COMPETITION IN THE ARTIFICIAL INTELLIGENCE TECH STACK

Recent developments and emerging issues

Prepared by AGCM Staff for the discussion at the G7 Competition Summit 2024

I. Introduction

1. This document outlines recent developments in the artificial intelligence (AI) stack and the emerging competition issues. It has been prepared by the staff of the Italian Competition Authority (AGCM). This document is based on the publicly available studies and materials published by the G7 competition authorities[1] and does not necessarily reflect the views of any individual G7 competition authority, including the AGCM.

2. In order to inform the discussion at the G7 Competition Summit held in Rome on October 3-4, 2024, the AGCM established a G7 Competition Working Group on Artificial Intelligence (AI WG), which included one or more staff level representatives (managers, officials) from each of the G7 competition authorities. The AI WG served as a forum for mutual updates and exchange of knowledge on the respective research, studies and horizon scanning activities related to the AI industry and in particular to Generative AI.

3. At the conclusion of the Summit, the G7 competition authorities and policymakers issued a statement on competition in AI markets, which describes a shared commitment to enforce competition laws and policies necessary to ensure that principles of fair competition are applied to nascent AI markets.

II. Key Features and Current Trends in the Generative AI Technology Stack

4. The AI **tech stack** (**see Glossary**) has experienced rapid growth and significant advancements in recent years, driven by breakthroughs in machine learning algorithms, increased computational power including cloud services, and the availability of vast amounts of data. Generative AI refers to AI that can create new content, such as text, images, music, and videos,

[1] More specifically: the UK CMA Report AI Foundation Models Review: Short Version (September 2023); CMA AI Foundation Models: Technical Update Report (April 2024); US FTC Office of Technology Blog articles; Opinion 24-A-05 on generative AI and Opinion 23-A-08 on competition in the cloud sector of the French Autorité de la Concurrence; Japan FTC Report on Trade Practices in Cloud Services Sector (2022).

based on the data they have been trained on or through the continuous inference process where they analyse that data for a specific outcome.

5. **AI systems** are indeed becoming integral to many sectors, especially to digital markets, enabling businesses to automate processes, analyse vast amounts of data, and support informed decision-making. AI-powered chatbots, predictive analytics, and personalized recommendations are just a few examples of how these technologies are enhancing customer experiences and optimizing operations. As AI continues to advance, its applications primarily, but not exclusively, in digital markets are expected to expand further.
6. The generative AI stack encompasses the entire process of creating and delivering AI-generated content. **Foundation Models (FMs)** sit at the heart of this chain, acting as the powerful engines driving innovation. FMs are a type of AI technology that are trained on vast amounts of data that can be adapted to a wide range of tasks and operations. They provide the basis for generative AI systems. Once trained, FMs can be fine-tuned for specific tasks, making them highly customizable for various applications, services and content generation[2].
7. FMs development relies on: data to train and fine-tune the models, computational resources (chips, supercomputers, cloud services), technical expertise (highly skilled research scientists and engineers) and capital to fund the computational resources needed and the training costs. In turn, the development of Generative AI applications and services rely on access to foundation or more specialised models (e.g. for a specific task like coding or with data from a specific use case).

III. Requirements for the development of generative AI services

III.1 Access to data

8. Vast quantities of data are integral to build the knowledge of the FM in the so-called pre-training phase while in the fine-tuning and inference phases FMs are trained on a smaller but very specific datasets and are optimised for some specific tasks.
9. FMs developers are exploring a range of data types to support FMs:

    - ✓ Publicly available data: open-source datasets and web-scraped information remain significant contributors, particularly during the pre-training phase where data diversity is crucial.
    - ✓ Third-party proprietary data: there is a growing trend of accessing data through partnerships and licensing agreements; this is likely driven by concerns over intellectual property (IP) enforcement and the desire for higher quality data sets.
    - ✓ Synthetic data: artificially generated data, including that produced by other FMs, has been utilized in both pre-training and fine-tuning stages. However, the effectiveness and potential drawbacks of synthetic data, such as model collapse and data quality issues, remain areas of ongoing research.
    - ✓ Data feedback loops: the potential for real-time data gathered from downstream services to improve FMs can become increasingly important for inference. However, its current usage and impact on model performance are yet to be definitively established.

[2] Content can be delivered in multiple modalities i.e., text, images, videos, and 3-D representations (such as scenes and landscapes for video games).

Competition Questions

10. For innovation and competition to flourish, a variety of viable data sources appear paramount. However, concerns about access might emerge:
    - ✓ Data advantage for large digital companies: these companies might leverage their financial resources and market position to secure specialised data through partnerships, leaving smaller FMs developers at a disadvantage.
    - ✓ Web-Scraping Challenges: Future copyright and data protection issues could restrict access to web-scraped data, putting later entrants at a disadvantage compared to early movers.
11. At the deployment stage, data requirements seem to become more specific and demanding. Access to high-quality, specialised data could become a key competitive differentiator:
    - ✓ Exclusivity agreements and vertical relationships: Exclusivity agreements or vertical relationships across the AI value chain could restrict access to high-quality data for new entrants and increase the costs for training or fine-tuning FMs. Firms with access to proprietary or privileged data (including "return output/monitoring" data) might gain advantage in the development and deployment of specialised models and AI applications, e.g., in the health sector. If firms are vertically integrated and have access to data through other digital services they provide, especially consumer data, this could also put them at an unfair advantage against smaller firms without that access.
    - ✓ **Data feedback loop** advantage: If data feedback loops prove crucial for model improvement, new developers might be unable to fine-tune their models with the necessary data, hindering their competitiveness.

III.2 Access to chips

12. Semiconductor chips, commonly known as **chips** are essential for providing the computing power and memory needed to create and run software systems, including AI systems. These chips seem crucial insofar as AI systems require a lot of computing power for two main tasks: training and inference.
13. Data-parallel numerical computations required to train and run FM and AI systems it is not typically supported by conventional general-purpose chips like **CPUs** (Central Processing Units). Instead, most FMs are trained and run on **AI accelerator chips**, including chips for general computing (**GPUs**) and those for specific tasks (such as **FPGAs** - Field Programmable Gate Arrays - and **ASICs** - Application-Specific Integrated Circuits), which are used for the development and deployment of more complex FMs and AI systems. Their use differs according to computing requirements, in terms of architecture, workload performance, flexibility, tasks specificity and power consumption.

14. The recent surge in demand for Generative AI services and the market preference for specific chip developers has created a major shortage of these accelerator chips, making it harder for new companies to enter the market. Because of this shortage, developers of FMs and large cloud computing companies seem to be looking for ways to make their own chips or partner with others, thereby stimulating the creation of specialized chips designed to perform given AI tasks, and to reduce their reliance on external providers.
15. Notwithstanding the efforts of large digital companies to design custom chips for internal use, no significant demand shift has been observed so far, meaning that the market structure and dynamics in this layer appear unlikely to change drastically in the short term. Innovation efforts in new chips may as well reinforce a captive use in proprietary integrated ecosystems, namely cloud services.
16. While start-ups are investing in creating new AI chips, there may be an over-specialization risk for these newcomers. Designing a chip usually takes two to three years, which could be a long time in the fast-moving world of AI. By the time these new chips are ready, they could be specialized for less popular applications, even if they outperform current GPUs in some areas, making them less useful and less competitive in the market.
17. Similarly, the competitive landscape for **chip programming models**, which are necessary to instruct chips to execute tasks adapting to a wide range of software, is unlikely to change in the near term as they are largely on incumbent proprietary models. In future, open-source frameworks might become more available and interoperable, despite considerable switching costs.

Competition Questions

18. Agreements or partnerships for supplying or co-designing chips and their programming models could strengthen incumbents' market position. This might allow them to use their power in the market to limit competition, make it harder for customers to switch to other options, or discriminate among customers.

III.3 Access to cloud computing

19. The cloud industry provides on-demand computing power and storage solutions. It appears to be characterized by a critical size, high costs, and economies of scale and scope, benefiting major digital players. These players seem to enjoy conglomerate benefits namely from their advantages in cloud ecosystems.
20. **Cloud service providers (CSPs)** are integral to the Generative AI sector, as long as they leverage their extensive cloud infrastructure to offer vital compute and integrated services to FMs developers and deployers. For start-up or smaller AI developers, CSPs appear to be crucial as they provide resources beyond mere cloud computing, such as specialized hardware and software for development and deployment of FMs.

Competition Questions

21. CSPs might implement clauses or practices that could act as lock-in mechanisms, increasing barriers to switching providers or limiting commercially available options (barriers to migration and to expansion). These provisions may incentivize customers to consolidate most, if not all, of their cloud needs with a single provider, even if other

providers offer superior services in certain areas. Practices that seem to deserve a careful scrutiny include egress fees, which are charges users incur when moving their data out of one cloud provider to another. Other examples include cloud credits (that is, allocations of cloud services accessible for free for a certain period), discounts and minimum spend and technical barriers to interoperability and data and app portability.

22. As for FMs development, there might be few viable alternatives without partnerships with CSPs. The scarcity of chips, along with the lack of sufficient computational infrastructure, could drive up computing costs and increase reliance on cloud infrastructure services, particularly for new entrants and small FMs developers. Exclusive cloud partnerships or unfair and discriminatory access conditions might further exacerbate this issue.
23. As for deployment, small FMs deployers might struggle to find reliable resources outside of CSPs, and their access to users and their choice of FMs might depend on the FM distribution platform of their CSPs. Exclusive partnerships between CSPs and FM developers could undermine competition among FM platforms, potentially limiting innovation and diversity in the AI market. Investment in public supercomputers, designed to tackle complex and computationally demanding problems, accessible under fair and non-discriminatory conditions to private players may alleviate reliance on private computing resources including cloud services.
24. Further, some CSPs are also AI developers, and so this vertical integration across the FM technical stack could mean that the CSPs reduce access for other firms they see as in direct competition with their AI services.

III.4 Access to talent

25. Another key input for generative AI is labour expertise. Developing a generative model can require a significant engineering and research workforce with particular - and relatively rare - skills, as well as a deep understanding of machine learning, natural language processing, and computer vision.

Competition Questions

26. It might be difficult to find, hire, and retain the talent required to develop FMs for smaller developers that lack the necessary funding. Large technology companies might have an incentive to buy start-ups to scoop up their talent or acquire their talent through deals or partnerships which are not necessarily subject to scrutiny from competition authorities.

III.5 Access to foundation models (FMs)

27. Access to FMs seems crucial for providing downstream generative AI services. The number and diversity of FMs globally are increasing, with variations in size, input, resource requirements, performance levels, and specialization.

28. Cases of strategic collaboration between big-tech ecosystems and FMs developers have been reported, both in the form of long-term partnerships under exclusivity conditions for cloud services and strategic acquisition of minority shareholdings.
29. Developing larger, more capable FMs is still ongoing but will require significant data, compute resources, and high training costs, so there seems to be a drive to create and deploy smaller models with extensive capabilities but fewer resource requirements. The development of smaller FMs is also likely to facilitate their deployment on consumer devices, reducing reliance on cloud infrastructure. This seems supported by the decreasing size of FMs and the increasing availability of specialized AI chips for personal devices like smartphones and tablets.
30. **Closed and open source FMs** exist on a spectrum, creating a diverse landscape of model development. Companies also choose different business models and monetization strategies as they develop their products.

    ✓ Access to pre-trained models that sit on the open source spectrum appears to be relatively common and has positively contributed to advancements in generative AI. In closed access formats, third parties can access FMs via APIs or plugins, typically relying on controlled, paid-for access and strong licensing conditions.

    ✓ Some of the most prominent FMs are open source, which appears to be fostering competition and innovation. These models could allow firms with innovative ideas to develop new models and improve existing ones. However, it appears that some high-performing open-source models are developed by well-funded firms with significant resources, which could make the open-source AI ecosystem reliant on large digital companies or financially strong firms.

31. Models are often released on various platforms, often hosted by large technology/digital firms. These FMs platforms or marketplaces support multiple modalities for accessing FMs, providing a broad range of models for deployers to choose from. Although still in the early stages, these platforms might exert some control over FM distribution in the future.

Competition Questions

32. Open source models might become closed in the future, which might lead to a concentration of models within a few firms and potentially stifle innovation.

IV. Development of Generative AI services and applications in other markets

33. The downstream layer of the Gen AI stack consists of applications and tools tailored to be used across a variety of tasks. FMs are used in a variety of applications across a wide range of industries improving existing products and services or creating new ones which may have, on the one hand, the potential to disrupt markets and existing market power and, on the other hand, potentially create new or entrench existing positions of market power for the firms that develop a specific product or service.
34. Firms with market power in upstream markets or vertically integrated firms might be able to leverage that market power through input foreclosure practices to restrict competition in FM deployment, leading to limited choice for downstream customers. For instance, the developer

of a foundation model may give its own downstream AI services exclusive access to the best version of the FM, or the owner of must-have applications and operating systems might tie downstream users to certain models and certain cloud platforms. Potential licensing restrictions — such as limitations imposed through APIs, restrictions on licensees' commercial applications or other development uses — could negatively impact innovation and shape consumers preferences limiting their choice.

35. Similarly, concentration and barriers to entry and expansion in downstream markets might arise where customer foreclosure strategies are implemented so that downstream customers have difficulty switching between FMs or FM and AI products and services e.g. because they are locked into ecosystems that only offer a limited range of FM deployment options or products and services. For example, the integration of generative AI tools in certain devices, such as smartphones or laptops, could consolidate the AI industry around some prominent digital companies.
36. All these strategies might increase dependency on particular foundation models for the development of downstream applications. As the foundation model's technology is integrated into more apps, services, and products, the control of key inputs may offer large players a favourable bargaining position and the ability to influence the evolution of technologies and innovation in secondary markets.
37. High-performing closed-source models may also create positive feedback loops through API access, attracting advertisers, app developers, users, and smart device manufacturers. Competing FMs developers might struggle due to lack of access to the model's code, training data, or detailed model parameters. In other words, a weakening of competition in the FMs markets could have a cascading impact on the fine-tuning models and, ultimately, on generative AI applications and services available to end users. Thus, the prevalence of proprietary FMs could affect innovation and diversity in AI applications, also impacting the broader AI industry and consumer choices.

V. Algorithmic collusion

38. The use of AI and algorithms could facilitate collusion between firms, making it easier for them to coordinate prices, share competitively sensitive information, and undermine competition. Competitors' joint use of common algorithms could remove independent decision-making. The opacity of these algorithms may potentially enable companies to collude while making it more difficult for competition authorities to detecting such behaviour[3].
39. Additionally, AI technologies could be used to engage in surveillance pricing and set individual prices by using expansive and highly personal and sensitive information, which could be detrimental for consumers.

VI. Other non-competition concerns: consumer protection

40. The new applications of AI seem to have high potential to enhance consumers' experiences with products and services. At the same time, AI-generated outputs could introduce or reinforce biases, mislead consumers, shape their preferences, and prevent them from making informed choices. For instance, when seeking product recommendations, consumers might get chatbot answers driven by commercial incentives rather than the best options for their

[3] See the joint study of *Bundeskartellamt* and *Autorité de la concurrence*, Algorithms and Competition, November 2019.

needs. Additionally, consumers may not know what data is being collected about them to drive these recommendations, and they may have no choice in how their data is used.

41. Other potential issues include the possibility of AI reducing the cost of creating fake reviews and other mechanisms to influence consumers, which could undermine feedback mechanisms aimed at improving consumer's ability to make an informed choice. All these risks could impact overall market confidence—a key component of competition.

VII. Concluding observations

42. Across the various layers of the AI stack, market concentration within each layer and vertical integration throughout the stack seem to present potential upstream risks that could negatively affect the sector development as well as the downstream layers, ultimately leading to consumer harm. A small number of large technology/digital companies dominate the field, leveraging their substantial resources and key inputs access to maintain a competitive edge.
43. With respect to the development and deployment of FMs, closed ecosystems might influence the direction and rate of innovation in FM markets, insofar as they could affect their contestability and the distribution of remuneration across the value chain.
44. Promoting a contestable and competitive FM environment seems to rely on ensuring fair access to essential inputs, like data and computational infrastructure, in order to maintain and promote choice and innovation. If access to these inputs is restricted or offered on unfair commercial terms, this could lead to significant concerns. First-mover advantages, such as network effects, feedback loops, and platform effects, combined with vertical integration in related markets, might reduce the incentives for incumbents to compete on merits and instead increase their ability to engage in anticompetitive behaviours.

Glossary

AI accelerator - Specialised computer chips designed to process AI and machine learning computations faster than generic chips.

AI system - A machine-based system 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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment (OECD Recommendation on Artificial Intelligence (amended in May 2024)).

AI tech stack - The different layers of technologies and components that make up an AI system.

ASICs (Application-Specific Integrated Circuits) - Integrated circuits that are specifically designed and tailored for a particular application or use.

Chip programming models - Tools or frameworks that help developers write software to run on specific types of computer chips. These models are important because they help make sure that software runs as fast and efficiently as possible on the chosen hardware. Chip programming models are like special sets of instructions that help developers create software that runs really well on specific types of computer chips, making technology work faster and more effectively.

Chips - The essential electronic components of all computing technology, including generative AI systems. Chips provide the computing power and memory functions to develop and deploy software systems, including Generative AI systems which need computing power to perform two main functions: training (applying machine learning and deep learning architectures to a given set of data) and inference (by creating new content).

Closed-source models are usually developed privately within companies, and access to the models, as well as information about them, is more controlled and shared only to the extent that the company chooses.

Cloud service providers (CSPs) A CSP (cloud service provider) is a company that provides cloud computing services. Cloud computing services allow businesses and individuals to access computing resources, such as CPUs, GPUs, and storage, on demand.

Computational resources - The amount of processing power required to train and deploy AI models. Compute can be provided by a variety of resources, including CPUs, GPUs, and specialised AI accelerators.

CPU - A CPU (central processing unit) is the main processing unit of a computer. It is responsible for carrying out the instructions that are stored in the computer's memory.

Data feedback loops or effects – They refer to the ability of FMs and FM developers to use data generated by their usage to improve their performance.

Foundation Models (FMs) are a type of AI technology that are trained on vast amounts of data that can be adapted to a wide range of tasks and operations. Most FMs are currently being developed using a deep learning model called a transformer.

FPGAs (Field Programmable Gate Arrays) - A type of semiconductor that can be programmed and reprogrammed according to customer's design and device needs.

GPU - A GPU (graphics processing unit) is a specialised processor that is designed for graphics processing. GPUs are increasingly being used for AI applications, as they can provide significant performance improvements over CPUs for certain types of AI workloads.

Open-source models – FMs that are freely shared, and can be used at no cost, subject to their licenses (which can prohibit commercial use). An open-source release can consist of the underlying code, model architecture, and training data, enabling others to replicate the training process. In some cases, it also includes the weights and biases (i.e., the 'knowledge') of the model, such that others can use or fine-tune the model without conducting their own pretraining.