Technology Platform***

Coreweave (US), Lambda (US), Vast.ai (US), Ori (UK)

How does it work?

AI infrastructure refers to the hardware and software used to create, train, and run AI models. It includes:

- **Hardware:** all physical and electronic components (chips, servers, data centers, networks, etc.) needed to store data, train algorithms, and deploy AI systems.
- **Software:** intangible computer instructions used to develop AI systems, from chip programming models to machine learning frameworks, development libraries, or data management and analysis tools, and many more.

Infrastructure is used to **collect, store, and manage data**, a key input for the development of gen AI models (more on this in the following chapter).

The choice of infrastructure depends on the specific needs of each AI project. We distinguish two main approaches:

- **On-premise:** the infrastructure is installed and managed by the company on its site/buildings. This option offers total control over the physical and digital assets, including the data, but requires a significant investment in hardware, software, and personnel to ensure the management and maintenance of data centers.
- **Cloud-based:** the infrastructure is hosted by a third party, the so-called "cloud provider". This option offers scalability and flexibility without the burden of managing a data center.

Not all cloud providers operate at the same scale. They can offer a combination of IaaS, PaaS, and SaaS, and their scale can vary considerably.

- **Hyperscalers** are global players such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). They offer a wide range of services (PaaS and SaaS) and a massive infrastructure capable of training large foundation models.
- **Mid-tier or specialized cloud providers** are players that offer a smaller infrastructure or that can focus on specific services (such as private cloud, sovereign cloud, cloud for a vertical industry, etc.).

💡 **Different clouds for different needs:**

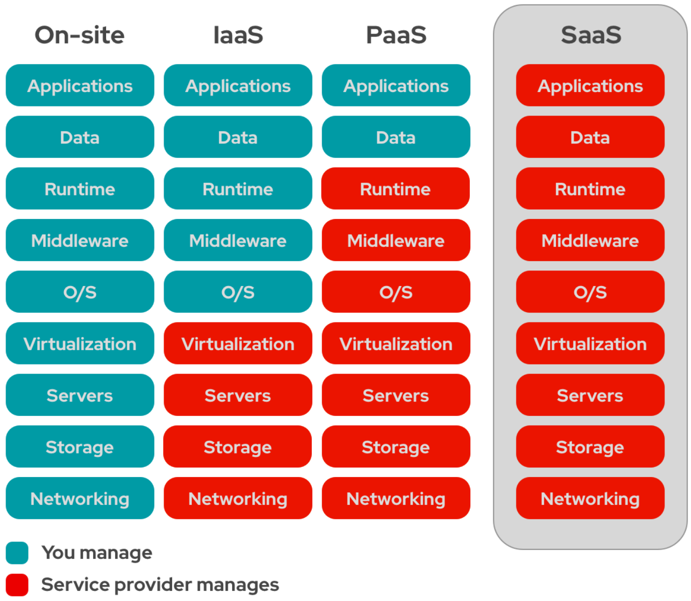
Today, cloud providers offer companies three tiers of services, from infrastructure all the way to app development and deployment:

Infrastructure as a Service (IaaS), providing some or all of the basic infrastructure in a virtualized form. Users are responsible for managing their own operating system, applications, data, and configurations.

Platform as a Service (PaaS), provides the user with a development and execution platform to create and deploy their own applications.

Software as a Service (SaaS), provides software applications hosted on the cloud and accessible to users via the Internet, on a subscription basis.

IaaS is the level that allows the user to maintain the most control over data and infrastructure, while SaaS is the one where control is most delegated to the third party.



On-site	IaaS	PaaS	SaaS
🟦 Applications	🟦 Applications	🟦 Applications	🟥 Applications
🟦 Data	🟦 Data	🟦 Data	🟥 Data
🟦 Runtime	🟦 Runtime	🟥 Runtime	🟥 Runtime
🟦 Middleware	🟦 Middleware	🟥 Middleware	🟥 Middleware
🟦 O/S	🟦 O/S	🟥 O/S	🟥 O/S
🟦 Virtualization	🟥 Virtualization	🟥 Virtualization	🟥 Virtualization
🟦 Servers	🟥 Servers	🟥 Servers	🟥 Servers
🟦 Storage	🟥 Storage	🟥 Storage	🟥 Storage
🟦 Networking	🟥 Networking	🟥 Networking	🟥 Networking

🟦 You manage

🟥 Service provider manages

Infrastructure control by the user across IaaS, Paas and SaaS service models

Source: RedHat

It should be noted that some hyperscalers, like Microsoft's Azure, and some mid-tier providers, like OVHcloud and Scaleway, have a IaaS and PaaS offering as **private cloud**, granting clients exclusive use of a certain part of the provider's infrastructure.

The majority of startups and scale-ups interviewed in our research rely on public cloud services. Several among them apply a so-called "**multi-cloud strategy**", meaning that they rely simultaneously on different providers. **Some of them do so opportunistically, based on the availability of cloud credits** (that is, free-tier offerings), while **others do so consciously out of performance and confidentiality considerations**. It is generally acknowledged by interviewed companies that US-based hyperscalers currently are the best-performing cloud providers on the market. However, their obligation to abide by US extraterritorial legislation on data access (CLOUD Act, FISA's Section 702) raises concerns among startups' EU-based corporate and public sector customers dealing with sensitive data, like financial and health data. As a result, startups have adopted several strategies to limit this risk.

A significant number of startups adopt a multi-cloud strategy based on a **mix of American and European cloud providers,** using one or the other depending on their clients' confidentiality requirements. Other startups offer the possibility to **deploy their software on the client's infrastructure** (on-premise or private cloud) and a minority rely on a **private cloud** or have their own **on-premise infrastructure** (often complemented with a public cloud offering for testing, research or non-strategic activities).

What are the competitive bottlenecks?

The dominance of popular chip programming models

Chip programming models, the “operating system” of chips, is the first infrastructure layer where competitive risks emerge. As was the case with Microsoft Windows on personal computers in the 1990s and early 2000s, there is a risk for certain chip programming models to dominate the market and lead to user lock-in. A case in point is Nvidia's chip programming model CUDA, which since its launch in 2007 has become the *de facto* standard for GPU acceleration in AI. Nvidia's success with CUDA is due to its ease of use, early investment in complementary tools, libraries, and AI frameworks, and collaborations with universities and tech companies. Due to its proprietary nature, however, code bases developed on CUDA remain difficult to port to other programming models, leading to lock-in.

To mitigate dependence on CUDA, competitors like Intel are investing in alternative programming models, open source frameworks are increasingly available and interoperable across chip programming models, GPU-neutral languages are emerging, and AMD GPUs, which are not dependent on CUDA, are gaining traction[38]. However, another risk emerges here: **if Nvidia doesn’t succeed in adapting to this new, more competitive reality, it may revert to anti-competitive practices like Microsoft did in the early 1990s** to secure its monopoly over the operating system and browser markets.[39] **Competition authorities should therefore remain vigilant as to the openness of this market:** Nvidia’s GPUs should remain open to chip programming models other than CUDA, portability and interoperability across models should be strengthened and alternatives to CUDA should be given a fair chance to emerge.

Chip providers inroads in the cloud market

Just as hyperscalers are entering the chips market to limit their dependence on Nvidia, so is Nvidia entering the cloud market to diversify its client base. Nvidia has recently started investing in AI-specialized cloud providers like Coreweave through its corporate investment arm[40]. Such 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[41]. This could lead to unfair pricing competition with general purpose cloud providers, especially mid-tier actors that lack the financial firepower of hyperscalers.

[38] 1Kg (2023). *Nvidia’s CUDA Monopoly*. Medium

[39] *Complaint : U.S. V. Microsoft Corp.* US Department of Justice, Antitrust Division

[40] Archibald L. (2023). *How NVIDIA Fuels the AI Revolution With Investments in Game Changers and Market Makers*. Nvidia Blog.

[41] Venturo B. (2023). *CoreWeave + NVIDIA*. Coreweave

Lock-in practices in the cloud market

The public cloud market is concentrated in the hands of US-based hyperscalers Amazon Web Services (AWS), Microsoft's Azure, and Google Cloud Platform (GCP) which account for 65% of the global market[42]. The level of concentration is even higher in France, where these three companies account for 71% of the market[43]. Such level of market concentration has[44] and continues[45] to raise concerns about fair competition on several fronts. Aside from the privileged partnerships to access computing power (notably GPUs) discussed in Chapter 1, hyperscalers have also put in place commercial and technical practices to artificially cement their dominance in the market.

Among the commercial practices put in place by hyperscalers to lock-in their business users, cloud credits and egress fees stand out.
Cloud credits are allocations of cloud services accessible for free for a certain period. While they are undeniably of great help to startups, especially in the early days of their development, in time they may lead to dependence on a certain cloud provider. It should also be noted that **hyperscalers are best positioned to offer cloud credits for prolonged periods**, given their superior financial resources compared to mid-tier cloud providers, thus distort competition with smaller players. Interestingly, interviewed startups reported that **cloud credit offerings have accelerated substantially over the past few months** in line with the excitement around AI and that AWS, Microsoft's Azure, and GCP are clearly competing to attract as many startups as possible.
Egress fees, instead, are a type of switching charge that users have to pay to migrate their data from one cloud provider to another but that are not related to costs borne by the provider and are thus abusive. While not all cloud providers charge them, egress fees appear in the leading hyperscalers' contractual clauses.

Among the technical practices put in place by hyperscalers to lock-in their business users, lack of interoperability and access restriction to certain software stand out. Hyperscalers may refuse to enable interoperability with other cloud providers and deny access to certain Application Programming Interfaces (APIs), making it challenging for application developers to create software that works across multiple platforms. They can also restrict access to certain software in case of contract termination[46].

While the good news is that these practices are, at least partly, prohibited by the EU Data Act, there is a **risk of lock-in in the period between now and the entry into force of the Data Act** (September 2025 for some provisions and January 2027 for others). Hyperscalers have no incentive to deprive themselves of this revenue source ahead of time, so cloud credits and egress fees - with their anticompetitive effects - are likely to remain in place for the time being. There already is evidence for this: Google Cloud Platform, for example, announced put an end to its cloud switching charges only to clarify shortly after that the change only applied to a selection of customers[47]. AWS also removed egress fees for some

[42] (2023). *Cloud Infrastructure Services Market*. Synergy Research Group
[43] Markess by Exaegis, *Part de marché 2021 des acteurs du IaaS / PaaS en France*, 2022
[44] Bass, D., Berthelot, B. and Deutsch, J. (2023). Microsoft, OVH Prepare to Settle Cloud Complaint to EU. Bloomberg
[45] Elias, J. & Goswami, R (2023). Google accuses Microsoft of unfair practices in Azure cloud unit. CNBC
[46] Autorité de la Concurrence (2023). L'Autorité de la Concurrence Rend Son Avis Sur le Fonctionnement Concurrentiel du Secteur du Cloud. Autorité de la Concurrence.
[47] The Stack (2024). *Google Cloud is NOT magicking away data egress fees*. The Stack

customers[48]. Moreover, **important issues, like the maximum legitimate duration of cloud credits or abusive restrictions to access software, have not been regulated at EU level altogether**, opening the door for national regulation and thus for distortions within the Single Market and for the continuation of abusive practices by hyperscalers.

💡**What the Data Act says**

Recital 38. Cloud credits shall not lead to user lock-in. *[...] Customers benefiting from free-tier offerings should also benefit from the provisions for switching that are laid down in this Regulation, so that those offerings do not result in a lock-in situation for customers.*

Article 23. Removal of obstacles to cloud switching. *Providers of data processing services shall take the measures provided for in Articles 25, 26, 27, 29, and 30 to enable customers to switch to a data processing service, covering the same service type, which is provided by a different provider of data processing services, or to on-premises ICT infrastructure, or, where relevant, to use several providers of data processing services at the same time. In particular, providers of data processing services shall not impose and shall remove pre-commercial, commercial, technical, contractual and organisational obstacles, which inhibit customers from:*

a) terminating, after the maximum notice period and the successful completion of the switching process, in accordance with Article 25, the contract of the data processing service
b) concluding new contracts with a different provider of data processing services covering the same service type
c) porting the customer's exportable data and digital assets, to a different provider of data processing services or to an on-premises ICT infrastructure, including after having benefited from a free-tier offering;
d) in accordance with Article 24, achieving functional equivalence in the use of the new data processing service in the ICT environment of a different provider of data processing services covering the same service type;
e) unbundling, where technically feasible, data processing services referred to in Article 30(1) from other data processing services provided by the provider of data processing services

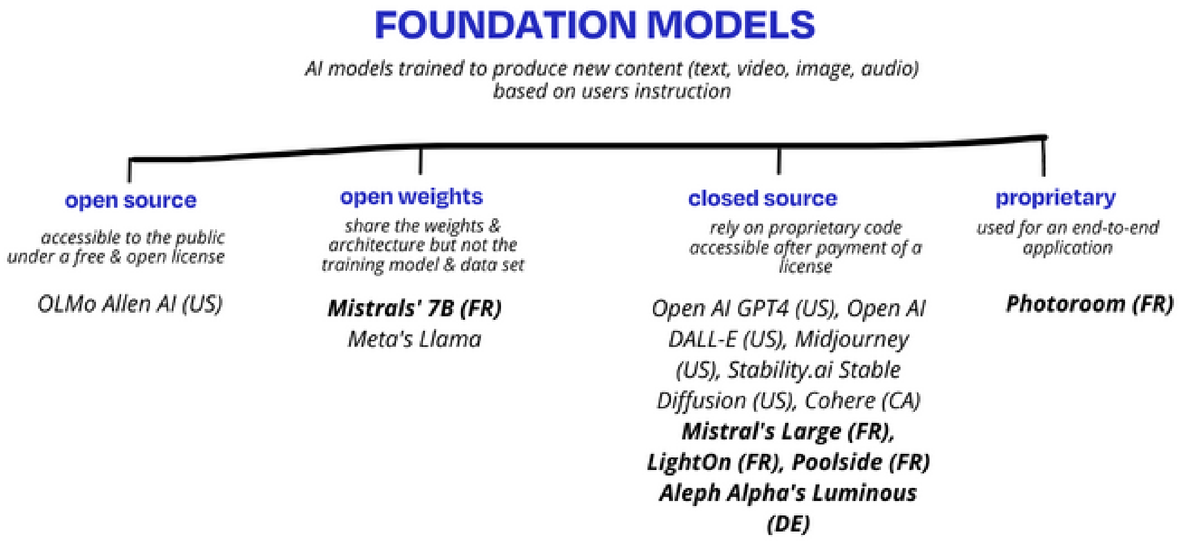
Art 29. *Gradual withdrawal of switching charges.*

1. *From 12 January 2027, providers of data processing services shall not impose any switching charges on the customer for the switching process.*
2. *From 11 January 2024 to 12 January 2027, providers of data processing services may impose reduced switching charges on the customer for the switching process.*
3. *The reduced switching charges referred to in paragraph 2 shall not exceed the costs incurred by the provider of data processing services that are directly linked to the switching process concerned.*

[48] Stormacq, S. (2024).*Free data transfer out to internet when moving out of AWS*. AWS News Blog

Layer 3 - Foundation models: a market with high barriers to entry

Companies in the "foundation model" layer of the gen AI supply chain



FOUNDATION MODELS

AI models trained to produce new content (text, video, image, audio) based on users instruction

open source

accessible to the public under a free & open license

OLMo Allen AI (US)

open weights

share the weights & architecture but not the training model & data set

Mistrals' 7B (FR)
Meta's Llama

closed source

rely on proprietary code accessible after payment of a license

Open AI GPT4 (US), Open AI DALL-E (US), Midjourney (US), Stability.ai Stable Diffusion (US), Cohere (CA) **Mistral's Large (FR), LightOn (FR), Poolside (FR) Aleph Alpha's Luminous (DE)**

proprietary

used for an end-to-end application

Photoroom (FR)

How does it work?

Generative AI foundational models are **systems that can produce new content or data,** such as text, images, and audio, based on a given set of data and instructions by the user ("prompt"). The ability to create new content comes from the application of machine learning and deep learning algorithms to a certain dataset (training). This "teaches" the AI model to create content that is similar to the original. Once the model has learnt this, it can be applied to any dataset (inference) to produce content relevant to the dataset.

Foundation models had been research projects for years before OpenAI became a commercial entity in 2019 and made a foundation-model based application - ChatGPT - available for the general public in late 2022.

Foundation models can vary in size and scale. **General-purpose models, like Large Language Models,** are trained on massive amounts of data and apply up to billions of parameters. They are generalist, so they can produce a wide range of content with a varying degree of precision on specific topics. **Specialized models** are trained on smaller but very specific datasets and are optimized for some specific tasks.

They can be open source, open weights, or closed source:

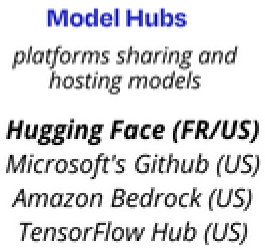
- **Open source models** are made accessible to the public under a free and open license that allows for the access, usage, modification, and distribution of the model. Open source models make the model architecture, training code, parameters ("weights"), the training dataset, and information on the model usage publicly available, enabling the maximum level of customization by users.

- **Open weights models** are made accessible to the public under a license that allows some re-use and are thus less open than fully fledged open source models. Open weight models make the parameters ("weights"), architecture, and information on model usage publicly available, but they do not share the training dataset nor the training code.

- **Closed source models** rely on proprietary source code that is only accessible after the payment of a license, subscription or other form of financial compensation.

Developers can either **build their own generative AI models** or **use existing models.** Developing a generative AI model requires three main inputs: computing power, data, and talents. Existing models, instead, can be accessed directly through an API or indirectly through a model hub. **Model hubs are specialized platforms that share and host foundational models.**

Companies at the "model hub" sub-layer in the gen AI supply chain



Model Hubs

platforms sharing and hosting models

Hugging Face (FR/US)
Microsoft's Github (US)
Amazon Bedrock (US)
TensorFlow Hub (US)

High barriers to entry

28 Out of the 30 startups interviewed as part of this research shared the view that developing a generative foundation model was not in their strategy as the process is too expensive (from a financial, technical and human resource standpoint). The Return on Investment (ROI) is unclear for them given the availability of off-the-shelf proprietary solutions and customizable open source and open weight options on the market. The 4 interviewed investors also indicated that they were not planning to invest in general-purpose foundation models, due to the massive upfront capital investment required, the difficulty in making profit margins, and therefore obtaining a meaningful ROI. Both groups expressed a preference for smaller (in terms of parameters and data) and specialized models. As a result, a minority of interviewed companies are developing specialized and proprietary generative AI models (some of which have been on the market for more than 5 years now), and the rest are enhancing existing models (with techniques like fine-tuning or RAG).

What makes it so difficult to develop foundation models?

- **Computing power:** There's a direct **correlation between the size of a language model and the amount of computing power it needs**. To illustrate this, researchers estimate that training a large language model (LLM) similar to GPT-3 required approximately **1,024 NVIDIA A100 GPUs over 34 days**[49]. The cost of

[49] Narayanan, Deepak et al. (2021). *Efficient Large-Scale Language Model Training On GPU Clusters Using Megatron-LM*. Cornell University

training more advanced models is higher, as OpenAI's GPT-4 is significantly larger than its predecessor. While there's no official data on the total computing power utilized to train GPT-4, **OpenAI used Microsoft Azure's 10,000 GPUs supercomputer**[50]. Estimates from Intel suggest the training cost for Open AI could be in the billions of dollars[51]. Such massive use of GPUs, in turn, entails major electricity costs. For instance, training ChatGPT-3 required approximately 1,283 MWh[52], **equivalent to the average energy consumption of about 274 French households for an entire year.** Inference also requires energy intensiveness. The energy cost of a single query on ChatGPT is estimated at 2.9 Wh. When multiplied by the volume of queries per day recorded at the beginning of 2023, it amounts to 564 MWh per day and 206 GWh per year, **the annual electricity consumption of a country like the Central African Republic**[53]. These substantial resource requirements for larger models limit the accessibility of training general-purpose models to a few companies. However, **specialized models or fine-tuning large models require significantly fewer resources for training, usage, and adaptation, making them more attractive to the generative AI community.**

- **Data access:** training a generative AI model requires massive amounts of data (for Large Language Models (LLMs) we speak of trillions of tokens of data[54]). There are four main categories of training data:
  - **Commercial data**: this includes high-quality content like books, music, newspapers, scientific literature, etc. Much of the access to this data is concentrated through copyright, technical, or contractual means. For example, Google retains exclusive access to millions of digitized books until the end of the digitization contracts with the original rights-holders.
  - **User data**: this includes information on the use of a certain software or platform, like interaction and preference data. Meta, for example, retains information on Facebook user's ad preferences.
  - **Open data**: any type of data that can be freely accessed, processed, and reused. **Interviewed companies were unanimous in pointing out that currently available open data is insufficient to train AI models.**
  - **Synthetic data**: any type of artificially generated data. While perfect to protect privacy, **interviewed companies were unanimous in pointing out that synthetic data is inadequate to train AI models.**

[50] Langston, J. (2020).Microsoft announces new supercomputer, lays out vision for future AI work. Microsoft

[51] (2024). Intel's CEO Says AI Training Now Costs Billions. Wired

[52] Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, et Jeff Dean. (2021). Carbon Emissions And Large Neural Network Training.

[53] Randrianarisoa, M. (2023). *Voici la consommation d'électricité phénoménale de l'intelligence artificielle*. Révolution Énergétique.

[54] Premosa, G. (2023). *AI Tokens: The Building Blocks of Language Models*. Povio

Accessing the necessary amount of adequate and high quality data to train AI models is currently difficult for three main reasons: (1) a complex, fragmented and ineffective **legal framework** both in the EU and beyond, (2) a significant **concentration of the market for user data** and (3) the **entrenched position of holders of commercial data**. These will be discussed in the following section.

- **Data cleaning and labeling:** data can't be used directly to train or fine-tune an AI system: it must first be cleaned and labeled to make it readable and prevent, to the best extent possible, bias. According to interviewed startups, labeling data from scratch to fine-tune a specialized AI model may take up to 6 months. Additional iterations to improve model output, for example through Reinforcement Learning techniques, further lengthen the procedure.

- **Talents:** AI engineers capable of developing foundation models are scarce, so they are not only expensive but also hard to attract. Big Tech have an advantage as they can not only offer the highest salaries and most comprehensive employee benefits, but also because they can lock-in developers in their developer environment (AWS, GCP…) by creating a whole development ecosystem and professional recognition system around it (Google Developer Expert, Github Star…).

What are the competitive bottlenecks?

Unlike chips and infrastructure, which have been existing for decades, the market for foundation models is not even 2 years old. It is too soon to say whether anticompetitive practices have materialized, but it is already possible to point out to a number of risk:

Strings attached? Partnership between Big Tech and foundation models

Over the past year, a number of **strategic partnerships** have been announced between established tech companies and foundation model startups. Such partnerships are diverse in nature but tend to include a mix of **equity investments**, **access to chips and/or infrastructure** (in the form of preferred access clauses or exclusivity agreements) and featuring in the **large companies' marketplace or services**. Here's an overview of the main strategic partnerships recorded to date (March 1st 2024) :

Established company	Startup	Strategic partnership
Microsoft (US)	OpenAI (US)	Microsoft invests 13 billion USD (49% stake)[55] OpenAI has access to a dedicated Azure supercomputer (10,000 GPUs)[56] OpenAI is available on Azure's marketplace[57]

[55] Bradshaw, T et al. (2023). *How Microsoft's multibillion-dollar alliance with OpenAI really works. Financial Times*

[56] Maubant, T. (2023). *OpenAI réfléchirait à fabriquer ses propres puces d'IA.* ActuIA.

[57] Azure Marketplace, retrieved March 2024

Established company	Startup	Strategic partnership
Microsoft (US)	Mistral (FR)	Microsoft invests 15 million USD in convertible bonds (minority stake) Mistral has access to Azure infrastructure Mistral is available on Azure's marketplace[58]
Google (US)	Anthropic (US)	Google invests 2 billion USD (10% stake)[59] Anthropic available through Google Cloud Anthropic to use Google TPU v5e chips for AI inference[60]
Amazon (US)	Anthropic (US)	Amazon invests 4 billion USD (minority stake)[61] Anthropic to use AWS chips Anthropic available on AWS Amazon Bedrock[62]
Intel (US)	Stability.ai (US)	Intel invests 50 million USD Stability.ai to use Intel chips only[63]

All of the interviewed startups agree that strategic partnerships are healthy and welcomed for startups as long as they are not exclusive. Interviewed startups see the opportunity of being featured on the large company marketplaces as the main advantage of the partnership, as it gives their product exposure to a wide array of clients. The fact that both OpenAI and Mistral's models are available on the Azure Marketplace, for example, is positive as it shows that respective partnerships are not exclusionary. Moving forward, **authorities should ensure that all startups have an equal opportunity to be featured on marketplaces like Azure's,** that they can be **free to establish additional partnerships** with other distribution channels, including other cloud marketplaces, and **investigate whether cloud providers' marketplaces account as *gatekeepers*** as defined in the EU Digital Markets Act (DMA), as recommended in the European Parliament's 2023 report on competition in generative AI[64].

[58] Boyd, E. (2024). *Microsoft and Mistral AI announce new partnership to accelerate AI innovation and introduce Mistral Large first on Azure*. Microsoft Azure Blog

[59] Field, H. (2023). *Google commits to invest $2 billion in OpenAI competitor Anthropic*. CNBC

[60] Moss:, S. (2023). *Anthropic to use Google TPU v5e chips to train generative AI models*. Data Center Dynamics

[61] Amazon (2023). *Amazon and Anthropic announce strategic collaboration to advance generative AI*. Amazon

[62] Amazon (2023). *What you need to know about the AWS AI chips powering Amazon's partnership with Anthropic*. Amazon

[63] Smith, T. (2024). *Stability AI's Intel fundraise came with hefty hardware purchase commitments, sources say*. Sifted

[64] (2023). *European Parliament resolution of 16 January 2024 on competition policy – annual report 2023*. European Parliament.

While these strategic partnerships are certainly beneficial for AI startups in the short run, competition authorities should also remain vigilant as to their **potential side effects on foundation model startups in the long run.** For example, startups are faced with the risk of **lock-in with the established company's chips and/or infrastructure**. Moreover, the established company cloud platform could try to establish itself as the exclusive distribution channel for a certain foundation model. This could happen, for instance, if the current minority stakes of established companies become **majority stakes or are turned into fully-fledged acquisitions**.

On the latter point, it should be remembered that **acquisitions remain the most common exit opportunity for startups and one of the most preferred options by their investors**. Exits are an essential part of a startup lifecycle as they ensure a return to the investors that financed their growth. It should also be noted that in high-tech sectors like artificial intelligence, only a handful of corporations have the financial means and appetite to perform acquisitions and that these large companies are concentrated in America and Asia. Therefore, **in the event of acquisitions, we call on authorities to balance two equally important considerations:** on the one hand, the fact that acquisitions by established tech corporations will strengthen their vertical integration and thus further cement their market dominance; on the other, the need for startups to have attractive exit opportunities to pay back their investors. On the latter point, **policymakers should proactively ensure that alternatives to acquisition by Big Tech companies are available to startups.** Such alternatives include encouraging **acquisitions by European corporations** and creating a **European stock market** attractive enough to compete with the US.

Lock-in through vertical integration

Not all foundation model providers are new companies: incumbents are also active in this market. Google and Meta are cases in point. These companies are **vertically integrated from chips to infrastructure all the way to foundation models and applications**. This raises some risks.

First, vertically integrated companies could gain an **unfair advantage by combining user data from multiple products** that they offer to better sell their AI models.

Second, they could favor their own foundation models over those of competitors (**self-preferencing**).

Third, vertically integrated companies may be incentivized to **strategically limit access to their models to competitors in downstream or adjacent markets**. To prevent this, it is crucial to ensure that access to these general purpose AI models is granted under Fair, Reasonable, and Non-Discriminatory (FRAND) conditions for all market participants.

Keeping the distribution channels open: model hubs

Developers can today access foundation models via a variety of channels: the provider's website, cloud providers' marketplaces, and third party hubs. This diversity should be preserved to ensure user choice: **model marketplaces and hubs should not evolve into closed environments.**

There is a **risk of self-preferencing and user lock-in when off-the-shelf models are distributed via dominant cloud platforms**. The cloud provider could in fact try to favor the takeup of its proprietary foundation model over third party alternatives, or prevent user migration to a competitor. If the cloud platform controls everything from input data to model architecture, code, and weights it could technically impede interoperability or export of the model by the user.

Moreover, competition authorities should ensure that the **conditions (e.g. fees) required by hubs and marketplaces to distribute models are not abusive**: the sharing of value between platforms and third party model developers should be fair.

Consequently, **cloud providers should be the only way to access and deploy models that rely on their infrastructure and no exclusivity agreements should be allowed between model marketplaces/hubs and model providers**. In other words, competition authorities should remain vigilant as to model hubs not evolving **like app stores** in the mobile ecosystem and hubs thus not becoming "gatekeepers" as defined in the EU Digital Markets Act (DMA). In this regard, the existence of third party model hubs (e.g. Hugging Face) should be encouraged to prevent the abusive practices listed above and provide meaningful alternatives to marketplaces run by dominant companies (e.g. Amazon Bedrock).

💡**Focus: What the DMA says**

Art. 6(4): Third party app stores. *The gatekeeper shall allow and technically enable the installation and effective use of third-party software applications or software application stores using, or interoperating with, its operating system and allow those software applications or software application stores to be accessed by means other than the relevant core platform services of that gatekeeper. The gatekeeper shall, where applicable, not prevent the downloaded third-party software applications or software application stores from prompting end users to decide whether they want to set that downloaded software application or software application store as their default. The gatekeeper shall technically enable end users who decide to set that downloaded software application or software application store as their default to carry out that change easily.*

Art 6(5): Self-preferencing. *The gatekeeper shall not treat more favorably, in ranking and related indexing and crawling, services and products offered by the gatekeeper itself than similar services or products of a third party. The gatekeeper shall apply transparent, fair, and non-discriminatory conditions to such ranking.*

Diversifying data access options

As hinted previously, the market for data is characterized by a patchy legal framework, market concentration, and the presence of incumbents with entrenched positions. This leads to several competitive bottlenecks.

When it comes to the **legal framework**, the main bottleneck relates to the inconsistent **copyright rules that apply in different regions:**

- In the European Union, Directive 2019/790 (Copyright Directive) allows automated data mining. However, data holders choose to opt out, thus limiting access to valuable datasets.

- In the United States, the “fair use” doctrine creates an exception to copyright. However, its application to train generative AI has led to many not-yet-concluded litigations questioning its legitimacy.

- In Japan, a broad copyright exception allows for the training of both commercial and non-commercial generative AI models[65].

Data protection regulations also vary across regions, with the EU applying the most protective legal framework in the world. The uneven application of the EU General Data Protection Regulation (GDPR) and its strict interpretation by Data Protection Authorities, however, have made data sharing and access for European companies more difficult than in other regions. This has put them at a competitive disadvantage.

When it comes to user data, the market is concentrated in the hands of Big Tech companies (notably Amazon, Google, and Meta) running consumer-facing platforms and acting as gatekeepers for other companies to interact with their users. While the DMA partly addresses these issues, **according to interviewed companies gatekeepers continue to apply strategies to block or reduce access of third-parties to their users’ data**. This is done by adding new features or stopping technical support on key technologies, using potential risk to the protection of user personal data and security risk (most of the time exaggerated) as a legal ground to actually foreclose competitors.

The market for commercial data, instead, is in the hands of incumbents with entrenched positions. This includes Big Tech companies (like Google in the digitized books market), but also publishers (e.g. Axel Springer) and a vast network of right-holder intermediaries in the creative industries. **Startups entering the genAI market are faced with the risk of several anticompetitive contractual practices by right holders**, notably:

- **Abusive contractual terms**: a company with significant access to data (web index or search engines, for example) could deny or restrict access to data under its control. Similarly, such players could provide more favorable treatment to developers with whom a partnership has been established (for the provision of cloud or platform services, for example), or to their own in-house services. Also, companies having a

[65] Nishi, M. (2023). *Japanese Law Issues Surrounding Generative AI: ChatGPT, Bard and Beyond*. Clifford Chance

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 rights of data use 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 these data providers

- **Exclusive partnerships**: With exclusivity and priority clauses, large players could lock in data providers, preventing competitors' access. For example, this practice could take place in the form of high-priced contracts that *de facto* exclude smaller competitors with less financial resources.

Overall, interviewed startups and investors agree that the partnership model is not a long-term solution to the underlying problem: the **need to update copyright rules to the AI era and to generalize Fair, Reasonable, and Non-Discriminatory Contractual clauses in data access agreements**, as mandated by the EU Data Act.

💡 **What the Data Act says**

*Art 13: **Unfair contractual terms unilaterally imposed on another enterprise.***

1. A contractual term concerning access to and the use of data or liability and remedies for the breach or the termination of data related obligations, which has been unilaterally imposed by an enterprise on another enterprise, shall not be binding on the latter enterprise if it is unfair.

2. A contractual term which reflects mandatory provisions of Union law, or provisions of Union law which would apply if the contractual terms did not regulate the matter, shall not be considered to be unfair.

3. A contractual term is unfair if it is of such a nature that its use grossly deviates from good commercial practice in data access and use, contrary to good faith and fair dealing.

4. In particular, a contractual term shall be unfair for the purposes of paragraph 3, if its object or effect is to:
(a) exclude or limit the liability of the party that unilaterally imposed the term for intentional acts or gross negligence;
(b) exclude the remedies available to the party upon whom the term has been unilaterally imposed in the case of non-performance of contractual obligations, or the liability of the party that unilaterally imposed the term in the case of a breach of those obligations;
(c) give the party that unilaterally imposed the term the exclusive right to determine whether the data supplied are in conformity with the contract or to interpret any contractual term.

5. A contractual term shall be presumed to be unfair for the purposes of paragraph 3 if its object or effect is to:
(a) inappropriately limit remedies in the case of non-performance of contractual obligations or liability in the case of a breach of those obligations, or extend the liability of the

enterprise upon whom the term has been unilaterally imposed;
(b) allow the party that unilaterally imposed the term to access and use the data of the other contracting party in a manner that is significantly detrimental to the legitimate interests of the other contracting party, in particular when such data contain commercially sensitive data or are protected by trade secrets or by intellectual property rights;
(c) prevent the party upon whom the term has been unilaterally imposed from using the data provided or generated by that party during the period of the contract, or to limit the use of such data to the extent that that party is not entitled to use, capture, access or control such data or exploit the value of such data in an adequate manner;
(d) prevent the party upon whom the term has been unilaterally imposed from terminating the agreement within a reasonable period;
(e) prevent the party upon whom the term has been unilaterally imposed from obtaining a copy of the data provided or generated by that party during the period of the contract or within a reasonable period after the termination thereof;
(f) enable the party that unilaterally imposed the term to terminate the contract at unreasonably short notice, taking into consideration any reasonable possibility of the other contracting party to switch to an alternative and comparable service and the financial detriment caused by such termination, except where there are serious grounds for so doing;
(g) enable the party that unilaterally imposed the term to substantially change the price specified in the contract or any other substantive condition related to the nature, format, quality or quantity of the data to be shared, where no valid reason and no right of the other party to terminate the contract in the case of such a change is specified in the contract.

Abusive practices in hiring agreements

Over the years startups consistently indicate recruitment as one of their top three challenges to develop their business, with the short supply of required profiles and competition from Big Tech companies cited as the most recurrent obstacles[66]. The administrative complexity of cross-border hiring within the EU only exacerbates this problem[67]. While France Digitale's members have not directly experienced anti-competitive barriers to recruitment so far, it is relevant to note that Big Tech companies in the US are under increased scrutiny for two practices: **abusive non-compete clauses** and **no-poach agreements**.

A **non-compete clause** is a clause in an employment contract that forbids the employee to enter into or start a similar profession or trade that competes with the employer after the termination of employment[68]. While permitted under certain conditions both in the US and in Europe[69], **academic research has demonstrated that non-compete clauses have been used anti-competitively by high tech companies in the US[70]**, leading the Federal Trade

[66] EY x France Digitale (2023). *2023 Barometer - Social and Economic Performance of French startups*. France Digitale.
[67] *France Digitale (2023). Our Manifesto for the 2024 European Elections.* France Digitale
[68] Posner, E. and Volpin, C. (2023). No-poach agreements: an overview of EU and national case law. 4 May 2023, e-Competitions No-poach agreements, Art. N° 112194.Concurrences
[69] Legal and Administrative Information Directorate, Prime Minister (2023). *What is a non-compete clause?* ServicePublic.fr
[70] Balasubramanian N. et al (2022). *Locked In? The Enforceability of Covenants Not to Compete and the Careers of High-Tech Workers*. 57 J. of Hum. Res. S349.

Commission to propose a non-compete clause rule in 2023[71]. If adopted, the rule will ban non-compete clauses as an unfair method of competition unless the restricted party owns at least 25% of the business. Other restrictive employment clauses like Non-Disclosure Agreements and customer non-solicitation agreements, instead, will remain lawful.

A **no-poach agreement** is an agreement between companies not to hire away each other's employees and it is illegal in both Europe and the US except in the case of joint ventures[72]. Here, too, there is evidence of this practice in the high tech sector, with a case brought up by the US Department of Justice in 2010 against Apple, Adobe, and Google ending up in a settlement[73]. While there have been no cases in Europe of non-poach agreements among tech companies yet, we believe that this is an area that needs further scrutiny and that startup feedback in this regard would be particularly valuable to the regulator.

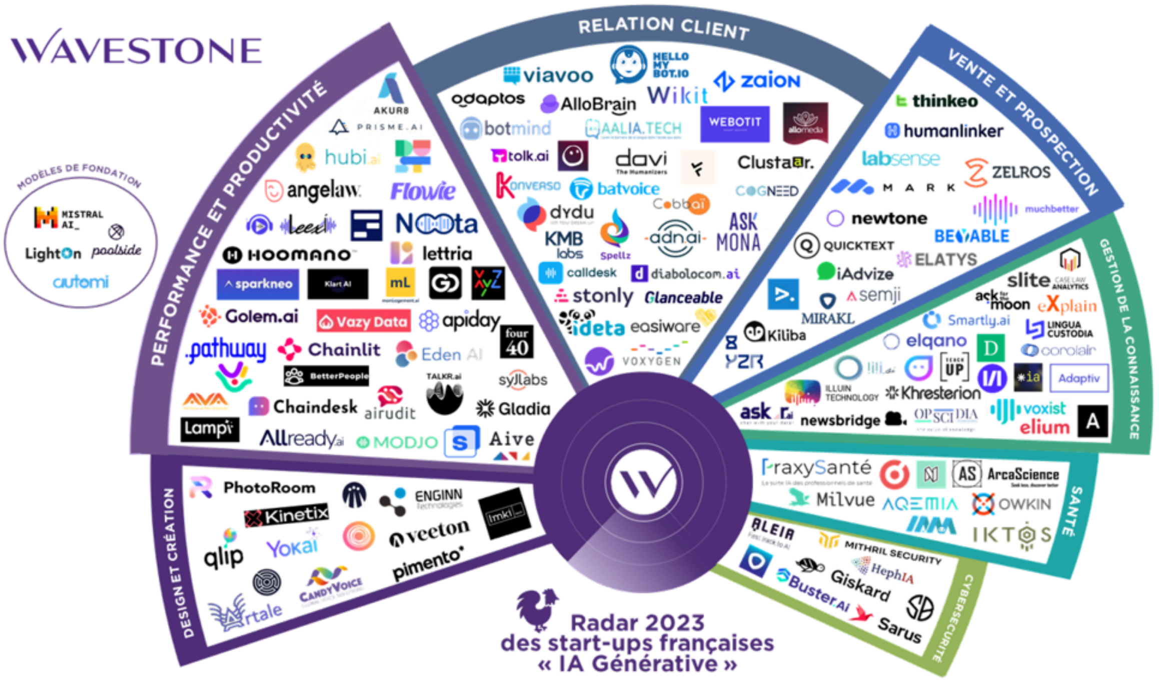
[71] Notice of Proposed Rulemaking, *Non-Compete Clause Rule* 88 Fed. Reg. 3482 (Jan. 19, 2023) (to be codified at 16 C.F.R. § 910 (2023)

[72] Posner, E. and Volpin, C. (2023), No-poach agreements: an overview of EU and national case law, 4 May 2023, e-Competitions No-poach agreements, Art. N° 112194, Concurrences

[73] *Idem*

Layer 4 - Applications: a vibrant market in the shadow of Big Tech

2023 Mapping of French generative AI startups



[Source: Wavestone](#)

How does it work?

Application providers play a crucial role in the development and deployment of applications for end users. Through applications like chatbots, document automatic completion, and image correction tools, foundation models are put to use in concrete use cases. Application providers can be divided into two subcategories:

- **End-to-end application providers:** Some application providers develop their own custom AI models to meet the specific needs of their applications.
- **Third-party model integrators:** Others use pre-trained AI models developed by third parties to add to the functionality of their applications. Developers of these applications can integrate the AI model in its original form or adapt it using various techniques (fine tuning, Retrieval Augmented Generation (RAG), etc.).

They can further be distinguished between general purpose applications (e.g. OpenAI's ChatGPT, which can answer questions on virtually any topic) and specialized applications. (e.g. Quicktext's Velma chatbot, which can only answer hotel-related queries).

What are the competitive bottlenecks?

Applications is the stage of the supply chain where competition is most intense and where the highest number of companies are active. In France alone, there are over 135 generative AI startups as of January 2024[74], up from 86 in the same period the year before[75]. There is a consensus among interviewed startups and investors that the most value is to be gained in the market for highly specialized, rather than general purpose applications due to the presence and predominance of incumbents in the general purpose market. Indeed, at this stage of the supply chain, startups face competitive issues with incumbents on two levels: app development and app distribution.

When it comes to app development, it should be noted that **vertically integrated companies (e.g. Google, Microsoft) are directly competing with startups**, **especially when it comes to general purpose apps**. Google, for example, is directly developing the Gemini chatbot, while Microsoft developed the Copilot AI assistant. There is a risk here that by integrating (**bundling**) such applications into its existing popular services (the 365 and Google suites, respectively), these companies may be giving an unfair advantage to their own products, a practice identified as **self-preferencing** and forbidden by the DMA. To ensure the contestability of this market and offer an equal chance for alternative services to emerge, the installation and integration of alternative applications should always be allowed on systemic platforms. To this end, an investigation to clarify whether services like Google Suite and Microsoft 365 qualify as *gatekeepers* as defined in the DMA should be carried out.

💡**What the DMA says**

Art. 6(4) Access to third-party software. *The gatekeeper shall allow and technically enable the installation and effective use of third-party software applications or software application stores using, or interoperating with, its operating system and allow those software applications or software application stores to be accessed by means other than the relevant core platform services of that gatekeeper. The gatekeeper shall, where applicable, not prevent the downloaded third-party software applications or software application stores from prompting end users to decide whether they want to set that downloaded software application or software application store as their default. The gatekeeper shall technically enable end users who decide to set that downloaded software application or software application store as their default to carry out that change easily.*
The gatekeeper shall not be prevented from taking, to the extent that they are strictly necessary and proportionate, measures to ensure that third-party software applications or software application stores do not endanger the integrity of the hardware or operating system provided by the gatekeeper, provided that such measures are duly justified by the gatekeeper.
Furthermore, the gatekeeper shall not be prevented from applying, to the extent that they are strictly necessary and proportionate, measures and settings other than default settings, enabling end users to effectively protect security in relation to third-party software applications or software application stores, provided that such measures and settings

[74] Hantouche, C. et al. (2024). Generative Artificial Intelligence: 2023 radar of French "GenAI" startups, Wavestone

[75] Tordeux, M. et al. (2023). *Mapping 2023 des startups françaises de l'IA*. France Digitale

other than default settings are duly justified by the gatekeeper.

***Art 6(5) Self-preferencing**. The gatekeeper shall not treat more favourably, in ranking and related indexing and crawling, services and products offered by the gatekeeper itself than similar services or products of a third party. The gatekeeper shall apply transparent, fair and non-discriminatory conditions to such ranking.*

***Art 6(6) Software switching.** The gatekeeper shall not restrict technically or otherwise the ability of end users to switch between, and subscribe to, different software applications and services that are accessed using the core platform services of the gatekeeper, including as regards the choice of Internet access services for end users.*

When it comes to app distribution, startups have a dual relationship with incumbents. On the one hand, they benefit and actively aim to be referenced in incumbents' marketplaces, including infrastructure providers, as it gives them exposure to their user base. A partnership like Mistral's with Microsoft Azure or Voxist's with OVHcloud is therefore highly welcomed among startups as a client acquisition avenue.

On the other hand, this raises questions as to the role of incumbents being an entry point to find clients for startups, which may in turn lead to their classification as gatekeepers under the DMA. This, in turn also raises questions, partly discussed above, as to which and how third-party services are referenced in incumbent platforms. So far, startups are working through bilateral partnerships. These partnerships must remain non-exclusive and are negotiated and drafted in FRAND terms.

Distribution through app stores, instead, should continue to be regulated under the DMA to ensure fairness.

Glossary

Application-Specific Integrated Circuits (ASICs): type of integrated circuit designed for a specific task or use case.

Bundling: the practice of selling multiple products and/or services in a single package deal.

Central Processing Units (CPUs): type of integrated circuits responsible for executing instructions and performing calculations.

Chip programming model: the “operating system” of chips, software used to program and control their behavior.

Coopetition: strategic alliance between rival organizations, often in the software and hardware sectors, in the hope of mutual benefits, sometimes on a specific project.

Core: unit within the chip capable of executing instructions and performing calculations. The utilization of multiple cores facilitates parallel processing, thereby enabling the chip to execute multiple tasks simultaneously.

Deep learning: Artificial neural networks are composed of multiple layers to learn and extract features from data. Deep learning is a subset of machine learning.

Electronic Design Automation: software tools to design electronic systems, including integrated circuits.

Fine-Tuning: process of further training a pre-trained model on a specific task to improve its performance for a particular purpose

Floating Point Units (FPUs): type of integrated circuit specialized for complex mathematical operations, mainly used to train AI models.

Foundry: a high specialized facility where integrated circuits are manufactured

Gatekeeper: a company providing core platform services

Graphic Processing Unit (GPU): type of integrated circuit with multiple cores, designed to handle multiple computing processes at the same time (“parallel processing”). Mainly used to train AI models.

Hyperscaler: a company that offers a wide range of services (Iaas, PaaS and SaaS) and a massive infrastructure capable of computing on a large scale.

Inference: the process by which a pre-trained model generative AI model creates new content based on an external given set of data.

IP provider: a company offering intellectual property (IP) that can be integrated into chip designs or components.

Model training: process of creation of a generative AI model by applying machine learning and deep learning architectures to a given set of data.

No-poach: agreement between companies not to hire away each other’s employees.

Prompt: input or instruction provided to an AI system to generate a response or perform a task.

Retrieval-Augmented-Generation (RAG): the process of giving to a pretrained model a particular dataset to train him for more accurate or optimized responses.

Self-preferencing: anticompetitive behavior in which a company prioritizes its own products or services over those of a third party.

Vertical integration: strategy where a company controls multiple stages of the supply chain of a product or service.

Video Random Access Memory (VRAM): type of memory specifically designed for use in GPUs optimized for high-speed data transfer.

Methodology

This study is based on desktop research and qualitative interviews with 36 companies: 30 startups and scaleups (including two chip designers), 4 venture capital funds and 2 European cloud providers.

List of interviewed companies

1. Alpha Intelligence Capital
2. Buster.ai
3. Case Law Analytics
4. Cleyrop
5. Criteo
6. Doctrine
7. Dust
8. Elaia
9. Emocio.hr
10. Explain
11. Flex.ai
12. Flowie
13. Giskard
14. Glanceable
15. Gleamer
16. Golem.AI
17. Hugging Face
18. La Forge
19. LinguaCustodia
20. IRIS
21. Lampi.AI
22. Malt
23. Mistral
24. PhotoRoom
25. :probabl
26. OVHcloud
27. Quicktext
28. ReciTAL
29. Scaleway
30. Serena
31. SiPearl

32. Stonly
33. Suzan.ai
34. Voxist
35. Welcome to the Jungle
36. XXII

Contacts/Authors

Marianne Tordeux Bitker, Public Affairs Director marianne@francedigitale.org
Agata Hidalgo, Public Affairs Lead agata@francedigitale.org
Gaël Gutierrez, Policy Analyst gael@francedigitale.org

About France Digitale

Founded in 2012, France Digitale is the largest startup association in Europe, bringing together over 2000 startups and investors (venture capitalists and business angels).

The association's goal is to help build Europe's future tech champions by uniting and raising the voice of those who innovate to change the face of the world.

France Digitale is co-presided by Frédéric Mazzella, Chairman and founder of BlaBlacar, and Benoist Grossmann, CEO of Eurazeo Investment Manager.

www.francedigitale.org