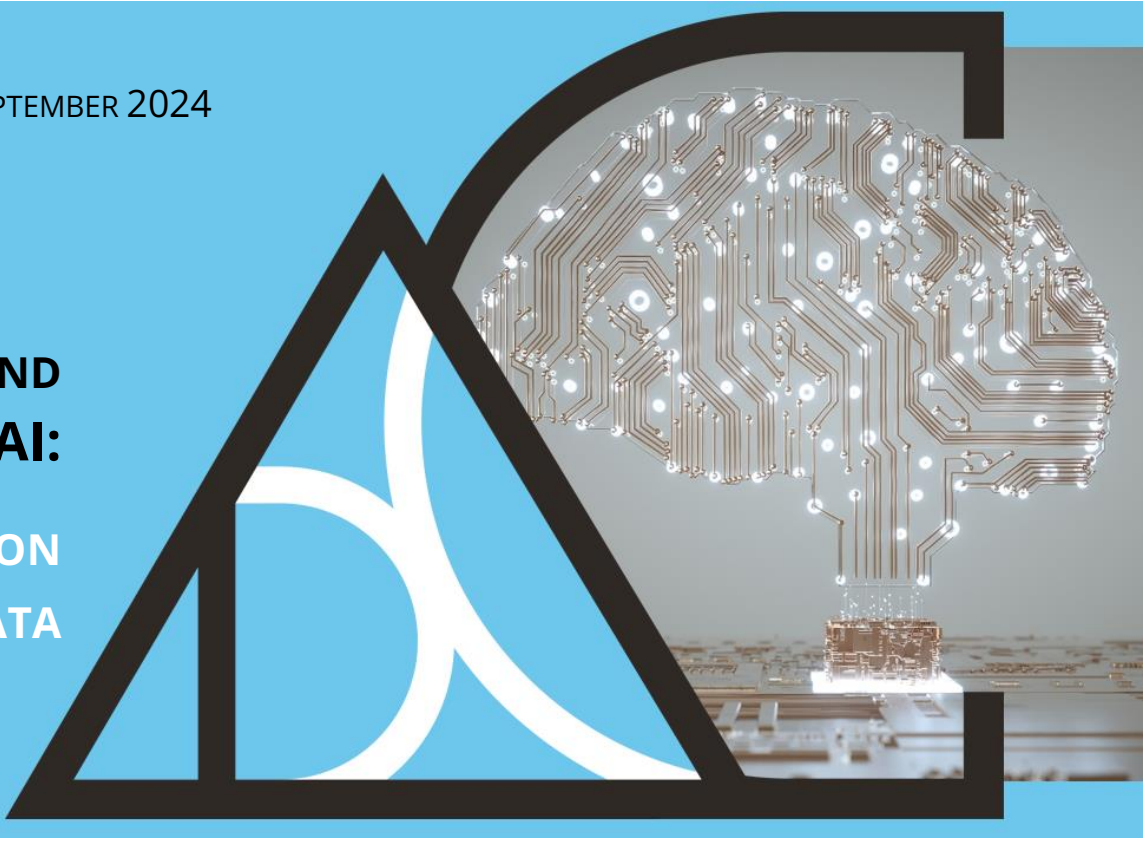
AdC Short Papers | September 2024



Competition and Generative AI:

Zooming in on Data

Since late-2022, generative artificial intelligence (AI) has disrupted the digital sector. This is a new type of AI capable of producing content similar to what a human would do.

In November 2023, the Portuguese Competition Authority (AdC) published an issues paper[1] tracking these developments, mapping the key determinants of competition in generative AI and identifying the main risks to competition in the sector.

This short paper expands on that exercise by considering how the sector has evolved since November 2023. It covers the use of data in the development of generative AI, and the increasing importance of data licensing agreements.

I. Introduction

Along with computing power and know-how, data is a key input in developing generative AI models. Data comes in many formats – text, image, video, audio, etc. – depending on the type of generative AI model being developed. During training, patterns in the data are embedded into the generative AI model, enabling it to generate new content on demand by replicating these patterns. This **training data** can originate from different sources and, as addressed below, may entail acquisition costs. In addition, some models may combine their embedded knowledge to external and verifiable sources of information (also known as grounding), such as search results, to improve their reliability, scope of knowledge and access more up-to-date information, while reducing the likelihood of model "hallucinations".[2,3] Finally, developers also collect **monitoring data** on the training and performance of their models, such as tracking user behaviour, to experiment and optimize future iterations of the models.[4]

A level playing field in the access to large, diverse, updated and high-quality datasets

[1] Available here.
[2] Hallucinations, in the case of Large Language Models, are responses by AI models that are inaccurate, misleading or non-sensical but presented as factual.
[3] In the case of Large Language Models, this is done via retrieval-augmented generation (RAG) techniques. This technique is used in services such as ChatGPT, Perplexity AI or You.com. See also the seminal paper Lewis et al. (2021) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Available here.
[4] See more in sections II, III and III.1 of the Issues Paper on Generative AI, here.

will be crucial to foster a competitive environment in generative AI markets.[5] Models typically perform better on tasks they have already been exposed to during training. Consequently, the richness of training data may significantly affect model performance,[6] especially with regards to long-tail prompts.[7] These are highly specific prompts that, individually, very few people make but, collectively, represent a large portion of total prompts. On long-tail prompts models are more likely to veer outside familiar territory and output generation errors, such as hallucinations in Large Language Models (LLMs).[8]

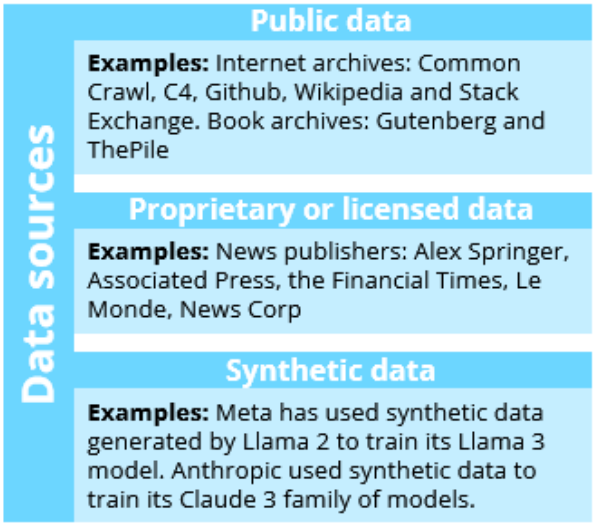
Pursuing this goal requires competition authorities to identify possible bottlenecks in access to data in the development of generative AI. It is important to determine whether training data is non-substitutable or hard to replicate by competitors, such that a small number of firms may be able to create and exploit bottlenecks in AI markets, in a way that harms competition, innovation and ultimately consumers. On the other hand, if similar knowledge can be extracted from different data sources, then the risk of bottlenecks is lower.

The identification of possible bottlenecks should consider recent trends regarding the use of data in generative AI:

- **Until recently, AI developers mostly used public data to train AI models. Generative AI developers have become increasingly less transparent over the training data they use**.[9] Because of this, it is more difficult to ascertain exclusivities in access to data, data licensing agreements and how essential each dataset is to develop performant models.[10]
- **Data licensing agreements seem to have become more prevalent**. These are agreements between sources of data – such as publishers, stock image repositories or social networks – and generative AI developers.
- **Synthetic data and data pre-processing seem to be playing an increasingly**

[5] See, e.g., Longpre et al. (2023) A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. Available here.
[6] See, e.g., Kandpal et al. (2023). Large Language Models Struggle to Learn Long-Tail Knowledge. Available here.
[7] The long tail refers to a statistical property where a large number of occurrences appear infrequently but still account for a significant portion of the total. This is, for example, a feature of e-commerce sales. There are many products which sell few units but, in total, these products account for a large portion of total sales.
[8] A similar effect happens with search engines, where the scale of data may improve the quality of search results especially in the case of long-tail queries or less visited webpages. See, e.g., discussion in Appendix I from the CMA's Online platforms and digital advertising market study, available here.
[9] For example, for GPT-3, OpenAI lists the names of the datasets it used. For GPT-4, OpenAI only says it uses both publicly available and data licensed from third-party providers. Lastly, for GPT-4o, OpenAI provides no details. The same evolution happened between Llama 2 and Llama 3, where Meta went from providing a brief description of each dataset to only mentioning training data is obtained from publicly available sources. This is also highlighted in Longpre et al. (2023). The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. Available here.
[10] The AI Act may have some relevant provisions on this regard. For instance, under the AI Act, providers of a high-risk AI system shall provide technical documentation of the model, including a description of the training datasets used, information about their provenance, scope and main characteristics; how the data was obtained and selected; labelling procedures and data cleaning methodologies (see, e.g. Annex IV referred to in Article 11(1) of the AI Act). This provision, under Chapter 3, shall enter into force on 2 August 2025 (Article 113).

important role in training efficient and performant generative AI models.[11]



The following chapters start from this baseline and map the determinants of competition involving data in generative AI.

II. Intellectual property issues over data

A significant portion of the training data for generative AI models is publicly available. This includes scraped webpages, images or videos, as well as book repositories. To the extent that training data is publicly available, access to data is more equalised among developers and the key barriers are the computing resources and the expertise needed to work with the data.

However, the publicly available data used by generative AI developers may be subject to intellectual property (IP) rights. As business models and commercial applications have begun consolidating and maturing, the holders of IP rights have started demanding compensation for the use of their content. IP holders argue that AI developers use their content without authorization during training and inference, and that AI systems can reproduce or generate derivative content based on their IP. As such, some IP holders have started to implement tools that protect their content against unauthorised use.[12]

There is, therefore, a legal risk for AI developers associated with the use of much of publicly available training data. It remains an open question whether AI developers may only use data covered by IP rights if authorized by IP holders. This will depend on existing IP legislation and on how it is interpreted. Numerous lawsuits have already been filed by IP holders against generative AI developers for copyright infringement.[13]

To mitigate the legal risk over copyright infringement, generative AI developers have begun entering into data licensing agreements with IP holders. ***Vis-à-vis* the AI sector, IP holders are producers and/or distributors of data**. IP holders often publish and distribute significant volumes of original content that is instrumental in training generative AI models (see Box 1).

[11] See sections II and III of this document.
[12] For example, researchers have developed a tool that alters images in training data and which, if used, harm the performance of the image generation AI model. The changes applied to the images are not perceptible to a human. See Shan et al. (2024). Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. Available here.
[13] For example, New York Times v. OpenAI; Author's Guild v. OpenAI; a group of Artists v. Stability AI and v. Midjourney; or Sony, Universal and Warner v. Suno and Udio.

Box 1 – Examples of licensed content

News publishers produce large volumes of relatively formal text which may be key for teaching LLMs the structure of different languages and facts about the world. For example, Alex Springer, Associated Press, the Financial Times, Le Monde, News Corp and Prisa Media have entered into data licensing agreements with OpenAI.

LLMs can learn informal language and expand the topics they are exposed to via **social networks** where users upload text in the form of posts and comments. For example, Reddit has entered into data licensing agreements with both Google and OpenAI, and there are reports of talks with Meta and Apple.

Likewise, access to **stock image repositories** may be crucial for image generation models, as these produce or distribute large volumes of images.

The same applies to **video-sharing platforms** and video generating models. Shutterstock, for example, has been licensing its images, videos and music to AI developers such as OpenAI or Meta.

There has been a wave of data licensing agreements. This suggests IP holders are open to licensing their data to AI developers. This could give some players in the digital sector additional monetisation strategies, as in the case of publishers, social networks, video-sharing platforms or other platforms with many users who upload content.

There are already a few public examples of data agreements that have been reported to be valued tens or hundreds of millions of dollars.[14] For example, Reddit has revealed that it is at the early stages of data monetisation and that its data will be key in training LLMs. It also mentioned it has entered into data licensing agreements totally valued at $203 million.[15] Platforms may also intensify data collection. For example, in May 2024, Meta announced that it planned to use publicly available posts from Facebook and Instagram users to train its future AI models.[16] However, data available in digital platforms may still be subject to IP rights.[17]

IP holders may prefer to license their data for grounding instead of solely for training AI models. Grounding requires recurrent use of their data, which could provide a continuous of revenue for IP holders, whereas, during training, data is typically used few times.

The shift from publicly available data to proprietary data and data licensing agreements creates barriers to entry and expansion and may reinforce market power of leading companies with access to data. Some AI developers may have better means to acquire data and to manage the transaction costs associated with data licensing. In addition, some IP holders may choose not to sell their data or may sell it selectively. Indeed, some IP holders

[14] For example, OpenAI will reportedly pay $250 million over five years to access News Corp's content, and the partnership between Google and Reddit is reportedly valued at $60 million.
[15] See Reddit's Form S-1, filed February 2024.
[16] See the blog post by Meta about the change, here.
[17] In June, Meta also announced it received a request from the Irish Data Protection Commission, to delay training LLMs using public content shared by adults on Facebook and Instagram (see here).

are also AI developers, due to their presence in other digital markets or their digital ecosystems. For this reason, they may find little incentive in sharing their data, especially if their data reinforces or sustains their position in some digital market.

Exclusivities in the access to training data may be especially harmful to competition, particularly if datasets are hard to substitute or to replicate. Generative AI models will likely differentiate on their reliability on the long-tail. This suggests that access to more data can improve model performance, even if there are significant redundancies in the data, or the AI can generalise beyond its training data. AI models may also differentiate by specialising on certain tasks, topics or domains. If such specialisations require specific datasets, differences in access to these specific datasets, namely via exclusivity provisions or discriminatory access, may create barriers to entry.

As such, exclusivities and preferential access can give AI developers undue competitive advantages, blocking competitors from using that data. This increases market power and hampers overall innovation in the AI sector. In addition, such exclusivities or preferential access may potentially infringe competition law both in Portugal and in the EU. This could happen, for example, if a firm has a dominant position in the relevant data market and it gives exclusive or preferential access to this data to its own offerings or to a third party, at the expense of competitors.

A consistent flow of fresh data can also play a relevant role in ensuring performant AI models. In some AI applications, the value of data may decrease with age, meaning that updated information is needed to gain a competitive edge over rivals. This may increase the value of newer datasets and place data holders in a better position to develop AI models that require this data.

Promoting competition and ensuring a level playing field in access to data requires a streamlined data licensing process. Bilateral and bespoke data agreements may increase transaction costs and entail significant barriers for entrants in the AI sector. In addition, discriminatory conditions in data agreements, such as exclusivities, may exacerbate barriers to entry and expansion. Similarly, price structures that require developers to pay for data upfront may also favour larger players with deeper pockets, especially in the case of foundation models.

A number of options could be considered to streamline access to data. For example, options such as serving data through open APIs, bundling licenses, and adopting pay-as-you-go pricing structures, to avoid scale effects, may be effective ways to mitigate these concerns. Making public datasets easily available and with no unnecessary restrictions, such as the public domain book repositories of national libraries or court rulings, may also contribute to reducing data-driven barriers to entry and expansion in generative AI.

Shift to proprietary data may reinforce concentration

The shift from publicly available data to proprietary data, as IP holders have begun demanding compensation, may reinforce data-driven advantages and market concentration.

Data exclusivities can be harmful to competition

Exclusivities and preferential access to data can be especially harmful to competition and possibly infringe competition law both in Portugal and in the EU.

Streamlining access to data to ensure level playing field

Streamlining access to data for developers will be key to ensure a level playing field (e.g., by serving data through open APIs, pay-as-you-go pricing structures or making public datasets easily available).

III.Synthetic data

Synthetic data refers to artificially generated data by an algorithm, namely generative AI models, which can then be used to train new generative AI models.[18] Conceptually, the use of synthetic data follows a principle similar to transfer learning (e.g., fine-tuning).[19] In transfer learning, new models are trained on top of prior models; in synthetic data the new model is trained using data created by the old model.

The use of synthetic data seems to have become more widespread, as many generative AI developers resort to this type of data in the development of their models. For example, Meta has used synthetic data generated by Llama 2 to train its Llama 3 model.[20] Likewise, Anthropic used synthetic data to train its Claude 3 family of models.[21] Nvidia has also developed a model specialised on generating synthetic data.[22] Lastly, OpenAI has created a guide on how to generate synthetic data using GPT-4.[23]

Synthetic data can serve as an alternative to real data, effectively reducing barriers to entry, data acquisition costs and be an alternative for developers that do not have access to specific datasets or want to increase the richness and diversity of the training data for certain subjects (data augmentation). This way, access to data is mediated by the generative AI models creating the synthetic dataset, which replicates the patterns in the original dataset. This can also be a way to protect privacy, trade secrets and other sensitive information.[24] In addition, synthetic data can be particularly useful as it is often more structured and easier to use in the development of AI models.

However, over-relying on synthetic data may hamper model performance, thereby limiting the effectiveness of synthetic data in reducing differences in access to data across AI developers. As the share of synthetic data in the training dataset increases, model performance may be degraded. This may happen because synthetic data is one step removed from

[18] Synthetic data can also be invaluable beyond generative AI development. For example, developers of machine learning models may resort to synthetic data if they find it challenging to access to real-world data. This has numerous applications, e.g. in areas such as healthcare or fraud detection. See, e.g., Ktena et al. (2024). Generative models improve fairness of medical classifiers under distribution shifts; or Benalcazar et al (2023). Synthetic ID Card Image Generation for Improving Presentation Attack Detection.
[19] See more in Section III of the AdC's Issues Paper on generative AI, here.
[20] See the blog post by Meta introducing Llama 3, here, and the paper presenting Llama 3, here.
[21] See Model Card for the Claude 3 family of models, by Anthropic, here.
[22] See the blog post by Nvidia presenting its Nemotron-4 340B family of models, here.
[23] See article from the OpenAI Cookbook, here.
[24] See, e.g., Afonja et al. (2024). The Crossroads of Innovation and Privacy: Private Synthetic Data for Generative AI. Available here.

the original dataset, meaning generation errors and biases from the original model will be passed to the new model as training data. Furthermore, synthetic data tends to be less diverse than the underlying real data.[25] To mitigate these issues, developers must use synthetic data in tandem with real data,[26] limiting the extent to which synthetic data may replicate or substitute real data.

Generative AI developers may also introduce limitations on the uses other developers may give to synthetic data. For instance, in its Terms of Use, OpenAI does not allow users to use the output of its models to develop models that compete with OpenAI.[27] In addition, the community license agreement of Meta's Llama 3 does not allow developers to use data generated by Llama models to improve other AI models.[28] Such terms may limit the commercial viability or the scale of models trained using synthetic data. In contrast, other models, such as Nvidia's Nemotron-4 340B, are open and introduce no restrictions to developers.[29]

Synthetic data will likely remain an alternative to real-world data, despite its limitations. Synthetic data may, at least partially and to a certain extent, substitute and replicate real-world data. As such, it is able to mitigate some of the differences in access to data and create a more level playing field in the sector. Still, they are not sufficient to ensure contestability AI developers with access to real-world data may enjoy a competitive edge over rivals.

Synthetic data may not be sufficient to ensure contestability

Synthetic data is increasingly used by developers and can reduce entry barriers and data acquisition costs, but it presents limitations and AI developers with access to real-world data may still enjoy a competitive edge.

IV. Data pre-processing

Data pre-processing, also known as data filtering or data selection, is an essential step of training any AI model and a key differentiating factor as it can improve model quality. The goal is to curate raw data into a higher quality dataset that is more suitable for training, in order to optimize model efficiency and performance.

All leading generative AI models resort to data pre-processing to some extent. For example, Meta created many data filtering

[25] The literature calls this problem "model collapse". This line of research is chiefly concerned with the possibility that developers unwittingly use synthetic data in their training data. This may happen because developers collect data from the Internet which may in turn be generated by AI. See an example of model collapse for LLMs in Shumailov et al. (2024). The Curse of Recursion: Training on Generated Data Makes Models Forget; and Shumailov et al. (2024). AI models collapse when trained on recursively generated data. Available here and here. The issue has also been found in generative image models – see, e.g., Alemohammad et al. (2023). Self-Consuming Generative Models Go MAD. Available here.

[26] See, e.g., Gerstgrasser et al. (2024). Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. Available here.

[27] Europe Terms of Use, by OpenAI. Available here.

[28] Meta Llama 3 Community License Agreement. Available here. This has changed with the release of Llama 3.1, in which case the community license agreement requires models developed using Llama 3.1 to be under the same community license agreement, display "Built with Llama" in the model publication materials and include "Llama" in the beginning of the AI model name. See the Meta Llama 3.1 Community License Agreement. Available here.

[29] See footnote 22.

pipelines for developing Llama 3.[30] Likewise, Google applies filters to training data in their Gemini models[31] and Microsoft has optimised the efficiency of its Phi-3 model by strategically curating its data, among other techniques.[32]

To maximize model performance and efficiency, AI developers must choose an optimal mix of data pre-processing techniques. There are many data pre-processing techniques, including removing low quality data, duplicate information or mixing data from different sources in specific ways (see Box 2).

Box 2 – Examples of data pre-processing techniques[33]

There are many data pre-processing techniques developers may employ in the case of LLMs. This is illustrative of the many decisions developers must make, which can have an impact on the efficiency and performance of the generative AI model. Some examples include the following techniques:

- **Language filtering** filters documents that only include desired languages, including code languages.
- **Heuristics** can be useful in removing large volumes of text that is not useful for training, such as documents with very few words, lines where words are repeated many times or lines with many symbols (e.g., # or -) relative to actual words.
- Developers also filter for **data quality**, aiming to select data similar to datasets they consider to be high-quality.
- If models are trained with a specific domain in mind (e.g., medicine or law), developers may filter for **domain-specific** data, by comparing it with datasets specialised on that domain.
- **Removing duplicated data** and near duplicates is an important step to make models efficient and performant.
- Filtering **toxic and explicit content** removes illegal and extremely undesirable content from the training data and leads models to produce less of that content.
- **Data mixing** assigns weights to each data source to give more or less importance to specific datasets, which has been found to significantly impact model performance.

Optimizing data pre-processing requires heavy experimentation from AI developers, which is computationally expensive, time consuming and requires a significant degree of expertise.[34] Typically, this optimization is based on the intuition and know-how of the individual developers and teams, following a learning-by-doing process. These techniques are

[30] See the blog posts by Meta introducing Llama 3 and 3.1, here and here. See also the paper presenting Llama 3, here, where Meta details many of the experiments it has conducted during the development of Llama 3.
[31] See the Technical Reports of Gemini 1 and Gemini 1.5, by Google, here and here.
[32] See the blog post by Microsoft introducing Phi-3, here.
[33] Examples taken from a survey of data pre-processing techniques in Albalak et al. (2024). A Survey on Data Selection for Language Models. Available here. More examples of data-filtering techniques are available in the paper presenting Llama 3 by Meta, here.
[34] See also section III.1 and III.3 in the AdC's Issues Paper on generative AI, here.

often described as an "art" and remain poorly documented.[35]

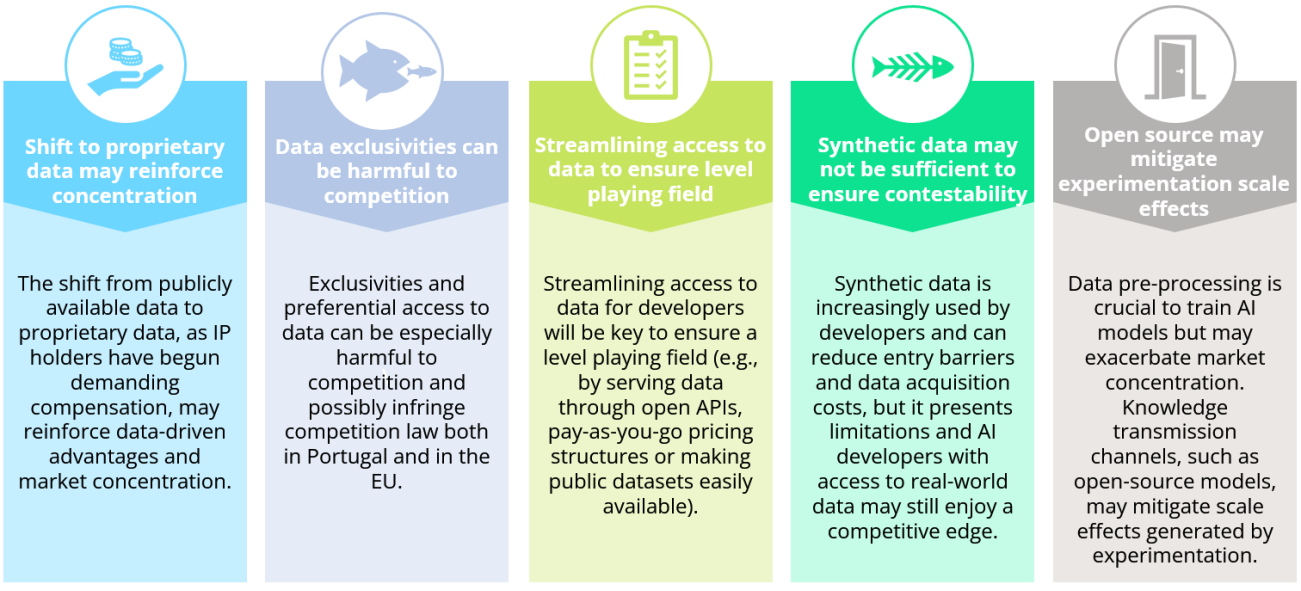
Experimentation is a structural factor of AI development that exacerbates the development costs and scale effects of generative AI models, **making AI markets more prone to concentration**. This is especially true for foundation models, where it may lead to scenarios where very few foundation models become a critical input in downstream markets. Therefore, the need for experimentation in AI development underscores the importance of access to key inputs, such as computing power, in reducing barriers to entry.

Open-source AI models may play a key role in reducing experimentation-driven barriers to entry in the sector. The key issue in experimentation is finding optimal model architectures or hyperparameters – and, in the case of data, the optimal mix of data pre-processing techniques. Once optimal configurations are found, this knowledge can be documented and shared across the industry. Due to its transparency, open-source models and detailed model documentation will be a key venue in reducing the need for experimentation and, therefore, development costs.

> **Open source may mitigate experimentation scale effects**
>
> Data pre-processing is crucial to train AI models but may exacerbate market concentration. Knowledge transmission channels, such as open-source models, may mitigate scale effects generated by experimentation.

COMPETITION AND GENERATIVE AI: ZOOMING IN ON DATA

KEY HIGHLIGHTS



[35] See, e.g., footnote 5.