cep Centrum für Europäische Politik
FREIBURG | BERLIN

cepInput

No. 6 | 2024 12 March 2024

Competition in Generative Artificial Intelligence

cepInput in Answer to the European Commission's Call for Contributions

Anselm Küsters and Matthias Kullas

© Figure generated by DALL-E 3 via ChatGPT with own prompt

Generative Artificial Intelligence (genAI) will reshape economies and the way we perceive and interact with reality. Ensuring competition in the genAI value chain for AI services is therefore not only an economic imperative but also a precautionary exercise to protect democratic values and ensure EU digital sovereignty. This cepInput assesses likely restrictions to competition in the three parts of the genAI value chain for AI services: infrastructure (large-scale data and computing power), training of foundation models, and B2B/B2C downstream services and applications.

- ► GenAI infrastructure: There is currently no indication of competition problems with regard to training data but exclusive licensing agreements with media firms should be closely monitored. Regarding computing power, a unique mix of high demand, too few suppliers, high switching costs, and growing vertical integration poses non-negligible risks to competition, such as in the form of self-preferencing or discrimination.
- ► GenAI training: Due to significant economies of scale and scope, such as high fixed costs for training, and "emergent capabilities" when scaling up (i.e. exponential learning effects), the market for foundation models is a "winner-takes-most" market. That makes it very hard for start-ups to compete with established providers of foundation models. However, future progress in AI architectures might reduce data requirements.
- ► GenAI deployment: With regard to down-stream genAI applications and services, start-ups must compete with large firms such as Microsoft and Google that have preferential access to advanced models and can leverage their market power and existing customers. Future hurdles might come from "app-store" models.
- ► In addition to understanding the dynamics of each part of the value chain for AI services, EU competition law enforcers must take an holistic view in order to understand the degree to which certain actors are already vertically integrated. A key example relates to Microsoft and its strategic partnerships with OpenAI and Mistral AI, which can be considered as one "undertaking" with strong market power over the whole genAI value chain for AI services.

Content

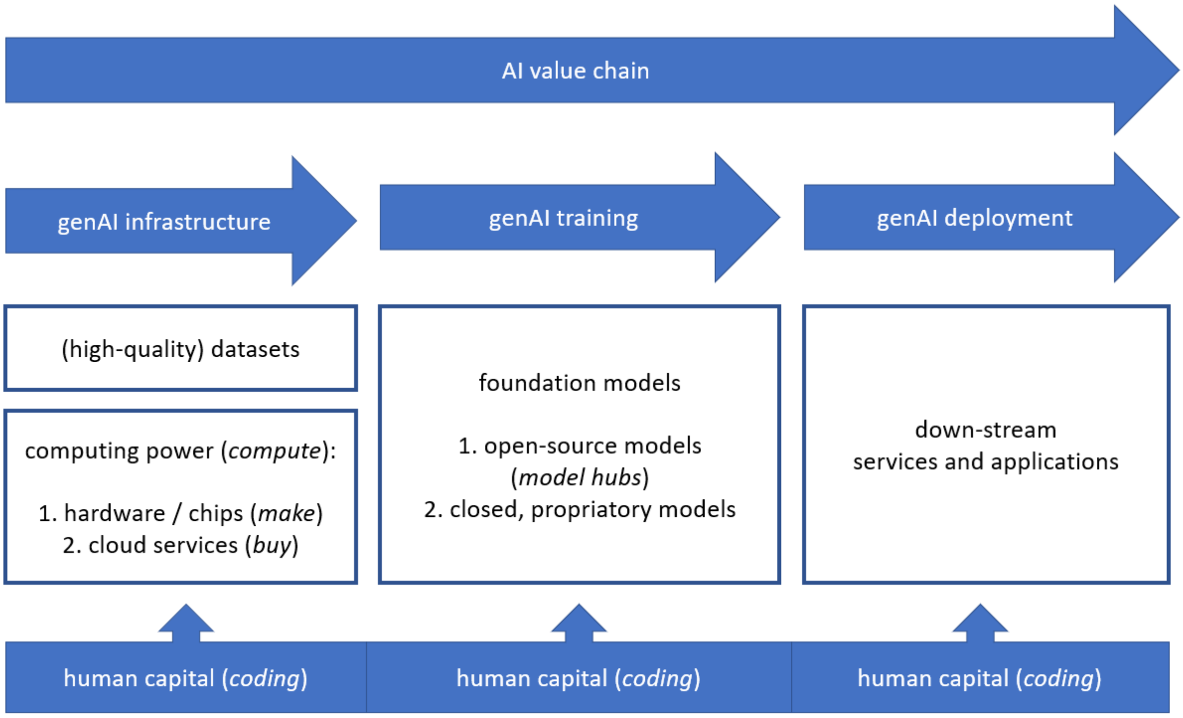
Figures

Tables

1 Introduction

Generative Artificial Intelligence (genAI) represents a transformative technology that is reshaping competition and innovation. It will impact the way we access information[1] and, increasingly, perceive and interact with reality as genAI agents are embedded in an ever-increasing array of products and services.[2] Therefore, assessing the state of competition in the value chain for genAI as a service (Figure 1) is not only an economic imperative in line with the competition law provisions of the European Union (EU), covering cartelisation, dominant positions, and mergers, but should also be seen as a crucial precautionary exercise to protect the Union's democratic values and ensure its digital sovereignty in the future.

Fig. 1: Value chain for generative AI as a service


AI value chain
genAI infrastructure
genAI training
genAI deployment
(high-quality) datasets
computing power (*compute*):
1. hardware / chips (*make*)
2. cloud services (*buy*)
foundation models
1. open-source models (*model hubs*)
2. closed, propriatory models
down-stream services and applications
human capital (*coding*)
human capital (*coding*)
human capital (*coding*)

Source: Own illustration.

Responding to the European Commission's call for contributions,[3] this cepAdhoc aims to provide a comprehensive analysis of the competitive dynamics within the value chain for generative AI as a service, also known as the AI "tech stack". We divide the value chain for generative AI as a service into three main parts: The first part relates to the genAI infrastructure, which consists of the market for large-scale (high-quality) datasets and the market for computing power. Both are essential inputs for the training and development of foundation models, which is the second part of the AI value chain. The third part is the market (or markets) for B2B or B2C downstream genAI services and applications that use a foundation model as input (Figure 1).[4] Each part of the value chain for genAI as a service

[1] For an example of how genAI might change Google traffic to newspaper websites, see: News Publishers See Google's AI Search Tool as a Traffic-Destroying Nightmare - WSJ. See also the excellent discussion at: Generative AI's end-run around copyright won't be resolved by the courts (aisnakeoil.com).

[2] This is the vision of, for instance, Suleyman (2023), The Coming Wave, London.

[3] The call for contributions can be found here: Competition in virtual worlds and generative AI (europa.eu).

[4] In the definition of the genAI value chain, we follow: Exploring opportunities in the gen AI value chain | McKinsey. In addition to the 6 dimensions highlighted by McKinsey, we add data and human capital.

requires computer specialists for coding, meaning that scarce human capital could become an additional bottleneck and thus a barrier to entry.

In particular, we screen all parts of this value chain for market concentration (oligopoly) and significant restrictions of competition, using the following three questions for guidance to see whether these markets tend to be subject to competition problems. In addition, we check whether vertical integration over several of parts of this value chain might lead to barriers to entry:

1. **Size matters: Are there economies of scale and scope, such as high fixed costs, or are there network effects that might lead to a tendency for monopolisation?**

2. **Switching costs: Are there lock-in effects that restrict competition?**

3. **Essential facilities: Is there a scarce factor of production such as human capital or patents, that restricts competition because only certain companies can access it?**

Given the rapid evolution of genAI technologies and their far-reaching impact, it is crucial for the Commission to anticipate and address potential competition concerns proactively. This cepInput provides a structured framework for analysing these concerns, offering recommendations to ensure a competitive and innovative environment for generative AI in Europe. After briefly outlining the need for agile competition law enforcement in the field of genAI as a complement to the EU's AI Act (section 2), the analysis goes on to assess the competitive dynamics in each part of the value chain for genAI as a service(section 3). Finally, we point to some more general concerns that are outside a narrow definition of competition law but which might also be relevant considering the Commission's objective to further strategic autonomy[5] in the digital realm (section 4). On this basis, we conclude (section 5).

2 Background: Competition law, digital markets, and AI

Over the last decade, digital markets have increasingly attracted the attention of competition law enforcers and legislators, particularly in the context of emerging technologies such as digital, two-sided platforms. This is largely due to the unique characteristics of digital markets, which often include rapid innovation, high scalability, lock-in-effects, and significant network effects. While digital technology helps to unlock innovation, the combination of these attributes poses challenges for competition law frameworks. Following its Communication "A competition policy fit for new challenges" of November 2021,[6] the Commission has therefore enacted several legislative initiatives in the last few years, including new rules for digital competition, especially through the Digital Markets Act (DMA), new guidelines for its traditional antitrust law toolbox (see, e.g., the revised Market Definition Notice which takes better account of digital markets[7]), and the Regulation on promoting fairness and transparency for business users of online intermediation services.[8] For the purpose of this analysis, emphasis will be placed on the Communication on fostering a European approach to AI,[9] a review of the Coordinated

[5] For a definition and discussion of this term, see: Armin Steinbach (2023), EU's Turn to 'Strategic Autonomy': Leeway for Policy Action and Points of Conflict | European Journal of International Law | Oxford Academic (oup.com).

[6] See: A competition policy fit for new challenges - European Commission (europa.eu).

[7] Commission (2024), Commission Notice on the definition of the relevant market for the purposes of Union competition law, Brussels, 8.2.2024, C(2023) 6789 final.

[8] For a discussion, see: Europe's Digital Sovereignty: How to Tame Digital Power (commongroundeurope.eu).

[9] Communication on Fostering a European approach to Artificial Intelligence | Shaping Europe's digital future (europa.eu).

Plan on AI,[10] the recent political agreement on the AI Act,[11] as well as the AI innovation package to support AI startups and SMEs[12]. Following the call for contributions on "Competition in Virtual Worlds and Generative AI", the Competition Commissioner stressed that since AI "will touch virtually every aspect of the economy", "we have to look carefully at vertical integration and at ecosystems", "take account of the impact of AI in how we assess mergers" and "think about how AI might lead to new kinds of algorithmic collusion".[13]

In the context of this general legislative overhaul and adaptation to the digital revolution, generative AI should not be seen as a paradigm shift in competition law enforcement as such, but as a further example of a new technology that poses novel problems, similar to previous digital technology shifts. We therefore welcome the fact that the Commission is engaging early on with this latest technological development, in order to prevent ex-ante the emergence of competition problems, such as ecosystems in other digital markets. The challenge for policymakers to keep pace with genAI became clear during the final negotiations of the EU AI Act, which necessitated the inclusion of new rules for foundation models. Since the technology's rapid development often outpaces regulatory responses, there is a need for more agile and informed competition law enforcement as a complement to the formulation of new rules.

However, when considering how to use the existing competition law toolkit in the context of genAI, several challenges emerge. To begin with, it is necessary to go beyond traditional market definition methods.[14] For genAI markets, one might rely on the revised market definition mentioned above, according to which "dependencies in relation to data", such as "the costs of data portability" and "access to data", can form significant barriers to substitution and switching costs. These barriers become even higher if new entrants and existing competitors have to make "specific capital investments or specific investments in production processes, learning and human capital".[15] In the context of genAI, it is crucial to consider factors like access to training data and computing power (which we analyse in the next section).[16] By focusing on barriers that reduce potential competition, instead of traditional, price-focused antitrust tools, enforcers can alleviate the problems with market definition by providing a more realistic picture of competitive constraints in digital markets.[17] For instance, when non-price parameters are particularly relevant for the assessment of substitution (as might be the case in the genAI value chain), the Commission can focus on "data portability and

[10] Coordinated Plan on Artificial Intelligence | Shaping Europe's digital future (europa.eu).

[11] Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future (europa.eu).

[12] Communication on boosting startups and innovation in trustworthy artificial intelligence | Shaping Europe's digital future (europa.eu).

[13] Speech from 19 February 2024, Brussels, at: Renew Europe event at the European Parliament (europa.eu).

[14] The seminal article is: Graef, Inge, Market Definition and Market Power in Data: The Case of Online Platforms (September 8, 2015). World Competition: Law and Economics Review, Vol. 38, No. 4 (2015), p. 473-506., Available at SSRN.

[15] Commission (2024), Commission Notice on the definition of the relevant market for the purposes of Union competition law, Brussels, 8.2.2024, C(2023) 6789 final, pp. 25f.

[16] As noted by the UK's Competition and Markets Authority (CMA), several developers of foundation models (FMs), such as Microsoft, Amazon and Google, "own key infrastructure for producing and distributing FMs such as data centres, servers, network infrastructure and data repositories". CMA (2023), AI Foundation Models: Initial Report, 18 September 2023, p. 16.

[17] Tone Knapstad (2023) Digital dominance: assessing market definition and market power for online platforms under Article 102 TFEU, European Competition Journal, DOI: 10.1080/17441056.2023.2280334.

licensing features"[18], which is relevant given OpenAI's current strategy to license high-quality news training data from large media companies (see our analysis in section 3.1 below).

Similarly, there is a risk that dominant firms in genAI markets may engage in anti-competitive practices such as self-preferencing, tying, or predatory pricing. For instance, "a dominant cloud platform could bundle or promote its own foundation model at the expense of rival offerings, while a dominant foundation model provider could cut off or degrade API (Application Programming Interfaces) access to rival application developers."[19] The Federal Trade Commission (FTC) recently discussed several examples for bundling and tying at the intersection of the cloud and AI markets. For example, big tech firms like Google and Amazon apparently sell access to their most powerful AI tools exclusively to developers that use the corporation's cloud services.[20] In the genAI market, this could manifest in terms of restrictive access to cloud infrastructure or certain "frontier models" that give downstream users state-of-the-art capabilities for developing or updating their AI-based products and services. To avoid this scenario, some researchers have recommended incentivising or even mandating open foundation models, i.e. models with widely available weights, because "closed model developers exert greater power in defining and restricting use cases they deem unacceptable, whereas downstream consumers of foundation models can better make these decisions for themselves with open models".[21]

Finally, it may be necessary to rethink certain elements of merger control. While the acquisition of innovative start-ups by established players has been a common feature in digital markets over the past decade or so, the current strategy in the genAI landscape seems to be the formation of strategic partnerships between established Big Tech firms and innovative start-ups. As of writing, such partnerships have been concluded between Microsoft/OpenAI and Mistral; Hugging Face/Amazon; Cohere/Google and Nvidia; Stability AI/Amazon; and Inflection AI/Microsoft and Nvidia.[22] As noted by the British competition authority, "[v]ertical relationships may also occur when companies use long-term partnerships and strategic investments as an alternative to outright acquisitions and vertical integration".[23] Since such partnerships often provide higher-priority access to computing power as well as cheaper rates, they may nevertheless stifle innovation and reduce competition, requiring careful scrutiny. As the EU Competition Commissioner recently made clear, the strong barriers to entry and the lack of AI talent make it unlikely that the next wave of AI disruption will be "driven by a handful of college drop-outs who somehow manage to outperform Microsoft's partner Open AI or Google's DeepMind".[24] From that perspective, one could qualify partnerships such as the one between Microsoft/OpenAI as a novel type of "killer acquisition" in EU competition law.[25] However, if these partnerships do not fall under the EU's Merger Regulation, as the Commission's preliminary review of the Microsoft/OpenAI partnership indicates, "the Commission should explore other avenues, including investigating them as anti-competitive agreements under Article 101 of the Treaty on the Functioning

[18] Commission (2024), Commission Notice on the definition of the relevant market for the purposes of Union competition law, Brussels, 8.2.2024, C(2023) 6789 final, p. 39.

[19] Max von Thun (2024), EU does not need to wait for the AI Act to act – Euractiv.

[20] See the discussion of the FTC meeting at: Generative AI Concerns Fuel Federal Trade Commission Scrutiny of Cloud Services — Center for Journalism & Liberty (journalismliberty.org).

[21] Rishi Bommasani, Sayash Kapoor, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Daniel Zhang, Marietje Schaake, Daniel E. Ho, Arvind Narayanan, Percy Liang (2023), Considerations for Governing Open Foundation Models, Governing-Open-Foundation-Models.pdf (stanford.edu), p. 5.

[22] Monopoly Power Is the Elephant in the Room in the AI Debate | TechPolicy.Press.

[23] CMA (2023), AI Foundation Models: Initial Report, 18 September 2023, p. 18, also p. 35.

[24] Speech from 19 February 2024, Brussels, at: Renew Europe event at the European Parliament (europa.eu).

[25] Radsch (2023), The real story of the OpenAI debacle is the tyranny of big tech | Courtney Radsch | The Guardian.

of the European Union (TFEU) or simply updating the Merger Regulation to ensure that it covers such investments and partnerships."[26]

Overall, the Commission can rely on the traditional three pillars of EU competition law – agreements, dominant positions, and mergers to counteract emerging anti-competitive behaviour in the field of genAI. In addition, it is worth pointing out that the DMA could be applied to many of the problems discussed below, too, if the Commission were to decide, for instance, to designate dominant cloud services as gatekeepers while also including foundation models under the purview of the legislation.[27]

3 Competition problems in the value chain for generative AI as a service

In the dynamically evolving landscape of genAI, a distinct value chain has started to emerge, consisting of three parts that contribute to their development and deployment. This value chain, while bearing similarities to traditional AI structures, introduces unique elements, particularly in the realm of foundation models. As the Commission's DG Competition seeks to understand genAI, we suggest analysing three different integral parts that underpin its value chain: Firstly, genAI infrastructure consisting of (high-quality) datasets and computing power (we further subdivide computing power into the markets for chips and cloud platforms). Secondly, the market for foundation models and, thirdly, the downstream services and applications that use the foundation models as an input. In the following, we aim to shed light on each component's role and the competitive dynamics within the genAI ecosystem.

3.1 GenAI infrastructure

3.1.1 Training data

In the early stages of model development (pre-training stage), in particular, genAI companies use large datasets, often comprising hundreds or thousands of gigabytes, to build the foundational knowledge of their models. GenAI companies have different ways of obtaining such datasets. They can either buy them, which is usually not a problem as there are many companies offering (high-quality) data,[28] or crawl the web themselves in order to obtain sufficient data.[29] Some datasets can be used even free of charge. A recent comprehensive review of the existing available dataset resources for LLM training covers 444 datasets, with the total data size surpassing 774.5 TB for training datasets.[30] Recently, the use of exclusive proprietary data has also become common practice. If the training data is derived from a platform service, such as a social media platform like "X", the owner of this platform can train his model on the existing data generated on this platform (here: X trained its "Grok" model on all tweets, while restricting access to the twitter archive for external developers). In addition, once a generative AI service has been set up, such as ChatGPT, the owner of this service can use the question-and-answer pairs (i.e. prompt by user and reply by model) to further develop his model. This constitutes a first-mover advantage.

[26] Max von Thun (2024), EU does not need to wait for the AI Act to act – Euractiv.
[27] Max von Thun (2024), EU does not need to wait for the AI Act to act – Euractiv.
[28] See: Rohstoff Daten: Eine falsche Analogie bremst die europäische Digitalpolitik aus - Tagesspiegel Background.
[29] Web crawling involves the use of automated programs to collect data from web pages for model training purposes.
[30] Liu et al. (2024), [2402.18041] Datasets for Large Language Models: A Comprehensive Survey (arxiv.org).

Popular datasets often used at this pre-training stage include the C4 dataset, compiled by AllenAI from the extensive Common Crawl collection, and The Pile, a collection of 22 high-quality datasets curated by EleutherAI from various sources including PubMed and GitHub.[31] In addition, the Project Gutenberg Corpus of over 50,000 public domain books and the LAION datasets of millions of image-text pairs are key resources for training language models and image generation algorithms. The ready availability of these data sources shows that the availability of data is not currently a sufficient barrier to entry. Although there are economies of scale at play, there is also a lot of competition among companies offering data. As there are no switching costs for companies that buy such data to train their foundation models, there are no apparent lock-ins.

However, firms like OpenAI are increasingly turning to proprietary data (e.g. from large media companies), which might be more difficult (or even impossible, in case of an exclusive license) to access for start-ups and emerging rivals. Privacy lawsuits by artists and newspapers such as the NYT are already showing that high quality proprietary data may have been used unlawfully to train the models because it was often available online and could therefore be scraped by means of web crawling, even by newcomers. If, in the future, such data is better protected and then exclusively licensed to a single provider of language models, such as OpenAI, this might limit the ability of new entrants to train their own models. In fact, there are currently several media stories circulating that suggest that this is becoming the dominant strategy of OpenAI, perhaps as a method to crowd out competitors. OpenAI has entered into a multi-year agreement with Axel Springer that reportedly does not commit either side to exclusivity, but does allow OpenAI to use the publisher's content, including articles from Business Insider and Politico, to train its AI models and improve ChatGPT. In return, Axel Springer will receive undisclosed payments from OpenAI and plans to support its AI-driven projects using OpenAI's technology.[32] A similar license has already been negotiated for some of the archive of The Associated Press.[33] There are also licensing talks going on with media companies like CNN, Fox Corp, and Time.[34] According to media information, OpenAI offers between $1 million and $5 million a year to license these training datasets.[35] Similarly, Google agreed on a $60 million deal with Reddit to use their data to train AI.[36] For OpenAI or Google, these costs are marginal, but they might deter new entrants as part of a "raising rivals' costs" strategy[37] and lead to essential facility problems. This problem looks larger in light of predictions that AI developers will run out of previously unused high-quality language data by 2026 (defined as the total stock of unlabelled data available on the internet), meaning that all commonly available data on the web will have been used for training at that point and that all additional data, which might be required for the next stage of genAI models, has to come from private, proprietary sources such as internal firm data.[38] This suggests that, unless new (synthetic) data sources emerge, the growth of large genAI models that rely on large online datasets may slow, and that OpenAI will increasingly rely on proprietary data licensed from a few dominant companies.

[31] For the argumentation in this paragraph, see: CMA (2023), AI Foundation Models: Initial Report, 18 September 2023, p. 11.

[32] OpenAI inks deal with Axel Springer on licensing news for model training | TechCrunch.

[33] ChatGPT-maker OpenAI signs deal with AP to license news stories | AP News.

[34] OpenAI Is in Licensing Talks With Major Media Outlets: Report (businessinsider.com).

[35] OpenAI's news publisher deals reportedly top out at $5 million a year - The Verge.

[36] Exclusive: Reddit in AI content licensing deal with Google | Reuters.

[37] Salop, S. C., & Scheffman, D. T. (1983). Raising Rivals' Costs. The American Economic Review, 73(2), 267–271.

[38] While low-quality language and image data will last much longer, between 2030 and 2050 and 2030 and 2060, respectively. Villalobos et al. (2022), Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning, 2211.04325.pdf (arxiv.org).

However, current research provides some theoretical reasons not to be overly concerned about data access problems. In particular, recent research strongly indicates that data efficiency and AI model architecture will significantly improve, meaning that advanced LLMs will require less and less data to achieve superior performance.[39] This contrasts with the conventional wisdom that scaling up the number of model parameters and the size of the training data dramatically increases the quality and capability of LLMs. For example, a synthetic dataset of short stories in easy language can be used to train and evaluate language models that are much smaller than state-of-the-art models (i.e. they have less than 10 million total parameters, compared to a total of about 1.76 trillion parameters for OpenAI's GPT-4[40]), or have much simpler network architectures (with only a single transformer block[41]), yet still produce consistent, long text and demonstrate reasoning capabilities.[42] Moreover, recent research suggests that when such smaller models are combined (called "blending"), they can even outperform the larger models. For example, the researchers note that blending three models of moderate size can rival or even surpass the performance metrics of a substantially larger model like ChatGPT. More generally, to gain an insight into the future direction of the field, it is helpful to consider the case of the recent "BabyLM Challenge", which investigated how to train a "BabyLM" on just 100 million words (inspired by the fact that current LLMs are now trained using >1000 times as much language data as a child).[43] The winning entries in this competition achieved superior performance compared to models trained on trillions of words.[44] Other notable submissions achieved significant results either by training on shorter input sequences or by using a student-teacher training approach, where a smaller (student) model learns from a larger pre-trained (teacher) model. While not all efforts were successful, this recent research progress strongly indicates that training data will not become a significant barrier to market entry for genAI companies.

All in all, we do not currently see competition problems on the market for high quality data because recent research indicates that advanced LLMs will require less and less data to achieve superior performance. Nevertheless, exclusive licensing agreements between providers of online-news or other owners of high-quality datasets should be closely monitored as this might make it more difficult for new competitors to obtain high quality data.

3.1.2 Computing power ("compute")

Once a prospective market entrant has collected or bought sufficient training data, it needs to find enough computing power to process the data and create a foundation model. Modern AI models require substantial computational resources for pre-training, fine-tuning, and inference, often necessitating specialized hardware such as GPUs (Graphical Processing Units) or TPUs (Tensor Processing Units) equipped with accelerator chips. While GPUs were initially designed and used for 3D graphics, they have now become mainstays of modern machine learning, due to their fast parallel computing. TPUs are, in contrast, less flexible, application-specific integrated circuits developed by

[39] See, e.g.: Williams (2024), 100x less compute with GPT-level LLM performance, Tech Radar.

[40] Blending Is All You Need: Cheaper, Better Alternative to Trillion-Parameters LLM (arxiv.org).

[41] This refers to the quintessential architecture that is responsible for the current AI hype and progress in developed, namely so-called pre-trained "transformer" language models. The seminal paper is: [1706.03762] Attention Is All You Need (arxiv.org).

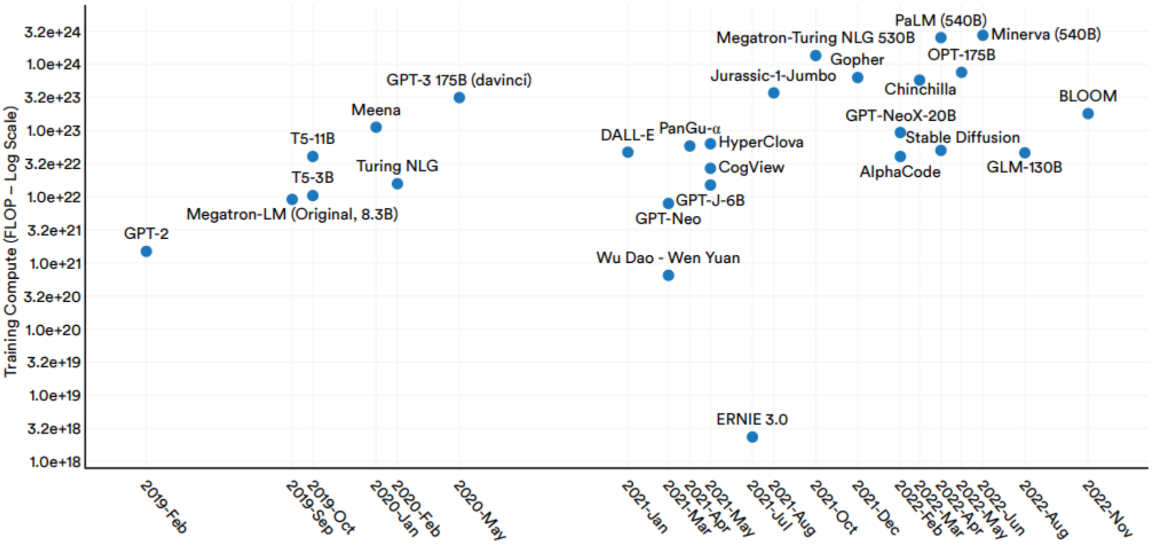
[42] [2305.07759] TinyStories: How Small Can Language Models Be and Still Speak Coherent English? (arxiv.org).

[43] The proceedings of the BabyLM Challenge can be found here: Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning - ACL Anthology.

[44] Using the LTG-BERT architecture. See the summary in: 2023.conll-babylm.1.pdf (aclanthology.org).

Google, which can be faster or slower than GPUs, depending on the use case.[45] These components are thus pivotal for handling the immense data processing demands of genAI models. Over time, the training compute for large language and multimodal models has tended to increase (Figure 2) – a trend that still holds.[46]

Fig. 2: Training Compute (FLOP) of Select Large Language and Multimodal Models, 2019–22



Source: Epoch, 2022 | Chart: 2023 AI Index Report.

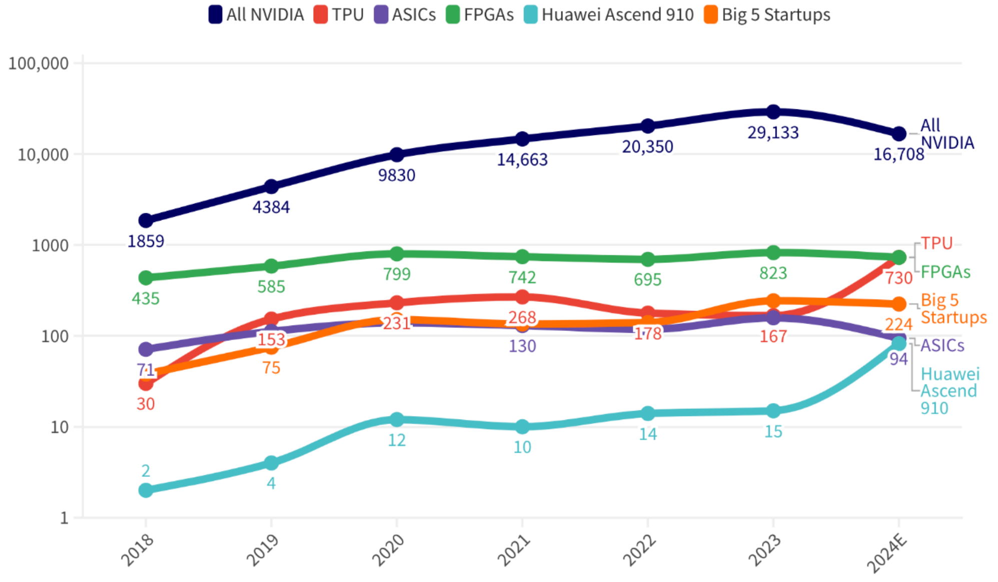
From the perspective of the genAI value chain, the market for suitable advanced processors is currently dominated by a few key players like NVIDIA (especially NVIDIA's H100 GPU) and Google, with manufacturing largely concentrated in the hands of entities such as Taiwan Semiconductor Manufacturing Company Limited (TSMC). The high entry barriers in terms of R&D investment and technical expertise make this segment relatively impenetrable for new entrants. Stakeholders note that computing power required for AI inference "can be particularly intensive at scale",[47] meaning that new entrants might struggle to scale their services as they face limitations in computing capabilities. In the future, some competition could come from Intel's recently announced new computer chips,[48] including Gaudi3, a chip for genAI software, as well as from AMD's new accelerators and processors geared toward running large language models.[49]

At the time of writing, NVIDIA – whose share price has jumped significantly during the current AI rush – is the most dominant firm. At a recent tech summit organised by the FTC, Corey Quinn, Chief Cloud Economist at the Duckbill Group, emphasised that "[a]ll roads lead to one company and that is NVIDIA" and noted that there is no transparency in how they distribute the key assets they hold.[50] To preserve competition in markets for computer chips used in datacentres, the FTC is currently considering blocking acquisitions planned by NVIDIA.[51] This dominant position of NVIDIA chips can be gauged from the State of AI Report "Compute Index", which tracks the utilisation of various AI chips in AI research papers (Figure 3). This is based on the assumption that usage of chips in AI research papers (representing early adopters and state-of-the-art science) is a leading indicator of industry usage.

[45] See: CPU vs GPU vs TPU: Understanding the difference b/w them (serverguy.com).
[46] See: HAI_AI-Index-Report_2023.pdf (stanford.edu), p. 61.
[47] CMA (2023), AI Foundation Models: Initial Report, 18 September 2023, p. 14.
[48] Intel unveils Gaudi3 AI chip to compete with Nvidia and AMD (cnbc.com).
[49] AMD releases new chips to power faster AI training - The Verge.
[50] Is "More Clouds" the Future We Want? A Dispatch from the FTC AI Tech Summit | TechPolicy.Press.
[51] See: Nvidia/Arm, In the Matter of | Federal Trade Commission (ftc.gov).

According to Figure 2, there were over 16,000 AI research papers in 2023 that made use of any type of NVIDIA hardware. By contrast, Google's TPU comes in at 730, and the total for the five big AI start-ups comes to 224 papers.

Fig. 3: Cited chip usage in AI papers (logarithmic representation)



Source: State of AI Report Compute Index and Zeta Alpha

Source: [State of AI Report Compute Index and Zeta Alpha](). Notes: This figure presents the number of open-source AI papers that cite the use of specific AI chips according to analysis by Zeta Alpha. The numbers for 2024 are extrapolated from Q4/23.

In this respect, the recent announcement by Meta's CEO, Mark Zuckerberg, regarding the development of the Llama-3 foundation model, underscores a significant shift in the dynamics of computational resources within the industry. Meta's planned acquisition of 350,000 H100 GPUs to support this objective is a strategic move. Given that there are only expected to be 2 to 2.5 million H100s globally by the end of 2024, Meta will control 14% of the global H100 GPU market.[52] Crucially, this substantial concentration of resources not only amplifies Meta's computational capabilities but also establishes a formidable economic barrier to emerging companies in the field of genAI. As noted by the CMA, the few companies dominating hardware are "creating dependencies for startups and other companies developing and deploying AI tools", not least because "[m]odern GPU architectures encounter bottlenecks with transformers" and "there is a shortage of server GPUs for AI purposes".[53] Tellingly, Sam Altmann, the head of OpenAI, is currently trying to raise $7 trillion for new AI chip project.[54]

Recently, the Commission has recognised this link between computing power and the development of AI more explicitly as a strategic priority. The EU has launched an antitrust investigation into the NVIDIA-dominated chip market and is planning comprehensive risk assessments for semiconductor and AI technologies as part of its economic security strategy.[55] According to an AI initiative package presented in late January 2024, the Commission aims to give European SMEs and start-ups privileged access to

[52] According to calculations in: Engels' pause; Silicon sensitivity; GPUs galore; Afrobeat ++ #457 (exponentialview.co).
[53] CMA (2023), AI Foundation Models: Initial Report, 18 September 2023, p. 33.
[54] OpenAI CEO Sam Altman reportedly seeks trillions of dollars for AI chip project (cnbc.com).
[55] EU Begins Early-Stage Probe Into AI Chip Market Abuses that Nvidia Dominates - Bloomberg.

the domestic supercomputer network.[56] To this end, the Commission will amend the European High Performance Computing Regulation. As a new pillar for the activities of the EU Supercomputer Joint Undertaking, it plans to establish AI factories whose use will be free of charge. However, according to media reports, there is more demand than there is capacity, so start-ups and SMEs have not been given enough of a chance.[57] In addition, the German AI Association points out that previous projects with European supercomputers used far fewer GPUs than are needed for the latest generation of AI (i.e. genAI), and therefore calls for a more powerful and flexible supercomputing infrastructure.[58] The trend towards smaller models described above, which is already evident in research but has not yet been sufficiently recognised in political and media discourse, suggests that these measures will be less important for a competitive market environment than commonly expected, at least in the medium and long term. In this regard, another notable line of research was recently presented by the Linux Foundation, whose novel AI model architecture (so-called “receptance weighted key value”) might be able to reduce the computational GPU requirements of training LLMs by up to 100 times.[59] So far, however, this approach has encountered certain technical challenges that are not yet completely solved.

Given the high costs and scarcity of advanced computational hardware, start-ups typically buy computing power from bigger companies using cloud platforms which offer businesses scalable and cost-effective access to the necessary computational power. Countries such as the US and Japan, as well as private companies such as Amazon, Google, and Alibaba, have long been ahead of their competitors thanks to extensive investment in the necessary cloud infrastructure.[60] The market for cloud services is currently dominated by a few major providers, with extensive infrastructure and privileged access to hardware, namely Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) (commonly referred as “hyper-scalers”). A recent UK report by Ofcom found that, in the UK market, AWS and Microsoft had a combined market share of 70% to 80% in 2021, with Google’s share between 5% and 10%.[61] As of early 2024, AWS maintains the highest market share globally at 32%, followed by Microsoft Azure (23%) and Google Cloud (10%), while Alibaba Cloud and Tencent Cloud are notable players in the Asia-Pacific market.[62] The higher figures for the UK, compared to the more equitable shares on the global level, suggests that the hyper-scalers are particularly dominant in Europe. This concentrated market structure might lead to problematic path-dependencies in the future, as significant investments in time are required to re-engineer an app or other AI services when moving, as a firm, from one cloud provider to another. Further lock-in effects are being created due to discounting schemes and high switching costs.[63] In this context, Microsoft’s USD 13 billion partnership with OpenAI has anti-competitive implications due to OpenAI’s dependence on Microsoft’s computing infrastructure.[64]

[56] Commission, COMMUNICATION on boosting startups and innovation in trustworthy artificial intelligence, Brussels, 24.1.2024 , COM(2024) 28 final.

[57] According to reports in: Europe.Table # 617 / 25. Januar 2024.

[58] LEAM-MBS_KIBV_webversion_mitAnhang_V2_2023.pdf.

[59] Williams (2024), 100x less compute with GPT-level LLM performance, Tech Radar.

[60] See the evidence collected in: Vili Lehdonvirta (2022), Cloud Empires: How Digital Platforms Are Overtaking the State and How We Can Regain Control.

[61] Ofcom (2023), Cloud services market study: final report, Cloud services market study final report (ofcom.org.uk), p. 3.

[62] Data taken from: 2024 Cloud Market Share Analysis: Decoding Industry Leaders and Trends (hava.io).

[63] Is “More Clouds” the Future We Want? A Dispatch from the FTC AI Tech Summit | TechPolicy.Press.

[64] Civil Society Groups Urge UK to Investigate Microsoft's Monopolistic Partnership with OpenAI — Open Markets Institute.

Since most SMEs need the advanced technology of cloud computing services to develop their own genAI services, the current hype around AI had the consequence that the big hyperscalers have recently been able boost their revenues through their cloud platforms – a trend that is likely to continue. During the second quarter of 2023, Microsoft's Intelligent Cloud division generated almost USD 24 billion in revenues, while AWS produced revenues exceeding USD 22 billion in the same period.[65] In the last quarter of 2023, Microsoft's revenue from its Azure cloud platform and related services rose by as much as 30%.[66] Google Cloud comes in third, with its revenues surpassing USD 8 billion in the second quarter of 2023. Of course, these recent increases cannot be explained solely by the fact that SMEs are developing their AI services, but it is unlikely to be a complete coincidence. Overall, the global hyperscale cloud market is expected to grow to USD 1,261.15 billion by 2028.[67]

To sum up, we see competition problems in the market for computing power due to a shortage of adequate chips. This shortage comes from the fact that the market for state-of-the-art chips is dominated by one company (NVIDIA). Due to technological restraints, this company cannot easily increase the production of these chips. The shortage of adequate chips and lack of entrants in the cloud computing market[68] has resulted in Big Tech companies getting most of this scarce resource. The fact that they provide start-ups with access to this resource via cloud services does not solve the problem, and even somewhat aggravates it, for which there are three reasons. Firstly, there are concerns that leading start-ups such as OpenAI, and incumbents like Big Tech firms, are more likely to get prioritised access to these cloud services for AI training, as they make deals to hold larger compute clusters.[69] Secondly, existing users of cloud services, including start-ups, face switching costs if they later want to change the provider. Thirdly, some of the cloud service providers develop their own chips and/or their own genAI models, increasing the risk of vertical integration and self-preferencing. For example, AWS is now starting to develop its own internal foundation models (e.g. Amazon Titan LLMs),[70] meaning that they will become competitors to external model developers. This is reminiscent of Amazon's behaviour on its marketplace, where it also provides the infrastructure while competing with some of the users. This problem is already present for Google and Microsoft, due to their unique position in the value chain for genAI services. The unique mix of high demand, too few suppliers of computing power, high switching costs, and growing vertical integration – even now – poses non-negligible risks to competition, for instance in the form of self-preferencing or discrimination. In other words, this market structure may lead to genAI-induced market power in markets, along supply chains or in ecosystems.

3.2 GenAI training (foundation models)

At the core of the genAI ecosystem are foundation models. These are extensive deep learning models pre-trained on vast datasets and capable of being adapted to a wide range of tasks. The development of such models requires expertise across multiple domains, including data preparation, model architecture, and continuous fine-tuning. In addition, the high fixed costs of training foundation

[65] Data taken from: Cloud hyperscaler quarterly revenue by vendor 2023 | Statista.

[66] Microsoft Azure revenue up 30%, with help from AI, as tech giant beats overall expectations – GeekWire.

[67] Global Hyperscale Cloud Market: Analysis By End-User, By Region Size and Trends with Impact of COVID-19 and Forecast up to 2028 (researchandmarkets.com).

[68] There is little venture capital for companies interested in building expensive cloud computing infrastructure. Is "More Clouds" the Future We Want? A Dispatch from the FTC AI Tech Summit | TechPolicy.Press.

[69] CMA (2023), AI Foundation Models: Initial Report, 18 September 2023, p. 36.

[70] AWS Bedrock distances firm from Microsoft, Google in generative AI race | ITPro.

models and their low marginal deployment costs lead to significant economies of scale, with unit costs declining as deployment increases. Finally, there are economies of scope and first mover advantages in this market, coupled with barriers such as scarcity of human capital, specific high-quality data, computing power and intellectual property, which further increase the benefits of concentration and might even lead to "natural monopolies".[71]

According to the AI Index 2023 Annual Report, there were, in 2022, 32 significant industry-produced foundation models compared to just three produced by academia.[72] 35 suppliers may not at first sound like a significant competitive problem, but competition problems could arise if the differences in quality are so great that downstream firms can only use one or two of those models to avoid being at a competitive disadvantage to others (winner-takes-most market). According to the State of AI Report 2023, the "uncontested most generally capable" foundation model is GPT-4 by OpenAI, beating every other LLM on both classic benchmarks and exams designed to evaluate humans (although other models have recently caught up, see Table 1 below).[73] Moreover, it must be noted that foundation models significantly improve their performance when they scale up (i.e. increase the amount of training data and/or the number of parameters in the neural net),[74] thus rewarding the providers of the largest models. In that sense, "size matters" in this market. This phenomenon is often called "emergent capabilities", defined as novel abilities that cannot be predicted simply by extrapolating from the performance of smaller models.[75] Put simply, this means that while a smaller model might struggle with a mathematics problem, a larger model could suddenly learn this feature and thus experience a "jump" in its capabilities. In other words, the emergent capabilities of foundation models suggest that additional scaling could further expand the range of capabilities of language models in unknown ways, which is certainly what many current market actors, above all OpenAI, are hoping for.

Nevertheless, counteracting this tendency towards concentration, there are currently several foundation models available that can be used for important (industrial) downstream tasks, some proprietary and "black box", others open source and transparent.[76] Hugging Face, via its "Hugging Face Hub", for example, offers over 60K models, all open source and publicly available, on an online platform.[77] Furthermore, a team at Stanford recently developed a "Foundation Model Transparency Index" assessing OpenAI (GPT-4), Anthropic (Claude 2), Google (PaLM 2), Meta (Llama 2), Inflection (Inflection-1), Amazon (Titan Text), Cohere (Command), AI21 Labs (Jurassic-2), Hugging Face (BLOOMZ; as host of BigScience), and Stability AI (Stable Diffusion 2), without finding a single provider that dominated in all categories.[78] Instead, each of these models have certain advantages and

[71] Vipra and Korinek (2023), Market concentration implications of foundation models: The Invisible Hand of ChatGPT | Brookings.

[72] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault (2023), "The AI Index 2023 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

[73] Welcome to State of AI Report 2023.

[74] On the emergent features of genAI, see: The Unpredictable Abilities Emerging From Large AI Models | Quanta Magazine. On the risks emerging during a polycrisis, see: Küsters (2023), cepAdhoc_AI_as_Systemic_Risk_in_a_Polycrisis.pdf.

[75] [2206.07682] Emergent Abilities of Large Language Models (arxiv.org).

[76] The release of foundation models is a gradient, see: Rishi Bommasani, Sayash Kapoor, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Daniel Zhang, Marietje Schaake, Daniel E. Ho, Arvind Narayanan, Percy Liang (2023), Considerations for Governing Open Foundation Models, Governing-Open-Foundation-Models.pdf (stanford.edu), p. 3.

[77] Introducing the Private Hub: A New Way to Build With Machine Learning (huggingface.co).

[78] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, Percy Liang (2023), The Foundation Model Transparency Index, [2310.12941] The Foundation Model Transparency Index (arxiv.org).

disadvantages. Moreover, the trend towards more efficient training methods and smaller, task-specific models may gradually democratize this level of the genAI value chain and thus erode the current market dominance of OpenAI and its GPT-4 model.[79]

Tab. 1: Crowd-sourced ranking of foundation models (top 10)

Rank	Foundation model	Arena Elo Score	Organization	License	Knowledge cutoff
1	GPT-4-0125-preview	1253	OpenAI	Proprietary	2023/4
2	GPT-4-1106-preview	1252	OpenAI	Proprietary	2023/4
3	Bard (Gemini Pro)	1224	Google	Proprietary	Online
4	GPT-4-0314	1190	OpenAI	Proprietary	2021/9
5	GPT-4-0613	1162	OpenAI	Proprietary	2021/9
6	Mistral Medium	1150	Mistral	Proprietary	Unknown
7	Claude-1	1149	Anthropic	Proprietary	Unknown
8	Claude-2.0	1132	Anthropic	Proprietary	Unknown
9	Gemini Pro (Dev API)	1120	Google	Proprietary	2023/4
10	Claude-2.1	1119	Anthropic	Proprietary	Unknown

Source: Own assessment based on: https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard. Notes: Arena Elo Score is a rating system used on the Chatbot Arena platform to gauge the skill levels of different LLMs by their outcomes in anonymous, randomized matches, similar to the Elo rating system in chess.

Over the past half-year, in particular, the rivalry among proprietary foundation models has surged, leading to a notable convergence in their capabilities. On the leaderboard of Chatbot Arena, which uses an ELO rating system (a widely-used rating system used in chess and other competitive games[80]) reflecting hundreds of thousands of human evaluations of LLMs, the different iterations of GPT-4 Turbo are still at the top, but in recent months, other models have caught up. Notably, Bard Pro is (as of writing) in second place. To put this in context, the "Chatbot Arena" website serves as a benchmarking platform for evaluating the performance of various LLMs in real-world scenarios, offering a useful proxy through which to assess the current market structure. By rigorously testing LLMs across a diverse range of tasks and challenges, the Arena allows users to receive objective, quantifiable metrics that reflect the capabilities, strengths, and weaknesses of each model. Although the platform only gives a dynamic snapshot of market evolution, the Commission can leverage the information from this benchmarking tool to make informed decisions and anticipate future trends – e.g. by tracking whether incumbent models improve their advantage over time or whether other models catch up – in the rapidly evolving foundation model sector. As can also be seen from the Table, Mistral has unveiled a model rumoured to rival GPT-4 in terms of performance. Not yet reflected in the Table is Gemini Ultra, which has just been made public by Google. Like OpenAI's state-of-the-art models, it is multimodal (meaning that it can mix text, images, and audio in the same session) and, currently unlike GPT-4 , it can handle an input of up to one million tokens.[81] Overall, GPT-4's leading position appears stable at

[79] For a good example, see: Phi-2: The surprising power of small language models - Microsoft Research.

[80] The scoring system updates the ratings based on the performance of each model in these battles, with improvements for better-than-expected outcomes and reductions for worse-than-expected outcomes, while ensuring a scalable, incremental evaluation of models handling open-ended problems.

[81] Introducing Gemini 1.5, Google's next-generation AI model (blog.google).

the moment but might soon come under threat, particularly from Gemini Ultra and Mixtral. In other words, it may be possible to rule out the scenario of one dominant model "to rule them all".

After pre-processing the training data and finding sufficient computing power, be it directly through hardware acquisitions or renting of cloud space, a prospective market entrant also needs human expertise to develop foundation models on this basis. GenAI has significantly influenced the job market in 2023, as evidenced by a more than 1000 % increase in genAI-related job posts on Upwork in Q2 2023.[82] Among current hiring plans, there is particular interest in Natural Language Processing, TensorFlow, image processing, and other AI-related expertise.[83] However, the number of global AI developers is limited, with estimates from 2019 ranging from 22,000 highly-trained AI specialists up to 300,000 AI researchers and practitioners working within broader technical teams.[84]

In this competition for human capital, European companies face a significant upward battle, as attracting AI talent is highly challenging and depends on a region's initial research sectors and business activities. The "superstar" and "winner-take-most" dynamics prevalent in digital economies imply that only a limited number of locations may predominantly facilitate medium-term AI-related growth.[85] An historical analysis found that 56% of the economically most impactful new technologies came from just two US locations, Silicon Valley and the Northeast Corridor, not least because these locations are where new technologies were pioneered and have remained crucial for the technology's high-skill jobs for decades.[86] This trend seems to hold for the genAI value chain, too: A Brookings report from 2023 found that nearly half of job postings for genAI positions over the previous 11 months were concentrated in just six large US metropolitan areas: San Francisco, San José (Calif.), New York, Los Angeles, Boston, and Seattle.[87]

Overall, we see that competition problems on the market for genAI training are arising for three reasons: Firstly, there are significant economies of scale and scope, such as high fixed costs for training, that make market entry expensive. Secondly, LLMs significantly improve their performance when scaling up, especially due to their "emergent capabilities", thus rewarding the largest providers. As a result, the market for foundation models is a "winner-takes-most" market as users typically want to use the best foundation model for their purpose. Therefore, it is not just a matter of fixed and marginal cost relations but of competition for the best quality. Since every additional user helps to improve this quality (by creating data than can again be used for fine-tuning the model), the revenue model might deviate from classical business models in the sense that firms try to scale quickly (comparable to social media companies in their early days). That makes it very hard for start-ups to compete with established providers of foundation models and reinforces our previous point that the strategic partnership between OpenAI and Microsoft should be seen as a form of vertical integration that could harm competition. Thirdly, there is a shortage of human expertise for developing foundation models which further limits competition in this market and makes it particularly difficult for European firms.

[82] 10 most in-demand generative AI skills | CIO.

[83] 10 most in-demand generative AI skills | CIO.

[84] Chapter 6: The war for talent (stateofai2019.com).

[85] The geography of AI | Brookings.

[86] The Diffusion of New Technologies | NBER.

[87] Building AI cities: How to spread the benefits of an emerging technology across more of America | Brookings.

Our assessment that at the moment, OpenAI seems to win this "winner-takes-most" market certainly holds true for the short and medium-term, but our review of the current literature and evidence of recent catching-up by competing foundation models (Gemini, Mistral) suggest that these competition restrictions might ease in the longer term.

3.3 GenAI deployment: Down-stream services

GenAI applications and services are specific software solutions developed on top of foundation models, and tailored for distinct tasks like customer service, content generation, or specialized consulting. The key to success in this realm lies in the ability to fine-tune foundation models with niche or proprietary data, allowing for the creation of highly specialized and competitive applications.[88] This sector will likely see the entry of both existing AI service providers and new niche players, not least because there are many open access models freely available (see above). However, the best, and thus most competitive, apps and services will require state-of-the-art-models, currently GPT-4. In this context, it will be increasingly difficult on the genAI deployment market to compete with Big Tech firms like Microsoft and Google, which can include such state-of-the-art-models in all their services and products due to their partnerships described above and often already have market power on these down-stream markets (e.g. for office software and search engines).

In this context, further competition problems might emerge in the future due to a new "app store" format recently pioneered by OpenAI, which invites developers and companies to build custom "GPTs" that add functionality to the chatbot (GPTs can be configured by setting specific prompts and parameters in ChatGPT to achieve a particular output).[89] This recent transformation of ChatGPT into an "app platform" might introduce a new dynamic into the distribution of market power in the down-stream sector for genAI-based services and products. By providing a centralised hub for deploying and accessing GPT-powered applications, ChatGPT facilitates a streamlined ecosystem where developers can reach a broad audience with minimal friction. However, this centralisation raises concerns about the potential for market power to be transferred down-stream, particularly if the platform begins to replicate the functionality of successful third-party applications (similar to what Amazon did to some of its most successful sellers[90]). In addition, the inherent network effects of the platform, characterised by the increasing value of the service with each additional user and application, could lead to significant lock-in effects, further driven by the fact that OpenAI could use the increased number of "prompts and replies" to further improve its model in future iterations (GPT-5, GPT-6 etc.). If such an "app store" format became commonplace, it would impact user retention and the attractiveness of individual LLMs, and consequently put up barriers to entry for potential competitors.

All in all, while there are numerous foundation models available that allow easy access to the market for down-stream genAI applications and services, the "winner-takes-most" dynamic means that only a handful of providers will be successful long-term. In this context, Big Tech's current dominant positions on many of these down-stream markets (and their partnerships with leading foundation model developers) mean that they will soon capture those markets and offer these services themselves. In other words, this creates two disadvantages for prospective market entrants: Firstly, start-ups have to compete with large firms such as Microsoft and Google that have preferential

[88] See: Exploring opportunities in the gen AI value chain | McKinsey.

[89] See the reporting in: OpenAI's New App Store Could Turn ChatGPT Into an Everything App | WIRED. For a more sceptical take, see: The GPT Store isn't ChatGPT's 'app store' – but it's still significant for marketers (econsultancy.com).

[90] See: Amazon copied products and rigged search results, documents show (reuters.com).

access to advanced models. Secondly, incumbents make the novel technology a feature of their older services and thus leverage their market power and their existing customers on these down-stream markets. In the future, a third hurdle might emerge from the spread of "app-store" models. These would create a scenario similar to Google's position on the market for smartphone software, whereby Google offers an operating system and an app store as well as important apps. We think it is likely that, without rapid intervention from the competition authorities, such a situation will also occur in the genAI ecosystem.

4 Resilience considerations

While the primary objective of EU competition law is to ensure economic freedom and protect consumer interests, the evolving nature of genAI with its grave geopolitical, political, and security implications requires a broader perspective. As has been recently noted by Paul Tucker, "if future consumer welfare (crudely, lower prices for equivalent quality goods) is the only test, then the regime will not take into account the risks and prospective costs to the stability of the political system of concentrated private economic power developing into concentrated political power."[91] Such thinking is not alien to EU competition law, which since its inception has been influenced by ordoliberal thinking about the interrelationships between economic power, political stability, and individual freedom.[92] However, there is a fine line to be walked. On the one hand, competition law enforcement should still be rules-based and not overly discretionary. On the other hand, prominent researchers have warned that "the role of antitrust in promoting competition could well be undermined if antitrust is called upon or expected to address problems not directly relating to competition".[93]

With respect to competition in genAI, this broader conception of competition law implies that competition enforcers should also consider general concerns, which are outside a narrow definition of competition law but which may also be relevant given the Commission's objective to promote strategic autonomy in the digital sphere ("digital sovereignty").[94] In fact, the EU Competition Commissioner herself recently noted that "digital markets are wide-reaching, sometimes affecting the economy in ways you might not have expected", implying that "that competition policy has to work together with digital regulation in this fast paced, dynamic economy".[95] Here, we point to several overlapping concerns at the intersection of competition and AI governance, namely existential security risks posed by AI systems, risks to political discourse and democracy, the rise of deepfakes and misinformation, the potential for malicious services, and military applications of AI. For each of these risks, DG COMP should consider whether more competition could reduce, or conversely even exacerbate, them. We assess these risks below.

In essence, our recommendation follows similar strategies that have been proposed for a modest re-conceptualisation of "green antitrust", a movement which advocates a relaxation of enforcement to

[91] See: An interview with Sir Paul Tucker - The Platform Law Blog.

[92] See: Anselm Küsters (2023), The Making and Unmaking of Ordoliberal Language. A Digital Conceptual History of European Competition Law, Studien zur europäischen Rechtsgeschichte 340, Frankfurt am Main: Klostermann 2023.

[93] Shapiro, Carl, (2018), Antitrust in a time of populism, International Journal of Industrial Organization, 61, issue C, p. 714-748, here: p. 716.

[94] Armin Steinbach (2023), EU's Turn to 'Strategic Autonomy': Leeway for Policy Action and Points of Conflict | European Journal of International Law | Oxford Academic (oup.com).

[95] Speech from 19 February 2024, Brussels, at: Renew Europe event at the European Parliament (europa.eu).

allow sustainability-enhancing collaborations with anticompetitive effects.[96] While sustainability should not be the primary objective of competition law (which is economic freedom as well as consumer law), one should nevertheless consider which type of enforcement actions could also lead to positive spill-over effects for sustainability or at least indirect environmental benefits.[97] Indeed, discussing the potential trade-offs between competition law enforcement and the EU green Deal, the Commission itself has asked whether "we can do more, to apply our rules in ways that better support the Green Deal".[98] Similarly, the Commission should ask in which technology-centred cases competition enforcement could help to reduce the broader non-economic risks of genAI, such as existential risks, and in which cases one should be more cautious about swift enforcement and look instead at possible negative side-effects outside the economic sector that would be fuelled by increased competitive pressure. In other words, the Commission should strengthen its enforcement in those cases that allow the punishment of conduct that is simultaneously anticompetitive and increases broader societal harms. Much can be achieved by proper selection of cases and priority setting,[99] without changing the substance of the law.

Existential Risks: As foundation models improve their performance when scaling up, they increase the likelihood of unpredictable phenomena, including severe risks.[100] The development and deployment of genAI systems throughout Europe could thus pose, or increase, potential existential risks, including the unintended consequences of autonomous decision-making. Competition can play a dual role here. On the one hand, increased competition could spur rapid advancements and diffusion of AI technologies, possibly outpacing regulatory and ethical frameworks such as the EU AI Act. On the other hand, a competitive environment could foster innovation in safety and ethical AI, as companies seek differentiation through responsible practices. Competition enforcers should encourage transparency (as mandated by the AI Act for high-risk models) and ethical standards as competitive factors. While regulation will always provide a second-best solution, more competition and clearer liability can optimise the trade-off between limiting risks and exploiting opportunities in line with preferences, as cep argued in a recent game-theoretical study.[101]

Risks to political discourse and democracy: GenAI technology might lead to centralised control over language technologies and corresponding influence over political discourse and the formation of public opinion in democracies. At a recent panel discussion in the European Parliament, industry experts emphasised the crucial role of competition law in protecting democracy in the AI sector, raised concerns about AI's potential to exacerbate existing problems in digital discourse, and called for regulation and enforcement ensuring competitive fairness to protect societal and democratic values.[102] Indeed, a key risk arising from the LLMs underlying genAI services is their potential to undermine democracy by concentrating language technologies in the hands of a few private

[96] See, e.g.: Sustainability, antitrust and the EU Green Deal | Netherlands | Global law firm | Norton Rose Fulbright.

[97] Marco Colino, Sandra, Antitrust's Environmental Footprint: Redefining the Boundaries of Green Antitrust (February 19, 2024). North Carolina Law Review, Vol. 103, No. 1, 2024, The Chinese University of Hong Kong Faculty of Law Research Paper No. 2024-01, Available at SSRN.

[98] conference 2021 - European Commission (europa.eu).

[99] See: Brook, Or and Cseres, Kati, Policy Report: Priority Setting in EU and National Competition Law Enforcement (September 28, 2021). Available at SSRN: https://ssrn.com/abstract=3930189.

[100] On the emergent features of genAI, see: The Unpredictable Abilities Emerging From Large AI Models | Quanta Magazine. On the risks emerging during a polycrisis, see: Küsters (2023), cepAdhoc_AI_as_Systemic_Risk_in_a_Polycrisis.pdf.

[101] Küsters and Vöpel (2024), Weniger KI-Risiken durch mehr Wettbewerb, cepInput.

[102] See the summary in: Ensuring competition in AI will also preserve democracy, experts say – Euractiv.

companies that could dictate the future of political discourse and public deliberation.[103] As described above, language models are trained on diverse data sources such as news and books, but who decides which data to include - and which to omit? If genAI applications such as ChatGPT become the next "internet platform", replacing traditional search services such as Google, they would have a huge impact on shaping public opinion and access to information. Findings from computer scientists show that current pre-trained language models have political biases that reinforce the polarisation present in their underlying training data, propagating social biases into hate speech and misinformation detectors.[104] A limited number of linguistic models with their inherent biases, driven by the ideological choices behind their development, therefore pose a risk to the democratic process by privatising what constitutes public discourse and potentially distorting the information that citizens can access. A good example is the motivation behind Elon Musk's Grok chatbot trained on Twitter data and with an explicit "anti-woke" bias. Another example concerns OpenAI's recent agreement with Springer, mentioned earlier, to the effect that ChatGPT users will get summaries of "selected" articles from Axel Springer's publications, sometimes even if they are behind a paywall.[105] The snippets will be accompanied both by attribution and links to the full articles. Accordingly, European competition authorities may in the future need to consider the implications of data selection, inherent biases, and the potential for genAI technologies to become dominant platforms for access to information, potentially requiring new frameworks to ensure diversity in digital communication spaces.

Deepfakes and genAI-fuelled misinformation: Alongside the broader issue of political discourse, the rapid advancement of AI in generating text and images might lead to escalating political misinformation, which is of relevance in light of the many elections taking place in 2024. According to the latest Freedom House report, at least 47 governments deployed commentators to manipulate online discussions in their favour during the coverage period, while generative AI tools were utilized in at least 16 countries to sow doubt, smear opponents, or influence public debate.[106] The use of AI by presidential candidates in Argentina during the 2023 campaign, the "first AI election",[107] further underscores the growing influence of this technology in political arenas. In light of genAI's capability to create realistic deepfakes and spread misinformation, which could undermine social trust and the integrity of information, a winner-takes-all dynamic in an oligopolistic market might exacerbate the problem by incentivising model developers in a "rat race"[108] to enable the production of more engaging, albeit potentially misleading, content. However, a competitive landscape could also encourage the development of counter-technologies for deepfake detection and information verification as an additional parameter for model providers to distinguish themselves through product quality.[109] As noted above, competition law enforcers should therefore monitor the genAI market structure to prevent monopolistic control over these technologies, ensuring accessibility and diversity in tools that combat misinformation. At the same time, existing research shows that the persuasive power of both political ads and online misinformation is often overstated and suggests that policy

[103] For this argument, see the essay: Whoever Controls Language Models Controls Politics – HANNES BAJOHR.

[104] [2305.08283] From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models (arxiv.org).

[105] OpenAI inks deal with Axel Springer on licensing news for model training | TechCrunch.

[106] The Repressive Power of Artificial Intelligence | Freedom House.

[107] Is Argentina the First A.I. Election? - The New York Times (nytimes.com).

[108] For this, see our study: Küsters and Vöpel (2024), Weniger KI-Risiken durch mehr Wettbewerb, cepInput.

[109] For this, see our study: Küsters and Vöpel (2024), Weniger KI-Risiken durch mehr Wettbewerb, cepInput.

should focus on preventing abuse in smaller, more local elections (where the impact is larger) and in mitigating bias.[110]

Malicious services: A recent in-depth analysis of more than 200 concrete malicious services, that utilize AI, highlighted the role of foundation models, especially those developed by OpenAI, in powering various illicit tools and services like BadGPT, XXXGPT, and Evil-GPT.[111] Based on data from 13,353 listings across nine underground marketplaces and forums, the findings revealed that a significant majority (93.4%) of malicious services offered malware generation capabilities, with some even evading virus detection. The study found that OpenAI's GPT-3.5 and GPT-4 models were the most commonly exploited backends in these malicious services, suggesting that these models currently seem to be the most effective ones and cannot at present be equated with free open source models developed by start-ups. The findings suggest that European competition authorities may need to scrutinise the dominance and use of OpenAI foundational models in the creation of malicious AI services, to ensure that market power is not abused in ways that harm consumers. The fact that illegal activities have been carried out with closed models, such as those of OpenAI, presumably via API access, and not with easily available open-access models, such as Llama2, may cast doubt on the common argument that ensuring open access and interoperability of models will help to reduce misuse.

Military Applications of genAI: Finally, the utilization of genAI in military applications raises ethical and security concerns. This has become an acute issue as OpenAI has recently removed its ban on military use of its AI tools, following joint AI research with the US Department of Defence.[112] A recent empirical study investigating the escalation risks of actions taken by large language models and the AI agents based on them, in simulated wargames, found that "models tend to develop arms-race dynamics, leading to greater conflict, and in rare cases, even to the deployment of nuclear weapons".[113] The EU AI Act exempts military applications from its focus. While increased competition in the genAI sector could lead to a proliferation of advanced military AI systems, potentially destabilizing global security dynamics, it could also stimulate innovation in defensive and security-enhancing AI technologies. While it is certainly beyond the remit of competition law enforcers to navigate this delicate balance, the current geopolitical context will likely lead to calls for consideration of the strategic implications of genAI for national and international security when deciding on concrete cases.

All in all, our overview of "resilience", i.e. non-economic factors that might play a role in future genAI cases, is based on the argument that: instead of focusing only on the possibility that increased competition through antitrust enforcement could either completely block the development of European genAI services or solve all of genAI's emerging problems (be they technical or societal), the EU Commission should reflect on ways in which the law's enforcement could have positive (indirect) repercussions on the larger issues that are at stake. Is it conceivable that increasing competitive pressure in the oligopolistic genAI market could lead to safer models by reducing "rat race"

[110] See the literature survey in: Scott Babwah Brennen & Matt Perault (2023), The new political ad machine: Policy frameworks for political ads in an age of AI, GAI-and-political-ads.pdf (unc.edu).

[111] Zilong Lin, Jian Cui, Xiaojing Liao, and XiaoFeng Wang (2024), Malla: Demystifying Real-world Large Language Model Integrated Malicious Services, 2401.03315.pdf (arxiv.org).

[112] OpenAI quietly removes ban on military use of its AI tools (cnbc.com).

[113] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, Jacquelyn Schneider (2024), Escalation Risks from Language Models in Military and Diplomatic Decision-Making, 2401.03408.pdf (arxiv.org).

incentives?[114] Whether or not this assumption holds up to more theoretical and empirical scrutiny, actively addressing these larger trade-offs, by issuing guidance and analyses, might help to manage expectations and shape the AI policy discourse.

5 Conclusion

The value chain for genAI services is a multifaceted ecosystem that can be divided into three parts. The first part relates to genAI infrastructure, which consists of the market for large-scale datasets and the market for computing power. Both are essential inputs for the training and development of foundation models, which is the second part of the AI value chain. The third part is the market (or markets) for B2B or B2C downstream services and applications that use a foundation model as input. For each of these different parts, this cepInput has investigated potential competition problems. In particular, we have looked at whether size matters (e.g. high fixed costs), whether there are significant switching costs, and whether there are scarce resources that limit competition. What are the key findings based on our analysis?

1. GenAI infrastructure

As regards training data, we do not currently see competition problems because recent research indicates that advanced LLMs will require less and less data to achieve superior performance. Nevertheless, exclusive licensing agreements between providers of online-news or other owners of high-quality datasets should be closely monitored as this might make it more difficult for new competitors to obtain high quality data.

As regards computing power, we see competition problems due to a shortage of adequate chips. This shortage comes from the fact that the market for state-of-the-art chips is dominated by one company (NVIDIA). Due to technological restraints, this company cannot easily increase the production of these chips. The shortage of adequate chips and the buying-power and influence of Big Tech companies in a limited cloud service market means that incumbents are getting most of this scarce resource. The fact that they provide start-ups with access to this resource via cloud services does not solve the problem and even somewhat aggravates it, for which there are three reasons: Firstly, there are concerns that leading start-ups such as OpenAI, and incumbents like Big Tech firms, are more likely to get prioritised access to computing power via cloud services for AI training, as they make deals to hold larger compute clusters. Secondly, start-ups face switching costs, if they want to change the cloud service provider. Thirdly, some of the cloud service providers develop their own chips and/or their own genAI models, increasing the risk of vertical integration and self-preferencing. This unique mix of high demand, too few suppliers of computing power, high switching costs, and growing vertical integration poses – even now – non-negligible risks to competition, for instance in the form of self-preferencing or discrimination.

2. GenAI training (foundation models)

As regards foundation models, we see competition problems for three reasons: Firstly, there are significant economies of scale and scope, such as high fixed costs for training, that make market entry expensive. Secondly, LLMs significantly improve their performance when scaling up, especially due to their “emergent capabilities”, thus rewarding the largest providers. As a result,

[114] For this, see our study: Küsters and Vöpel (2024), Weniger KI-Risiken durch mehr Wettbewerb, cepInput.

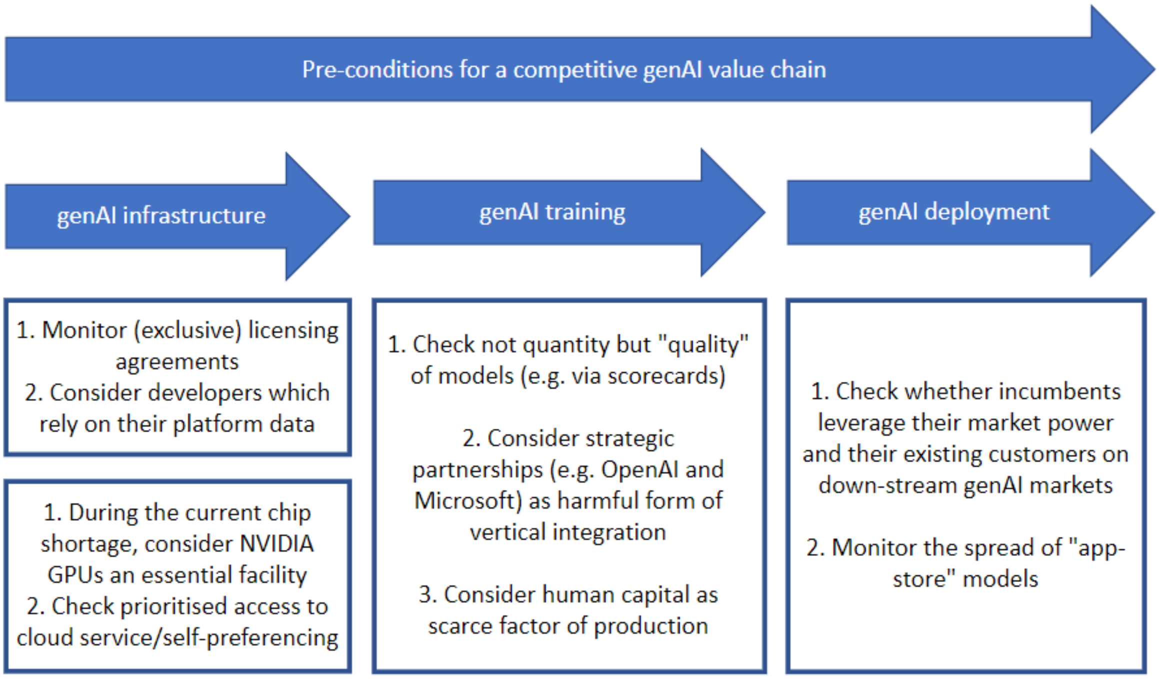
the market for foundation models is a “winner-takes-most” market as users typically want to use the best foundation model for their purpose. That makes it very hard for start-ups to compete with established providers of foundation models and reinforces our previous point that the strategic partnership between OpenAI and Microsoft should be seen as a form of vertical integration that could harm competition. Thirdly, there is a shortage of human expertise for developing foundation models which further limits competition in this market and makes it particularly difficult for European firms. This assessment certainly holds true for the short and medium-term, but our review of the current literature and evidence of recent catching-up by competing foundation models (Gemini, Mistral) suggest that these competition restrictions might ease in the longer term.

3. GenAI deployment (applications and services)

As regards down-stream genAI applications and services, the “winner-takes-most” dynamic means that only a handful of providers will be successful long-term. In this context, Big Tech’s current dominant positions on many of these down-stream markets (and their partnerships with leading foundation model developers) mean that they will soon capture those markets and offer these services themselves. In other words, this creates two disadvantages for prospective market entrants: Firstly, start-ups have to compete with large firms such as Microsoft and Google that have preferential access to advanced models. Secondly, incumbents make the novel technology a feature of their older services and thus leverage their market power and their existing customers on these down-stream markets. In the future, a third hurdle might emerge from the spread of “app-store” models. These would create a scenario similar to Google’s position on the market for smartphone software, whereby Google offers an operating system and an app store as well as important apps. We think it is likely that, without rapid intervention from the competition authorities, such a situation will also occur in the genAI ecosystem.

Ultimately, understanding the dynamics of each tier of the value chain is crucial for EU competition law enforcers when designing their strategy for tackling the genAI market. In addition, however, they also need to **take an holistic view in order to understand the degree to which certain actors are already vertically integrated**. A key example relates to Microsoft and its strategic partnerships with OpenAI and Mistral AI, which can be considered as a single market player or, for the purposes of competition law, a single “undertaking”. In our assessment, this undertaking has a strong position in the value chain due to its possession of high-quality data, cloud servers, human capital, control over several leading foundation models, and existing market power in down-stream office services, where genAI applications will became very relevant and commercially lucrative. Similar, but to lesser degree, this also applies to Google, Meta, and Amazon.

Fig. 4: Pre-conditions for a competitive genAI value chain


Pre-conditions for a competitive genAI value chain
genAI infrastructure
genAI training
genAI deployment
1. Monitor (exclusive) licensing agreements
2. Consider developers which rely on their platform data
1. During the current chip shortage, consider NVIDIA GPUs an essential facility
2. Check prioritised access to cloud service/self-preferencing
1. Check not quantity but "quality" of models (e.g. via scorecards)
2. Consider strategic partnerships (e.g. OpenAI and Microsoft) as harmful form of vertical integration
3. Consider human capital as scarce factor of production
1. Check whether incumbents leverage their market power and their existing customers on down-stream genAI markets
2. Monitor the spread of "app-store" models

Source: own representation.

As this technology continues to evolve, it will be imperative for DG COMP, through its analysis and formulation of guidelines, to play a pivotal role in actively shaping the competitive landscape of this transformative technology. While the primary aim of EU competition law is to safeguard economic freedom and consumer welfare, the unique characteristics of the genAI value chain demand a broader approach. By considering existential risks and the impact on information integrity, and perhaps even the implications for security, DG COMP can align its enforcement actions with the Commission’s broader objectives of digital sovereignty and societal well-being. This approach requires a nuanced understanding of the interplay between competition and the aforesaid externalities, ensuring that competition enforcement contributes positively to managing the risks and harnessing the opportunities presented by genAI. To support the Commission in this task, we end this paper by providing a “map” which, based on the foregoing analysis, identifies the pre-conditions for a competitive genAI market along the whole value chain (Figure 4).

Centrum für Europäische Politik
FREIBURG | BERLIN

Author:

Dr. Anselm Küsters, LL.M., Head of Division Digitalisation & New Technologies
kuesters@cep.eu

Dr. Matthias Kullas, Head of Division Internal Market & Competition
kullas@cep.eu

Centrum für Europäische Politik FREIBURG | BERLIN
Kaiser-Joseph-Straße 266 | D-79098 Freiburg
Schiffbauerdamm 40 Räume 4205/06 | D-10117 Berlin
Tel. + 49 761 38693-0

The **Centrum für Europäische Politik** FREIBURG | BERLIN, the **Centre de Politique Européenne** PARIS, and the **Centro Politiche Europee** ROMA form the **Centres for European Policy Network** FREIBURG | BERLIN | PARIS | ROMA.

Free of vested interests and party-politically neutral, the Centres for European Policy Network provides analysis and evaluation of European Union policy, aimed at supporting European integration and upholding the principles of a free-market economic system.