**Microsoft Corporation's Response to the European Commission's**
**Call for Contributions on Competition in Generative AI dated 9 January 2024**

(*Submitted 11 March 2024*)

Like other general-purpose technologies in the past, AI is creating a new sector of the economy. This new "AI economy" is creating not just new opportunities for existing enterprises, but new companies and entirely new business categories. Today, datacenters around the world house millions of servers and make computing power broadly available to organizations large and small and even to individuals as well. Already, many thousands of AI developers – in startups, enterprises, government agencies, research labs, and non-profit organizations around the world – are using the technology in these datacenters to create new foundation models and AI-based applications.

There are many new generative AI entrants like Anthropic, Cohere, Aleph Alpha, and Mistral AI that are developing and operationalizing generative AI-based technology for the public and private sectors. In addition, Microsoft, along with other large technology firms are dynamically pivoting to meet the AI era. The competitive pressure is fierce, and the pace of innovation is breathtaking.

All organizations in almost every industry will use foundation models to operate more efficiently and drive product development. There is a rapidly growing number of use cases and demand for models powering those use cases. This environment provides a strong platform for supporting broad investment in differentiated AI foundation models. Although some of the largest general purpose foundation models have captured public attention, smaller specialized foundation models are playing an increasingly important role in recognition that they can offer the performance needed at a much better cost proposition to customers.

Partnerships have played and will continue to play an important role in enabling this entry. Today, only one company – Google – is vertically integrated in a manner that provides it with strength and independence at every AI layer from chips to a thriving mobile app store. Everyone else must rely on partnerships to innovate and compete. Since 2019, Microsoft has collaborated with OpenAI on the research and development of OpenAI's generative AI models, developing the supercomputers needed to train those models. The ground-breaking technology ushered in by our partnership has unleashed a groundswell of innovation across the industry. And over the past five years, OpenAI has become a new, significant, independent competitor in the technology industry. It has expanded its focus, commercializing its technologies with the launch of ChatGPT and the GPT Store and providing its models for commercial use by third-party developers.

Innovation and competition will require an extensive array of similar support for proprietary and open-source AI models, large and small, including the type of partnership we announced last month with Mistral AI, a leading open-source AI developer based in France. We have also invested in a broad range of other diverse generative AI startups. In some instances, those investments have provided seed funding to finance day-to-day operations. In other instances, those investments have been more focused on paying the expenses for the use of the computational infrastructure needed to train and deploy generative AI models and applications. We are committed to partnering with market participants around the world and in ways that will accelerate local AI innovations.

Against this dynamic backdrop, it is important that competition authorities remain engaged, while giving the evolving markets and technologies the room to develop and emerge. Industry can play an important role in promoting innovation and competition by providing the access needed for organizations and individuals to develop and use AI in ways that will serve the public good. Microsoft's AI Access Principles[1] are one such approach. And governments must also work together with industry to develop policies that will ensure competition remains vibrant and that all members of society benefit.

* * * * * * *

**1. What are the main components (*i.e.*, inputs) necessary to build, train, deploy and distribute generative AI systems? Please explain the importance of these components.**

1.1 The following components of the AI technology stack underpin the building, training, deployment, and distribution of foundation models ("**FM**s"), which enable generative AI systems[2]:

- **Semiconductors.** Semiconductors, specifically Graphical Processing Units ("**GPU**s"), play a crucial role in the training and deployment of FMs. They facilitate the complex computations required in AI processes and enable the practical application of AI technologies in various industries. They are capable of handling thousands of threads simultaneously, making them ideal for the parallel processing requirements of machine learning algorithms.

  NVIDIA is the undisputed leader in the production and supply of GPUs, which are required for the training, building and deployment of AI models and FMs in particular. Its share is estimated to be over 70% and could be as high as 90%, with competitors still years behind.[3] Governments have stepped in to spur new entry and increase production. There are start-ups emerging who are innovating on new approaches. For example, Groq is a U.S.-based start-up that is developing a Language Processing Unit, a semiconductor that it claims "*provides the fastest inference for computationally intensive applications with a sequential component to them, such as AI language applications.*"[4] Existing traditional semiconductor firms, such as Intel and AMD, are accelerating their research, development, and production of GPUs to compete with NVIDIA.[5] And cloud-computing providers, like Amazon and Microsoft, are designing and developing their own AI-optimized

---

1 Available at *here*.

2 As defined by the European Commission ("**Commission**") in the context of this consultation alternately as "*AI system[s] that [are] able to produce new content, such as texts, images or other media*" (Consultation), and "*AI systems that generate, in response to a user prompt, synthetic audio, image, video or text content, for a wide range of possible uses, and which can be applied to many different tasks in various fields*" (Press Release). *See* European Commission, "Competition in Virtual Worlds and Generative AI: Calls for contributions", accessed March 8, 2024, *available here*, p.1 ("**Consultation**"); *See also* European Commission Press Release, "Commission launches calls for contributions on competition in virtual worlds and generative AI", January 9, 2024, *available here* ("**Press Release**").

3 *See* Wccftech, "Nvidia's global AI chip market *share* reaches a whopping 90%; analysts say competitors years from catching up", January 24, 2024, *available here.*

4 *See* Groq, "Why Groq?" accessed March 8, 2024, *available here*.

5 *See* Kif Leswing, "Intel unveils new AI chip to compete with Nvidia and AMD", December 15, 2023, (Intel released Gaudi3 a chip aimed at competing with rival chips from Nvidia and AMD that power AI models), *available here*; *See also* Asa Fitch, "AMD Rolls Out New Chips, Aiming for Nvidia's AI Crown", December 6, 2023, *available here*.

chips to provide themselves with more options to support and meet the demand for AI infrastructure that they offer to third-parties.

As an AI infrastructure provider, Google is uniquely situated. It has developed Tensor Processing Units ("**TPUs**"), an AI-optimized semiconductor designed for use with Google's TensorFlow framework software. It reportedly began using TPUs internally in 2015 and used them to train and serve AI-powered products like YouTube, Gmail, Google Maps, Google Play, and Android. In 2018, it made them available for third-party use as part of its AI infrastructure in Google Cloud Platform. Gemini, Google's most capable and general FM was trained on and is served using TPUs.[6]

- **Data**. FMs are trained on large datasets of existing content. This content can be text, images, music, or any other type of data that the model is designed to generate. The model learns the patterns in the training data and then uses that knowledge to generate new content that is similar in style and quality.

  Large language models ("**LLM**s") are trained on a lot of data; OpenAI's GPT-3, for example, was trained on 570 gigabytes of text data. Google's Gemini model was also reportedly trained on double the number of tokens as OpenAI's GPT-4.[7] Data is available from web scraping and crawling, public data sets – such as The Pile and Common Crawl (for text), ImageNet and Open Images (images), and LibriSpeech (for audio) – as well as books, news articles, scientific journals, and user generated content on social media sites, blogs, and forums, subject to appropriate privacy and usage considerations, and crowdsourcing. Model developers may also rely on proprietary data sets for model training. Some of the largest proprietary data sets include Google Search Index (estimated at between 500 and 600 billion pages) and YouTube (estimated at 14 billion publicly visible videos).

  Once the FM is trained, subsequent users may choose to "fine-tune" the model. Fine-tuning means to adapt an existing model to a specific task or domain by training it on a smaller and more relevant or specialized dataset. Fine-tuning allows users to customize the model to their specific needs and improve its performance on the targeted task. For example, by fine-tuning models on data regarding specific diseases, researchers and doctors can improve diagnostic accuracy and identify abnormalities more quickly.

  FMs may also be trained on synthetic data, as long as the synthetic data is realistic and representative of the real data. Synthetic data is data that is artificially created by a computer, rather than collected from the real world. Synthetic data can be used to augment existing data, fill in missing data, or create entirely new data sets for training FMs.

- **AI infrastructure**. The training process for FMs requires AI infrastructure, *i.e.* computational power delivered by GPU-powered servers. The number of servers

---

6 *See* Google Cloud, "Enabling next-generation AI workloads: Announcing TPU v5p and AI Hypercomputer", December 6, 2023, *available here.*

7 *See* Wire 19, "Will Google Gemini outdo GPT-4?", December 12, 2023, *available here*.

needed depends on the amount of data. Smaller models require less computational power than larger models.

FM developers can procure AI infrastructure from several sources. Companies may self-supply the necessary computing infrastructure for training their generative AI models; Google (Gemini), Amazon (Titan), Meta (Llama 2 model) and Aleph Alpha (Luminous) are examples of companies that have chosen to do so.[8] Companies can also procure necessary AI infrastructure from cloud-computing providers. There are at least ten global at-scale cloud-computing providers, including Amazon Web Services, Google Cloud Platform, IBM Cloud, Oracle Cloud, Microsoft Azure, Huawei, Bytedance, Baidu, Tencent, and AliCloud. Cohere and Anthropic, two AI start-ups, use Google Cloud Platform to train their large language models.[9] Specialized cloud providers, such as NVIDIA and Coreweave, have emerged, offering GPU-accelerated computing resources. Mistral trained its Mistral 7B on a Coreweave cluster.[10] In addition, governments around the world are investing in public computing infrastructure for training. CINECA – based in Italy – offers Leonardo, the sixth most powerful high-performance computing cluster in the world. It is open and available free of charge for industrial and scientific computing.[11] Finland's LUMI and Spain's MareNostrum, which has just opened in December 2023, are additional examples. The European Union recently announced a multi-billion-euro strategic investment to upgrade this existing network of high-performance supercomputers to support generative AI start-ups.[12]

1.2 FMs are incorporated and used by developers of AI-powered applications and services. Those applications are diverse and span a wide range of use-cases. Consumer-oriented chatbots and AI-powered search engines, like ChatGPT, Google Gemini, Perplexity AI, and Character.ai, have captured the most attention. But FMs are also being incorporated to deliver features and capabilities in existing applications, by businesses and researchers to analyze financial, medical, manufacturing and other data sets and derive insights, or to create AI-powered applications for internal business or organizational uses.

1.3 In building AI-powered applications, developers may draw on advanced general-purpose models, such as Claude 3 from Anthropic, Gemini from Google, and GPT-4 from OpenAI. But how they are used will vary based on the specific needs of the application. In some cases, a single FM might be used standalone to deliver a user experience with the choice dictated by the capabilities required and the price point. The most advanced FMs are expensive to run, and smaller FMs may provide the functionality needed for the use case at a lower price point. In other cases, FMs are

---

8 *See* MSN, "Mark Zuckerberg Says Meta Will Own Billions Worth of Nvidia H100 GPUs by Year End," January 18, 2024, available here. (Reporting on Mark Zuckerberg, Meta CEO's statement that Meta will have 350,000 Nvidia H1 GPUs and almost 600,000 H100 compute equivalent GPUs by end of 2024 and quoting him, "We're building massive compute infrastructure to support our future roadmap" for artificial intelligence).

9 *See* Multilingual, "Cohere and Google Cloud Announce Multi-Year Technology Partnership", November 18, 2021, ("Many of Cohere's products will be developed and deployed on Cloud TPUs, Google's supercomputers that are optimized for large-scale ML. . . ."), *available here*.

10 *See* Mistral AI, "Mistral 7B", September 27, 2023, ("We are grateful to CoreWeave for their 24/7 help in marshalling our cluster), *available here*.

11 *See* Leonard Cineca, "Leonardo Pre-Exascale Supercomputer", accessed March 8, 2024, *available here*. *See also* Barsotti, Marco Hugo, "Cosa fa Leonardo: in Italia uno dei supercomputer piu potenti al mondo", February 4, 2024, *available here*.

12 *See* "Commission launches AI innovation package to support Artificial Intelligence startups and SMEs", January 24, 2024, *available here*.

combined with other models to deliver the necessary performance at the lowest cost. For instance, Salesforce reportedly routes queries to its AI assistant "Einstein CoPilot" to different FMs, including its own, and those from OpenAI, Google, Anthropic, and Cohere, depending on the use case requirement.[13] Similarly, Microsoft uses multiple models to power its own AI assistant, Microsoft Copilot.

1.4 Many Open-Source FMs are made available and accessible for use by developers via hosting and collaboration platforms.[14] Hugging Face.com is one of the most popular of such platforms. It hosts more than 350,000 Open-Source models and facilitates collaboration on the research and development of those models.[15] Developers can download FMs from the platform and then deploy and use them on their own AI infrastructure or AI infrastructure offered by a third party, such as a cloud-computing provider.

1.5 Many cloud-computing providers offer or plan to offer a range of FMs for use by third-party applications developers. Microsoft Azure, for example, makes available a set of public "Cognitive Services" APIs, which draw on a combination of proprietary AI models to deliver specific functionality, *e.g.*, search, vision, speech. It also makes available more than 1,600 Open-Source and proprietary FMs for use by developers directly via a public API including models from OpenAI, Mistral, Meta, and Deci. The model providers set the price paid for the use of their model by developers. Amazon and Google each have similar services. IBM's watsonx.ai also offers a range of third-party FMs, including Mistral's Open-Source Mixtral-8x7B model.[16] Similarly, Oracle's OCI Generative AI service supports Meta's Llama 2 model and Cohere's models.[17] And other cloud-computing providers – such as OVHCloud – are planning to offer third-party FMs on their AI infrastructure.

1.6 Smaller, faster, and more cost-efficient FMs may be deployed and used on a personal computer or mobile computing device. Examples of these include Gemini Nano and mobile-optimized Llama 2.

1.7 Tools and Frameworks and Safety and Security Systems are also important. To deploy and use FM applications, developers typically rely on specialized tooling as well as first- or third-party safety and security systems that ensure the model is used responsibly. For example, orchestration tools help application developers to automate, coordinate, and manage workflows and pipeline infrastructure, track and monitor models for further analysis, and make the machine learning process more efficient. Examples of such tools include, Apache Airflow, Prefect, and Databricks Workflows.

1.8 It is important that FMs are used responsibly and are compliant with governing laws and regulations. First- and third-party safety and security systems serve as a wrapper that can enable responsible and compliant use. Microsoft Azure AI Content Safety is an example of a service that identifies harmful online content, categorizes it, and assigns

---

13 *See* Sean Michael Kerner, "What makes Einstein Copilot a genius? Salesforce says it's all about the data," VentureBeat, February 24, 2024, *available here*.

14 "**Open-Source**" refers to a software development approach where all or part of the software's source code is released openly to the general public. It may be developed collectively by decentralized voluntary communities, more formal organizations, or both.

15 Hugging Face Hub documentation, available here.

16 *See* IBM, "IBM Announces Availability of Open-Source Mistral AI Model on watsonx, Expands Model Choice to Help Enterprises Scale AI with Trust and Flexibility", February 29, 2024, *available here*.

17 *See* Oracle, "Announcing the general availability of OCI Generative AI", March 5, 2024, *available here*.

severity scores to support human content moderator workloads. It is designed to make applications and services safer from harmful user-generated and AI-generated content. Microsoft uses the service internally and makes it available for third parties to use with their own FMs.

**2. What are the main barriers to entry and expansion for the provision, distribution or integration of generative AI systems and/or components, including AI models? Please indicate to which components they relate.**

2.1 In addition to the above description of the AI technology stack in response to Question 1, the following is important to note:

- **Data is required to train and develop FMs but there are significant publicly available and Open-Source options.** As detailed in the response to Questions 1 and 7, data is used at the training stage to build the FM's knowledge. It is also used for fine-tuning, where the FM's accuracy is improved through dedicated training. The need for data is most pronounced at the training stage.

- **Significant compute capacity is required to train models, but this has not prevented multiple entrants from developing and deploying FMs.** Training for FMs requires a significant number of high-end accelerator chips (conventionally GPUs). Some FM developers use their own proprietary GPUs or TPUs (*e.g.*, NVIDIA, Google), while others are developing them (*e.g.*, AWS Trainium and Trainium2, Meta's training and inference accelerator, IBM Telum and Microsoft Maia), or have supported other silicon suppliers to enter the market (*e.g.*, Microsoft, Meta, Databricks, Essential AI and Lamini, among others, have confirmed they will deploy new AMD GPUs for AI workloads). As an alternative, FM developers also have the option to turn to cloud computing providers such as Google Cloud Platform (GCP), Amazon Web Services (AWS), Microsoft Azure, IBM Cloud, Oracle Cloud Infrastructure, and AliCloud; new entrant cloud providers, such as NVIDIA and CoreWeave, as well as smaller cloud providers, including Aligned, Arkon Energy, Cirrascale, Crusoe, Denvr Dataworks, and TensorWaves. In addition, there are other alternatives, such as publicly owned supercomputers. Access to compute capacity has not prevented significant number of recent entrants.

- Competition among cloud-computing providers to attract FM developers is intense. There are reports that Google has provided significant incentives to AI start-ups, such as Cohere and Anthropic, to use Google Cloud Platform infrastructure – reportedly going as far as to prioritize AI workloads from these start-ups over Google's own research teams. Outsourcing compute capacity has significant advantages for start-ups: less up-front cost, the ability to scale up and down, regional availability, and less distraction from building a datacenter. Moreover, there are a number of potential strategies to allow providers to optimize their compute requirements. For example, some providers deploy smaller models, which can provide answers to less demanding and/or narrower, more specialized prompts and reduce the need for compute power.

- **FM development requires significant technical expertise but start-ups are attractive for technical experts and innovation is not restricted to existing developers.** The significant number of start-ups developing FMs suggests that

innovation will not necessarily come from incumbent suppliers or larger firms. Start-ups can be particularly attractive for technical experts thanks to their nimbleness and ability to offer equity shares as part of their remuneration packages.

- **FM developers have shown an ability to raise funding from venture capitalists, as well as other types of investors.** The CMA Foundation Models Initial Report found that there "*has been significant investment into organizations that develop [foundation models] from a range of businesses including venture capital*", and further that "*currently small players are able to secure funding from investors"* leading to a *"notable increase in [foundation models] in recent years*."[18] NVIDIA, in particular, is reported to have invested in "more than two dozen" companies in 2023, "*from big new AI platforms valued in the billions of dollars to smaller start-ups applying AI to industries such as healthcare or energy*."[19]

**3. What are the main drivers of competition (*i.e.*, the elements that make a company a successful player) for the provision, distribution or integration of generative AI systems and/or components, including AI models?**

3.1 See responses to Question 1 and Question 2.

**4. Which competition issues will likely emerge for the provision, distribution or integration of generative AI systems and/or components, including AI models? Please indicate to which components they relate.**

4.1 Competition in the AI economy today is fierce, and the pace of innovation is rapid. Over the past decade, the most popular digital services – which have and continue to give rise to significant competition concerns – have been two-sided B2C platforms, characterized by strong direct and indirect network effects. In contrast, almost all of the layers of the AI technology stack are one-sided, positively priced inputs sold to other business customers. Consumers enter only at the top of the stack as customers for *some* of the thousands of AI-powered applications to be developed. It is not clear at all whether there will be a long-term, winner-take-all dynamic in generative AI.

4.2 Indeed, over the past year, FM developers have leapfrogged one another, touting their claim to leadership with each new model release, *e.g.*,:

- **Google**: "*Gemini 1.0 Ultra surpasses state-of-the-art performance on a range of benchmarks including text and coding.*" (December 2023)[20]

- **Mistral**: "*Mistral Large achieves strong results on commonly used benchmarks, making it the world's second-ranked model generally available through an API (next to GPT-4).*" (February 2024)[21]

---

18 *See* CMA, "AI Foundation Models: Full report", September 18, 2023, *available here*, para. 3.45.

19 *See* Tim Bradshaw and Ivan Levingston, "Nvidia emerges as leading investor in AI companies," Financial Times, December 11, 2023, *available here*.

20 *See* Google DeepMind, "Gemini", accessed March 8, 2024, *available here*

21 *See* Mistral, "Au Large", February 26, 2024, *available* here.

- **OpenAI**: "*GPT-4 considerably outperforms existing large language models, alongside most state-of-the-art (SOTA) models which may include benchmark-specific crafting or additional training protocols*." (March 2023)

- **Anthropic**: "*[Claude 3] Opus, our most intelligent model, outperforms its peers on most of the common evaluation benchmarks for AI systems[ . . .]*." (March 2024)[22]

4.3 The ability of FMs to remain competitive and for new firms to enter will depend in part on whether bottlenecks in data emerge in the future. So far, FM developers – whether following a closed or Open-Source approach – have been able to rely on a significant corpus of freely available public data to train and build their models. But the broad availability of such data may be threatened by two trends. First, intellectual property laws and their application will determine what content is available and at what price. In the United States, courts are currently considering whether the use of publicly available copyrighted content may be fairly used without permission or compensation for the training of FMs. If the answer is no, then the terms on which such content is made available may undermine innovation and competition. Will content creators make their content available broadly on reasonable and non-discriminatory terms? Or will they agree to exclusivity provisions or high prices that put their content out of reach for most FM developers?

4.4 Second, as the advancement and competitiveness of models requires more and different types of data, proprietary data sets may become increasingly important. For example, video content is necessary to expand the capabilities of AI models beyond text to understand and generate images and video. YouTube provides an unparalleled set of video content; it hosts an estimated 14 billion videos. Google has access to such content; but other AI developers do not.

4.5 Consumer oriented AI-powered personal assistants may face different, but unfortunately familiar, competition issues. These assistants are putting competitive pressure on traditional search as the gateway to the Internet and information and services online. To reach consumers where they are, such assistants will need to be present on their mobile phones. But Google and Apple own and control those platforms via Android and iOS and they already have voice assistants present on them: Google Assistant and Siri. They are well positioned to evolve and leverage their respective existing voice assistants into leadership positions in generative AI. Google is replacing Google Assistant with Gemini, its leading generative AI model. And Apple has announced that it will "deeply implement" generative AI, based on its Ajax model, into Siri and other Apple apps.[23] New entrants and competitors of Google and Apple will not enjoy the same advantages.

## 5. How will generative AI systems and/or components, including AI models likely be monetized, and which components will likely capture most of this monetization?

5.1 There is potential for monetization across the various layers of the AI technology stack described in response to Question 1 above.

---

22 *See* Anthropic, "Introducing the next generation of Claude", March 4, 2024, *available here.*

23 *See* Mark Gurman, "Inside Apple's Big Plan to Bring Generative AI to All Its Devices,", Bloomberg, October 22, 2023, *available here.*

5.2 The AI space is nascent, with innovation happening at a rapid pace at every "layer" of the stack. New FMs are being released every week. In the last couple of months alone: Google announced the Gemma open models family, its Gemini chat-bot, as well as its USD 2 billion investment in Anthropic; Adept AI released Fuyu-Heavy; Mistral launched Mixtral 8x7B; NVIDIA launched a chat-bot, OpenAI announced Sora; and the list of Open-Source models available keeps growing. Most of the promising start-ups, such as Mistral, Aleph Alpha, and Adept, did not even exist two years ago.

5.3 In this context, it is uncertain which monetization models will be adopted across the AI stack or even which layer of the stack is likely to capture most of this monetization. Some layers of the AI stack are established spaces, such as the AI infrastructure layer, whereas others are just emerging, such as FM applications.

5.4 Hardware and FM application layers of the AI stack clearly illustrate the differences in the potential for monetization and the available methods across the AI stack:

- Companies active at the hardware layer of the AI stack, such as GPU suppliers, have already developed a profitable monetization model and are currently capturing most of the monetization in the AI stack. The main supplier of GPUs for AI purposes is NVIDIA, who is responsible for over 70% and potentially as high as 90% of the demand for AI-focused GPUs.[24] NVIDIA has emerged as a major beneficiary of the development of AI generative models. In 2023, NVIDIA's AI and data center business generated more than USD 45 billion accounting for 79% of NVIDIA's total revenues.[25] Industry sources report that NVIDIA makes "*nearly 1,000%" (about 823%) in profit percentage for each H100 GPU accelerator it sells*".[26] As a result, NVIDIA's stock price has more than quadrupled over the past three years, making NVIDIA one of the most valuable companies in the world. Given that hardware is indispensable to build FMs, the monetization opportunity at the hardware layer of the AI stack is expected to continue, as demand is likely to surge and supply unlikely to be able to catch up.

- Unlike in the hardware layer, there is no settled monetization model for the newest FM-based chat-bots and AI assistant applications. As these applications are nascent, different companies are still exploring different monetization methods. Some players are offering both free and subscription-based versions of their chat-bots and AI assistants. For example, free versions of ChatGPT, Gemini, and Perplexity aim to meet everyday consumer needs with unlimited interactions. At the same time, the subscription versions of these chat-bots are powered by the most recent and powerful versions of FMs and offer additional tools. Others, such as Inflection, are still looking for a suitable monetization strategy and are currently offering only a free version of their chat-bots. Generally, advertising is currently not a widespread monetization model for such applications, and it is unclear whether the industry will deploy it in the future. The current sentiment is that these applications are unlikely to use advertising as a monetization model, given it would disrupt the natural flow of conversation that AI chat-bots and AI assistants try to

[24] *See* Wccftech, "Nvidia's global AI chip market *share* reaches a whopping 90%; analysts say competitors years from catching up", January 24, 2024, *available here.*

[25] *See* The Motley Fool, "Nvidia Is Making a Lot of Money Selling Artificial Intelligence (AI) Chips, but You May Be Surprised About How Much It Could Make From This Traditional Market", March 2, 2024, *available here.*

[26] *See* Tom's Hardware, "Nvidia Makes Nearly 1,000% Profit on H100 GPUs: Report", August 17, 2023, *available here*.

offer. This is particularly so as users expect seamless, non-intrusive interactions with the chat-bot or AI assistant of their choice.

5.5 As the AI field continues to evolve, different business models and revenue streams may emerge, influenced by market demand, customer perception, and technological innovation.

**6. Do open-source generative AI systems and/or components, including AI models compete effectively with proprietary AI generative systems and/or components? Please elaborate on your answer.**

## I. Open Source FMs compete effectively with proprietary FMs

6.1 Open-Source FMs and systems compete effectively with proprietary versions. Indeed, Open-Source FMs are often similar to proprietary FMs in terms of performance. In addition, because so many people contribute to Open-Source models, they can be developed faster than closed models. For example, Mistral announced that its Open-Source model Mixtral 8x7B outperforms GPT-3.5 on most standard benchmarks.[27] Mistral opined "*as with the Web, with web browsers (Webkit), with operating systems (Linux), with cloud orchestration (Kubernetes), open solutions will quickly outperform proprietary solutions for most use cases. They will be driven by the power of community and the requirement for technical excellence that successful open-source projects have always promoted.*"[28]

6.2 Several approaches to developing and releasing models may be considered Open-Source, including openly releasing model weights, datasets, and/or algorithms, and using different types of Open-Source software licenses. At present, numerous Open-Source FMs have been released by both decentralized communities and private sector companies. We expect this trend to continue.

6.3 Some well-known Open-Source FMs, that developers are using to develop their own tailored downstream FM applications (including chat-bots and AI assistants), include:

- **Gemma Series.** Google's Gemma series includes the Gemma 2B and Gemma 7B models, built using the same research and technology as the closed Gemini models.[29] Gemma models offer "state-of-the-art performance at size" and are capable of running on a variety of devices, including laptops.[30]

- **Llama 2.** In February 2023, Meta launched Llama and shortly followed up with the Open-Source Llama 2 on July 18, 2023.[31]

- **BLOOM.** A multilingual large language model with 176 billion parameters, capable of generating text in 46 natural languages and 13 programming languages. BLOOM was developed by a global community of AI researchers (the BigScience initiative) and is available publicly for download, study, and use.

---

27 *See* Mistral AI, "Mixtral of experts", December 11, 2023, *available here.*

28 *See* Mistral AI, "Bringing open AI models to the frontier", September 27, 2023, *available here*.

29 *See* Verge, "Google Gemma: because Google doesn't want to give away Gemini yet", February 21, 2024, *available here*.

30 *See* Google, *"*Gemma: Introducing new state-of-the-art open models*",* February 21, 2024, *available here*.

31 *See* Meta, "Meta and Microsoft Introduce the Next Generation of Llama", July 18, 2023, *available here*.

- **Mistral**, a French-based company, released Mistral 7B in September 2023, a small language model available under the Apache 2.0 license. Mistral 7B is a refinement of previously available small language models like Llama 2, but is claimed to outperform the competition on all standard English and code benchmarks. In December 2023, it released the Mixtral 8x7B large language model.[32]

- **OpenLLaMA**. An Open-Source reproduction of Meta's Llama model, developed by Berkeley AI Research, this project provides permissively licensed models with 3 billion, 7 billion, and 13 billion parameters, and is trained on one trillion tokens.

- **Stable Diffusion**. Founded in 2020, Stability AI has leveraged a cluster of 4,000 NVIDIA GPUs hosted on AWS to train a number of different FMs including Stable Beluga, Stable LM, and Stable Diffusion XL.

- **Falcon Series**. Developed by Abu Dhabi's Technology Innovation Institute (TII), Falcon-Series consists of Falcon-40B, Falcon-7B, and Falcon 180B (which reportedly is ranked #1 on Hugging Face's leaderboard for pre-trained Open LLMs and outperforms Llama 2).[33] The series has a unique training data pipeline that extracts content with deduplication and filtering from web data.

- **MPT Series**. A set of decoder-only large language models, MPT-Series models have been trained by MosaicML on one trillion tokens spanning code, natural language text, and scientific text. These models come in two specific versions: MPT-Instruct and MPT-Chat.

6.4 A diverse set of platforms, targeting researchers and technical users along with commercial organizations, have emerged to host and enable collaboration on Open-Source AI models, datasets and applications. For example:

- Microsoft's **GitHub** enables developers to create, store, manage and store their code openly, hosting more than 8,000 "open-source generative AI projects" which range from commercially backed large language models like Meta's Llama to experimental Open-Source applications.[34]

- **Hugging Face**, a French American company, is the leading platform and currently hosts over 350,000 free and accessible Open-Source models.[35] More than 50,000 organizations are using Hugging Face.

## II. Partnerships and investments are crucial for the continued success of Open-Source FMs

6.5 Creating Open-Source generative AI models is not cost-free. Unlike the brainpower and effort needed to support other Open-Source software development, the training data and infrastructure costs for training the generative AI models cannot be crowdsourced easily.

---

32 *See* Mistral AI, "Mixtral of experts", December 11, 2023, *available here*.
33 *See* TII, "Falcon Models", accessed March 6, 2024, *available here*.
34 *See* GitHub Blog, "A developer's guide to open source LLMs and generative AI", October 5, 2023, *available here*.
35 *See* Hugging Face, "Hugging Face Hub documentation", accessed March 6, 2024, *available here*.

- To secure the necessary data, Open-Source AI companies and projects have leveraged vast volumes of public content on the web and used open data sets – such as The Pile[36] – licensed for training purposes. However, as mentioned in response to Question 7 below, access to proprietary data can be a competitive advantage.

- To pay for the infrastructure cost for the model training, Open-Source developers have raised funding from venture capital funders and other technology companies. In particular:

  - **BLOOM.** BigScience's LLM was trained on the Jean Zay supercomputer in France with a compute grant valued at EUR 3 million.[37]

  - **Mistral.** During its latest December 2023 funding round, Mistral achieved a valuation of approximately USD 2 billion, raising EUR 385 million from various investors (including NVIDIA and Salesforce) with Andreessen Horowitz (a16z) leading the Series A investment.[38] Mistral AI previously raised a USD 113 million seed round shortly after its inception. Microsoft made a small investment of EUR 15 million in Mistral alongside these investors.[39]

  - **Stable Diffusion.** In addition to the USD 101 million raised in a Series A funding in October 2022, Stability AI raised new funding of USD 50 million in the form of a convertible note in November 2023.[40]

6.6 Open-Source FMs currently compete effectively with proprietary FMs. To stay competitive, it is essential they continue to have the flexibility to obtain funding and form partnerships to get access to the necessary inputs.

**7. What is the role of data and what are its relevant characteristics for the provision of generative AI systems and/or components, including AI models?**

7.1 FM developers need data at both stages: (i) training, where data is used to build the FM's knowledge; and (ii) fine-tuning, where the FM's accuracy is improved through dedicated training. Data can be either proprietary (*e.g.*, video hosting platforms such as YouTube, image repositories, and content websites) or Open-Source (including other FMs).

- **Training**. The availability of data is the most important at the training stage where the model learns the patterns in the training data and then uses that knowledge to generate new content. Such data, however, does not need to be proprietary – training can be successfully carried out with publicly available datasets.[41] For example, Llama (Meta), GPT-3 (OpenAI) and Stable Diffusion (Stability AI) have been trained entirely on Open-Source data.

---

36 *See* Pile, "What is the Pile?", accessed March 6, 2024, *available here.*

37 *See* BigScience Blog, "Introducing The World's Largest Open Multilingual Language Model: BLOOM", accessed March 6, 2024, *available here*.

38 *See* CrunchBase, "OpenAI Competitor Mistral AI Nabs $415M Round, Hitting $2B Valuation", December 11, 2023, *available here*.

39 *See* TechCrunch, "Microsoft made a $16M investment in Mistral AI", February 27, 2024, *available here.*

40 *See* Verdict, "UK's Stability AI gets financial boost from Intel", November 11, 2023, *available here.*

41 These datasets include C4, The Pile, Project Gutenberg Corpus, LAION-400M, Red Pajama, RefinedWeb, and Starcoder.

- **Fine-tuning**. In the fine-tuning phase, where the model's accuracy is improved through dedicated training, data is often human-generated either in-house or sourced from specialist third-party data providers such as Scale AI, Prolific, Surge AI, SuperAnnotate, Dataloop, and many others. Open source alternatives can also be used to fine-tune generative AI models.

7.1 Publicly available and Open-Source options for data include:

- **LAION**: FM developers use the open LAION-5B dataset for building large pre-trained vision language models. The dataset contains 5.85 billion image-text pairs generated from CLIP, with descriptions in English and foreign languages, thereby catering to a multilingual domain. Other LAION datasets include LAION-COCO, the world's largest dataset of 600 million generated high-quality captions for publicly available web-images and LAION-400M which contains 400 million image-text pairs.

- **COYO-700M**: COYO-700M is a large-scale dataset that contains 747M image-text pairs as well as many other meta-attributes to increase the usability to train various models. COYO-700M is an open dataset available for download on Hugging Face.

- **PMD**: Public Multimodal Dataset (PMD) is a collection of publicly-available image-text pair datasets, which is available on Hugging Face. PMD contains 70 million image-text pairs in total with 68 million unique images. The dataset contains pairs from a number of other datasets including Conceptual Captions, Conceptual Captions 12M, WIT, Localized Narratives, RedCaps, COCO, SBU Captions, Visual Genome, and a subset of YFCC100M dataset.

- **ImageNet**: ImageNet contains over 14 million images with annotations categorized according to the WordNet hierarchy. Since 2010, the dataset is used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a benchmark in image classification and object detection. The publicly released dataset contains a set of manually annotated training images.

- **Open Images**: Open Images is a dataset of over 9 million images annotated with labels spanning thousands of object categories.

- **Common Crawl**: Common Crawl is a non-profit organization, which provides web crawl data for free. The corpus contains petabytes of data collected over 12 years of web crawling, including over 50 billion web pages. The corpus contains raw web page data, metadata extracts and text extracts.

- **Dolma**: Dolma consists of 3 trillion tokens from a diverse mix of web content, academic publications, code, books, and encyclopaedic materials.

- **C4**: An English-language dataset (approximately 750 GB) prepared by the Allen Institute for AI through cleaning the Common Crawl dataset that has been built through 12 years of web scraping.

- **The Pile**: The Pile is an 825 GB dataset constructed by EleutherAI from 22 diverse high-quality subsets (both existing and newly constructed) many of which derive from academic or professional sources.

- **Hugging Face**: Hugging Face lists over 700 open computer vision datasets for image classification alone and over 400 datasets for image detection.

7.2 In addition to Open-Source options, FMs may also be trained and fine-tuned on synthetic data. Synthetic data is data that is artificially created by a computer, often by FMs themselves, rather than collected from the real world. Synthetic data can be used to augment existing data, fill in missing data, or create entirely new data sets for training of AI models.

7.3 Access to data can influence competitive dynamics in the AI space and is important for the development of FMs. As high quality, diverse data is necessary for training FMs, companies that have unfettered access to such data can enjoy a competitive advantage.

7.4 In light of the above, it is important that the Commission monitors the use and accessibility of data for FM development. Maintaining current levels of intense competition and innovation in AI hinges on ensuring that all FM developers have an opportunity to access quality data, and no single company has too great of a data advantage for FM development.

**8. What is the role of interoperability in the provision of generative AI systems and/or components, including AI models? Is the lack of interoperability between components a risk to effective competition?**

8.1 FMs are interoperable. Developers can combine FMs, whether Open-Source or proprietary, to deliver an optimal user experience:

- Models-as-a-Service ("**MaaS**") solutions currently offered by several major cloud-computing providers, such as Microsoft Azure AI Studio, AWS Amazon Bedrock, and Google Vertex AI, make it easy for developers to use multiple FMs together by offering access to thousands of models via a public API.

- Orchestration tools, such as LangChain, Dust, GPT Index, Fixie.ai, and Cognosis, can help developers automate and manage the use of multiple FMs in a single workflow.

**9. Do the vertically integrated companies, which provide several components along the value chain of generative AI systems (including user facing applications and plug-ins), enjoy an advantage compared to other companies? Please elaborate on your answer.**

**10. What is the rationale of the investments and/or acquisitions of large companies in small providers of generative AI systems and/or components, including AI models? How will they affect competition?**

10.1 The responses to Questions 9 and 10 are provided together in the following paragraphs.

**I. <u>The generative AI space is characterized by innovation and entry at a pace hardly seen before</u>**

10.2 The AI space is greatly competitive and vibrant. Most of the developments have occurred since Microsoft's investments in Open AI. OpenAI's partnership-fostered

innovation in FM and FM applications has triggered a strong competitive response by a wide range of competitors. The industry is now seeing numerous players partnering up to compete and innovate at every layer of the AI stack.

10.3 For instance, competition in the development and supply of FMs is evolving rapidly with innovation and market entry from a range of sources, as well as a growing number of increasingly powerful Open-Source options. Most prominently, FM developers include numerous start-ups, such as Anthropic, Stability AI, Cohere, Mistral, Adept, Character.ai, Cohere, and others, alongside traditional tech players. Importantly, all of these start-ups relied on different forms of investments and partnerships that enabled them to enter and expand in the space. Salesforce, Oracle, IBM, NVIDIA, Intel, Meta, Apple, Amazon, and Alphabet, to name just a few, each have venture capital funds that regularly invest in technology start-ups, including generative AI. Some examples of investments include:

- **Amazon Web Services** has invested USD 4 billion into Anthropic for a minority stake in the start-up.
- **Google** has also invested in Anthropic as well as Character.ai, an AI chat-bot start-up.
- **Salesforce** has invested in multiple AI start-ups including Cohere, Humane, Tribble, and Anthropic.[42]
- **NVIDIA** has invested into Cohere and Mistral, among others.
- **Microsoft** has invested in OpenAI, but also in a broad range of diverse generative AI start-ups, such as Adept and Mistral. In some instances, those investments have simply provided seed funding to finance day-to-day operations. In others, those investments have been more targeted providing the compute infrastructure necessary for the training and deployment of generative AI models and applications.

10.4 The competitive landscape today is entirely different from just a year ago and can be expected to change equally profoundly in the next 1-2 years. Only in the past three months, there have been the following major developments:[43]

- On December 11, 2023, Mistral released Mixtral 8x7B, a high-quality model with open weights.[44] Mistral claims that this model "*outperforms Llama 2 70B on most benchmarks with 6x faster inference*" and that it "*is the strongest open-weight model with a permissive license and the best model overall regarding cost/performance trade-offs."* Mistral was founded in April 2023.
- On December 19, 2023, Anthropic announced changes in their Commercial Terms of Service that enable their customers to retain ownership rights over any outputs they generate through their use of Anthropic's services and protect their customers

42 *See* "Salesforce Ventures doubles down on generative AI fund", June 12, 2023, *available here*.

43 *See* CMA, "Microsoft / OpenAI partnership merger inquiry", December 8, 2023, *available here*.

44 *See* Mistral, "Mixtral of Experts", December 11, 2023, *available here*.

from copyright infringement claims.[45] This announcement came less than a month after the release of Claude 2.1, Anthropic's latest AI assistant.[46]

- On January 24, 2024, Adept AI released Fuyu-Heavy, a multimodal model designed specifically for digital agents.[47] It claims to have on par performance with Google's Gemini Pro on standard text-only evaluations and to be outperforming it by other commonly used benchmarks.

- On January 29, 2024, Meta's CEO Mark Zuckerberg has announced a commitment to developing Artificial General Intelligence.[48]

- On January 30, 2024, Amazon announced the preview of Amazon Q data integration in AWS Glue.[49] Amazon Q is a chat experience powered by Amazon Bedrock and AWS Glue is a serverless data integration service.

- On February 8, 2024, Google launched Gemini Advanced, which Google claims "*in blind evaluations [...] Gemini Advanced with Ultra 1.0 is now the most preferred chatbot compared to leading alternatives*".[50] It also announced that a new mobile app for Gemini and Gemini Advanced was starting to be rolled out.

- On February 13, 2024, NVIDIA announced Chat with GTX, a free, personalized AI chatbox.[51]

- On February 15, 2024, Google announced Gemini 1.5, their next-generation FM.[52] This tool was released in preview form and is capable of handling 1 million tokens, which is a significant increase when compared to Gemini 1.0 Pro (32,000 tokens) or GPT-4 Turbo (128,000 tokens).

- On February 15, 2024, OpenAI announced Sora, its AI model that can create realistic videos from text instructions.[53] According to OpenAI, "*Sora can generate videos up to a minute long while maintaining visual quality and adherence to the user's prompt*".

- On February 21, 2024, Google announced Gemma, a family of lightweight, state-of-the-art open models built from the same research and technology used to create the Gemini models.[54]

---

45 *See* Anthropic, "Expanded legal protections and improvements to our API", December 19, 2023, *available here*.

46 *See* Anthropic, "Claude 2.1," November 21, 2023, *available here*.

47 *See* Adept, "Adept Fuyu-Heavy: A new multimodal model", 24 January 2024, *available here*.

48 *See* The Finance Story, " Meta's CEO Mark Zuckerberg to build AGI, a type of AI that can perform any intellectual task like humans", January 29, 2024, *available here*.

49 *See* Amazon AWS, "New chat experience for AWS Glue using natural language – Amazon Q data integration in AWS Glue (Preview)", January 30, 2024, *available here*.

50 *See* Google, "Bard becomes Gemini: Try Ultra 1.0 and a new mobile app today", February 8, 2024, *available here*.

51 Chat with RTX uses retrieval-augmented generation (RAG), NVIDIA TensorRT-LLM software and NVIDIA RTX acceleration to bring generative AI capabilities to local, GeForce-powered Windows PCs. Users can connect local files on a PC as a dataset to an open-source large language model like Mistral or Llama 2, enabling queries for quick, contextually relevant answers. *See here*.

52 *See* Google, "Our next-generation model: Gemini 1.5", February 15, 2024, *available here*.

53 *See* OpenAI, "Sora", accessed March 8, 2024, *available here*.

54 *See* Google, "Gemma: Introducing new state-of-the-art open models", February 21, 2024, *available here*

10.5 Against this backdrop, any developer that fails to aggressively innovate will quickly become obsolete. And with rare exceptions, no company to date has managed to do it alone.

## II. <u>While some vertical integration is common in the AI space, strong competition and innovation have come from different forms of partnerships in the industry</u>

10.6 Many companies active in the AI space participate in multiple levels of the value chain (AI technology stack). For instance:

- **Integrated FM development and FM applications players**. Many industry players, including AI start-ups, are active both at the FM development and FM applications layer of the AI stack, such as OpenAI (*e.g.*, GPT-4 and ChatGPT), Anthropic (*e.g.*, Claude 3 and Claude), Google (*e.g.*, Gemini FM and the Gemini AI assistant), and Character.ai (*e.g.*, C1.1 and Character.ai). These FM developers can integrate their FMs directly into their FM applications.

- **Integrated infrastructure.** Players with wide-ranging existing offerings can also integrate FMs across their infrastructure: Apple (*e.g.*, iOS data, Siri, Ajax and internal AI infrastructure), Meta (*e.g.*, training data, internal AI infrastructure, Llama Model, downstream applications in Facebook, Instagram, and WhatsApp), Amazon (*e.g.*, Titanium GPUs, proprietary data, AWS cloud services, Titan Model, Marketplace), and Alphabet (TPUs, proprietary training data, Google Cloud Platform, Gemini, Tensor Framework, downstream applications in Search, Maps, YouTube, *etc.*).

10.7 Despite this, most of the companies in the AI space also depend on third parties for input and/or distribution of its downstream products. For instance:

- **Perplexity.ai**, an AI start-up, trains and builds its own generative AI models to run its search application but it also depends on third parties: it incorporates leading generative AI models to power its application (in particular, OpenAI's GPT-3.5, GPT-4, as well as Anthropic's Claude 2),[55] it relies on computing infrastructure from Amazon's AWS and Microsoft Azure to deploy its application,[56] and it also has to go through mobile app stores to distribute its application on iOS and Android platforms.

- **Amazon, Microsoft, and Oracle** all offer cloud infrastructure and tooling optimized for AI workloads, are investing in developing FMs, and have products and services that integrate those and other FMs in consumer and/or developer-facing applications. Despite this, Microsoft was not able to develop a cutting-edge FM in-house to date. Amazon equally considered it important to invest in start-ups and foster partnerships, as described above. These companies are also dependent on third parties for multiple inputs, such as semiconductors, data for training, other FMs to provide to customers and power their own applications, and mobile

---

55 *See* Perplexity, "What model does Perplexity use and what is the Perplexity model?", *available <u>here</u>*.

56 *See* AWS, "Perplexity AI at AWS re:Invent 2023", accessed March 6, 2024, *available <u>here</u>* ; *See also* Microsoft for Start-ups, "Perplexity powers its 'answer engine' with Azure OpenAI Service", June 15, 2023, *available <u>here</u>*.

platforms and mobile operating systems for distribution of their FM applications to consumers.

10.8 While vertical integration described above is common in the AI space, certain value chain levels can be more important than others. For example:

- Semiconductors are one of the most important inputs for the development and supply of FMs. The availability of AI semiconductors is limited; the capacity constraints are already being felt in the industry, with one leading supplier, NVIDIA, supplying most of the AI space. The market is responding: Many companies are developing these chips in-house and NVIDIA is increasing the supply. Given this, and as mentioned in response to Question 1 above, a vertically integrated player able to self-supply AI semiconductors, such as Google with its TPUs, has a competitive advantage for the years to come. Indeed, Google's VP and General Manager of Compute and Machine Learning highlighted that the latest TPU v5p chip design for FM training (reportedly up to 4.8x faster than NVIDIA's A100 and on par or superior to the flagship H100[57]) was a result of the hardware-software co-design only available to a truly vertically integrated player.

- Similarly, one of the main ways in which FMs are supplied to end consumers is through FM applications (such as chat-bots or virtual assistants). An FM developer who controls/owns mobile operating systems such as Android or iOS, which are likely to become the most important distribution channel for FM applications, has a significant competitive advantage compared to other FM developers. For example, Apple is reportedly aiming to operate "*generative AI through mobile devices, which would allow AI chatbots and apps to run on the phone's own hardware and software rather than be powered by cloud services in data centers*".[58] Mobile app stores and mobile operating systems distribution channels can give these FM developers preferential access to a large number of consumers. For instance, Google has incorporated one of the most powerful FMs, Gemini, into Google Assistant, enabling users to opt easily into Gemini as their mobile assistant. Gemini enjoys privileged access to connect with popular Google apps like Gmail, Maps, and YouTube. In addition, all Android OEMs are required to install Google Assistant and set it as the default on Android mobile phones. These contractual provisions guarantee instant distribution to and access to Gemini by over 3 billion consumers worldwide.[59] Most recently, Google announced bringing AI to local computers and smartphones, with its Gemma2B model being made available on Android smartphones, giving Google's Gemma2B the ability to reach billions of users worldwide.[60]

10.9 Almost no company is vertically integrated up and down the AI technology stack to the extent that they can do it alone. That means almost all of the industry players need to rely on partnerships to compete effectively. To date, companies have been entering at different levels of the technology stack, relying on the investments by and partnerships

---

57 *See* Tech Radar, "Google is rapidly turning into a formidable opponent to BFF Nvidia — the TPU v5p AI chip powering its hypercomputer is faster and has more memory and bandwidth than ever before, beating even the mighty H100", December 23, 2023, *available here*.

58 *See* ArsTechnica, "*Apple aims to run AI models directly on iPhones, other devices*", January 24, 2024, *available here*.

59 *See* The Business of Apps, "Android Statistics (2024)", March 1, 2024, *available here*.

60 *See* Google, "Gemma: Introducing new state-of-the-art open models", February 21, 2024, *available here*.

with those upstream from them. This is what has fostered innovation in the AI space at a pace hardly seen before.

## III. Partnerships and investments in the AI space should be encouraged: They render innovation and consumer choice, preventing any player from becoming 'too big too soon'

10.10 There is currently a strong bias towards investments and partnerships as opposed to acquisitions in generative AI. In fact, there were only a handful of acquisitions of AI start-ups by large technology companies in 2023.[61] This trend reflects the state of the technology and the marketplace. In particular:

- **AI start-ups primarily benefit from investments and partnerships.** They allow AI start-ups to take advantage of the financial resources and, in some instances, the compute infrastructure that large technology companies have. For example, Cohere, a leading AI platform for enterprise, trains, builds, and deploys its generative AI models on Oracle Cloud Infrastructure. And, in turn, Oracle embeds Cohere's models across its portfolio of cloud services.[62] Similarly, the strategic partnership between Anthropic and AWS gives the AI start-up access to AWS' purpose-built AI chips and AWS customers. Amazon will build on top of Anthropic's models to "*incorporate generative AI capabilities into their work, enhance existing applications, and create net-new customer experiences across Amazon's business.*"[63]

- **Large technology companies gain strategic optionality by investing in and partnering with AI start-ups.** By investing in a diverse set of AI start-ups, large technology companies can accelerate innovation across several different dimensions more quickly and broadly than through their efforts alone. And if an investment succeeds, it may boost a company's existing technology offerings. And start-ups have proven to innovate more rapidly and aggressively in this space: FMs and FM applications developed by OpenAI, Anthropic, Mistral, Cohere, and others, are at least as good as the technology developed in-house by more established tech companies.

10.11 Partnerships and investments have played an important role in accelerating innovation and entry in the era of generative AI. The alternative is hardly attractive. If investment by large technology companies had been or is cut off or unduly restricted, those companies would have instead redirected funding to internal R&D efforts only. Such competitive dynamics in a nascent space render fertile ground for one company to become 'too big too soon,' which would not be a desirable outcome of the AI race. Encouraging pro-competitive partnerships in the AI space is an effective way to prevent companies from becoming vertically integrated in a manner that would result in an anticompetitive advantage.

---

61 Databricks acquired MosaicML in June 2023 and Thompson Reuters purchased Casetext in Augus 2023. Apple acquired WaveOne, a small developer of algorithms for compressing video, in March 2023 and Google acquired Photomath, an AI math vendor at the same time.

62 *See* Oracle, "Oracle to deliver powerful and secure generative AI services for business", June 13, 2023, *available here*.

63 *See* Anthropic, "Expanding access to safer AI with Amazon", September 25, 2023, *available here*.

**11. Do you expect the emergence of generative AI systems and/or components, including AI models to trigger the need to adapt EU legal antitrust concepts?** The current antitrust framework and legislation is sufficient. Among others, the Digital Markets Act, the EU Merger Regulation, and Regulation 1/2003 contain potent tools to address any potential concerns.

**12. Do you expect the emergence of generative AI systems to trigger the need to adapt EU antitrust investigation tools and practices?**

12.1 The Commission, along with other antitrust authorities, should continue to monitor the generative AI space. The continued build-out of DG COMP's Data Analysis and Technology unit led by the Chief Technology Officer can be useful in that respect.

12.2 Microsoft welcomes and supports the Commission in using its investigative tools to learn more about the emerging field of generative AI, be that through this call for contributions or an eventual sector inquiry.

***