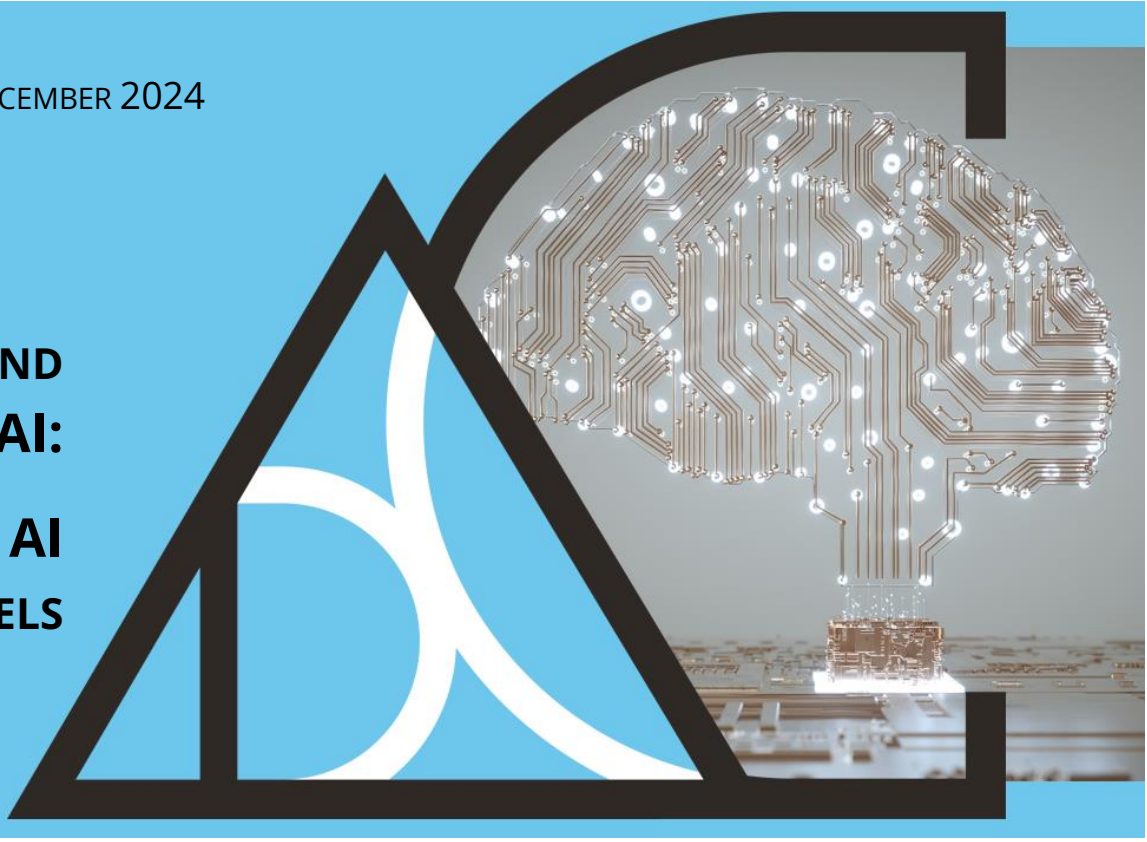
AdC Short Papers I December 2024



Competition and Generative AI: Opening AI Models

The Portuguese Competition Authority (AdC) has been following the generative artificial intelligence (AI) sector since late-2022. The AdC has published an issues paper[1] on AI in November 2023, and a short paper[2] on access to and use of data in generative AI in September 2024.

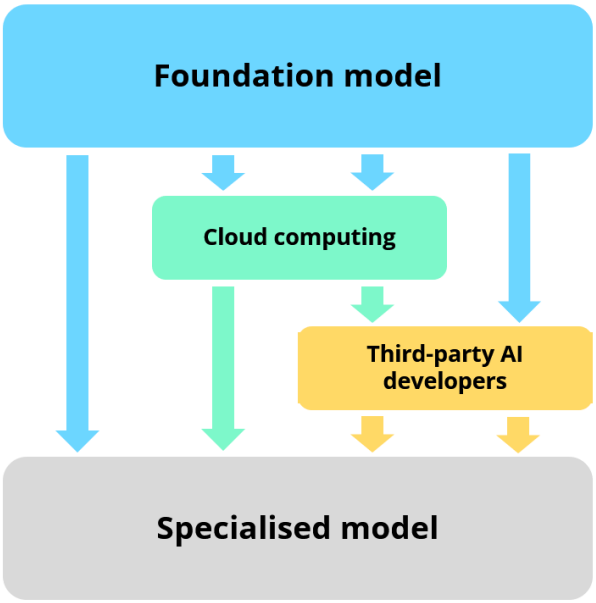
This short paper covers issues related to access to AI models by downstream third-party AI developers (third parties), as well as the role of openness of AI models in bringing in competitive discipline and innovation to generative AI markets.

I. Introduction

Generative AI is a disruptive technology that is able to produce new content – such as text, images, video or audio.

The generative AI sector can be broadly divided into two segments: upstream, where foundation models are developed, and downstream, where specialised models are built from foundation models[3]. Foundation model developers may decide (i) to vertically integrate their foundation models with downstream AI services, (ii) to provide access to third-party developers, (iii) and the level of access they give to their models. Access to foundation models is often granted via cloud services, which may in turn vertically integrate with specialised models, or provide services to third-party developers. The extent to which a foundation model is vertically integrated or made accessible to cloud providers or to third-party developers is the result of a strategic decision made by the foundation model developer (see figure below).

[1] Available here.
[2] Available here.
[3] The frontier between what is a foundation model and a downstream specialised model is not necessarily well-defined. The key characteristic of foundation models is that they are general in nature and may be adapted to many different tasks (e.g., via fine-tuning), producing a specialised model. However, the models that are widely considered to be foundation models, such as the GPT or the Llama families of models, are themselves built and adapted from primary foundation models. These primary models are then adapted to behave in specific ways (e.g., alignment to follow instructions) via fine-tuning, so the resulting models could be regarded as specialised models built from the primary foundation model. Still, they are considered to be foundation models, given their general nature and ability to be adapted to many different tasks.



Markets for foundation models have characteristics that make them prone to high levels of concentration. Developing and implementing foundation models require significant computing power, data and know-how. These inputs may be subject to bottlenecks and entail strong scale effects, which may be exacerbated by experimentation aimed at optimising models. Foundation models are also general in nature – once they are developed, they can be adapted to many different tasks, and the resulting specialised model is often considered better than one model built for that task from scratch. Lastly, AI models may benefit from network effects. By observing user behaviour, collecting user feedback or running live A/B tests on the platform, AI developers may improve future iterations of the models. Given these characteristics, markets may converge to scenarios where few foundation models support many specialised models.[4]

Competition and innovation in specialised models will critically depend on the degree and quality of access to foundation models. Foundation models are highly versatile, as they can be adapted to a wide range of applications. Broader and better access to foundation models expands the pool of developers working on specialised models and promotes innovation, namely by increasing the number of applications to which AI be adapted. Therefore, it contributes towards materialising the potential benefits of AI.

The openness of AI models may also create competitive discipline in AI markets.

The next two sections develop on these ideas to map the key determinants of competition related to access to AI models.

II. AI model openness in Generative AI

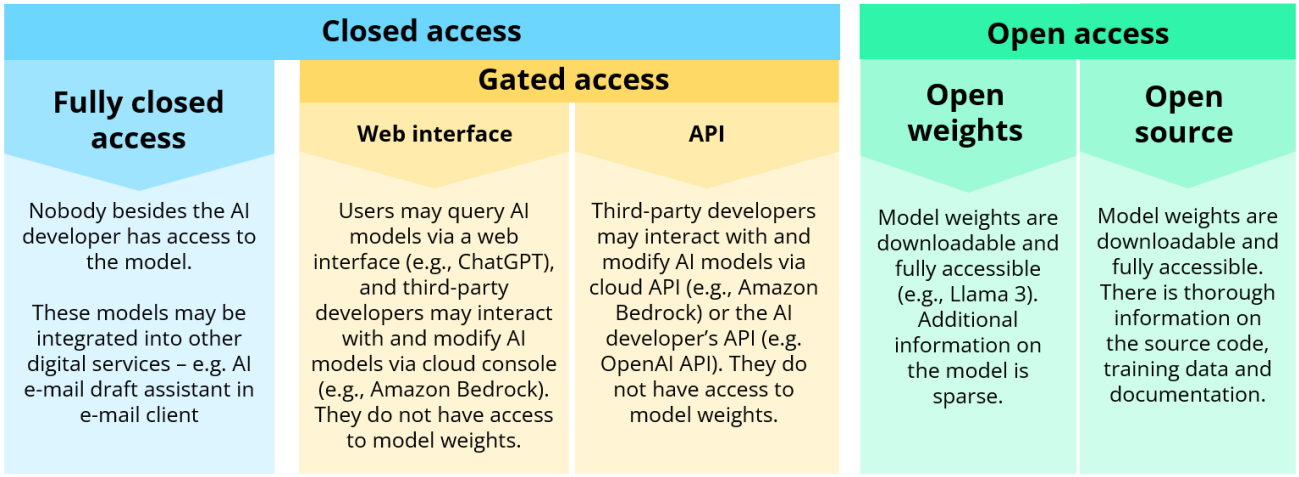
In generative AI, access refers to the extent to which users, third-party developers and researchers can interact and modify AI systems. For instance, a user querying and receiving responses from ChatGPT, via a web interface or an API[5], is accessing the Large Language Models (LLMs) by OpenAI. Similarly, when developers or researchers download model weights from a repository and are able to fine-tune it, they are also accessing the model.

The notion of openness regarding foundation models is not simple nor dichotomous, but rather a matter of degree. **AI models have multiple components which may be accessible to third parties, including model weights, documentation, source code or training data. There is, in this sense, a gradient of levels of access to AI models**, depending on the components which are

[4] See more in section III of the AdC's Issues Paper on Generative AI, available here.
[5] Application programming interface.

accessible to third parties. The gradient ranges from fully closed, gated access and fully open (see figure below)[6]. At one extreme, models are fully closed and only the developer can access any part of the model. This may happen, for example, because the model is integrated into another digital service. On the other hand, in open models, weights are accessible and, in fully open models, all its components are accessible to the public without restriction, including source code and documentation[7].

AI developers control how third parties interact with their models by setting up access points. Access to AI models is typically granted through gated access, such as APIs or web interfaces, which are frequently integrated into cloud services. For example, models by OpenAI, Anthropic, Cohere or Meta are available in Amazon's and Microsoft's cloud services. Gated forms of access often restrict direct visibility into the inner workings of the AI model, to the extent third parties do not have access to model components such as the model weights or architecture. However, third parties may still be able to customise the gated AI model in limited ways. For example, cloud-hosted AI services may allow third-party developers to tweak hyperparameters[8], do fine-tuning, continued pre-training or retrieval-augmented generation (RAG)[9]. Other services may be available for third parties, such as model distillation or synthetic data generation.



[6] This distinction and figure are adapted from Solaiman (2023). The Gradient of Generative AI Release: Methods and Considerations. Available here.

[7] Casper et al. (2024) provide an overview of the many components and access points an AI model may have. One can query the model or have access to sampling probabilities, gradients, model weights, the source code, documentation, hyperparameters, training data, model evaluations, among others. See Casper et al. (2024). Black-Box Access is Insufficient for Rigorous AI Audits. Available here.

[8] For example, in an LLM, third-party developers may have access to sampling probabilities or to the hyperparameters defining how new tokens are sampled (e.g. temperature sampling).

[9] Retrieval-augmented generation (RAG) allows developers to expand the knowledge of the AI model by accessing external sources of information, such as documents or search results. These techniques are used in services such as ChatGPT, Perplexity AI or You.com, namely for grounding. See also the seminal paper Lewis et al. (2021) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Available here.

The industry often simplifies this gradient of access by distinguishing two types of foundation models – open and closed (also referred to as proprietary) – depending on the accessibility to model weights. It should be noted this is an oversimplification and, as highlighted above, AI model openness is not binary, as there is a gradient of levels of access to AI models.

A key feature of open access AI models is that model weights are publicly available. **These models are often called "open" or "open weights"**. **Open source refers to the most open form of access in AI**, where, besides access to model weights, third-party developers have the right to use, study, modify and share AI models, as well as access to detailed information about the model, such as on the training data, model architecture or source code (see Box 1).

Box 1 – Notion of Open Source AI

Multiple concepts for "open source" in the context of AI models have been suggested, varying in scope and specificity. They typically emphasize that model weights must be freely accessible by third parties, along with detailed information on the inner workings of the AI model.

The **EU AI Act**, mandates that, for model distribution licenses to be considered free and open source, they must entail that model weights, information on the model architecture and on the model usage are made publicly available[10]. Open-source models are exempted from the obligations set forth in the AI Act, unless they are integrated into unacceptable or high-risk systems, or they are models with systemic risk.

The **Open Source Initiative (OSI)**[11] has published a definition for open-source AI[12], outlining criteria that must be met for an AI model to be deemed open source. According to OSI, AI developers must:

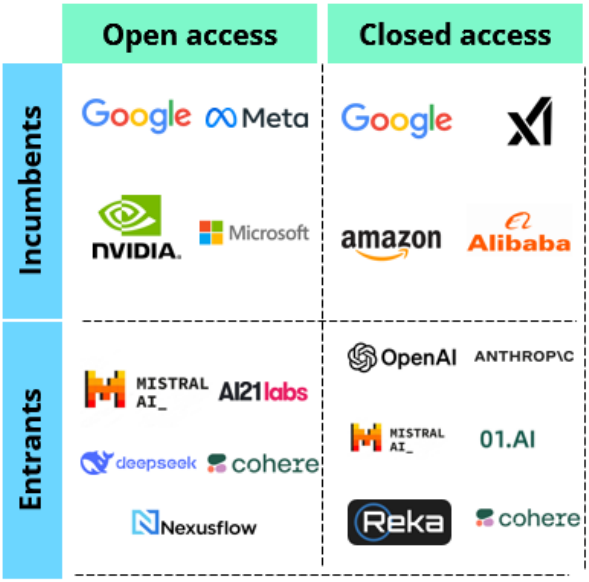
- Grant third-party developers the right to use, study, modify and share the AI model;
- Provide sufficiently detailed information on the data used, enabling other others to recreate the models using comparable or identical data;
- Share the source code used to train and run the AI system, including, for example, the model architecture or data pre-processing techniques; and
- Make model weights publicly available, including intermediate stages of training of the model.

Lastly, **developers** tend to refrain to call their open models "open source", with the exception of Meta, which focuses on access to model weights and model modifiability as the key features of open source[13].

There is currently a relatively large ecosystem of open weights models and developers, creating choice downstream. Hugging Face, for example, currently hosts more than 140 thousand large language models, even if many are based on a few foundation models[14]. In addition, in Chatbot Arena, there are over 30 organizations actively developing open LLMs.

[10] See pp. 102-104 of the AI Act. Available here.
[11] The Open Source Initiative is a standards organization that defines and maintains the most widely recognised criteria for open-source software, known as the "Open Source Definition". This definition establishes specific requirements for software to be classified open source, including the obligation for developers to make source code publicly available and to grant third parties the right to use, modify and redistribute the software freely. See more here and here.
[12] See the "Open Source AI Definition" by the Open Source Initiative here.
[13] The Llama models are widely marketed as open source by Meta. See, for example, the introductions to Llama 3.1 and 3.2, here and here, as well as Meta's statement on open source AI, here.
[14] For example, almost 40 thousand of the LLMs in Hugging Face have either "llama", "gemma", "phi" or "mixtral" in their name.

These include digital incumbents – such as Meta, Microsoft or Google – but also startups – such as Mistral, AI21 Labs, Deepseek or Cohere – see examples of LLMs in figure below. While these developers make model weights available to third parties, their distribution licenses do not necessarily align with the broadest definitions of open source, such as the Open Source Initiative's (see Box 1).



Open foundation models have typically lagged closed models in terms of performance, but there is evidence they are catching up. Open models seem to take up to a year to match the performance of the best closed models (see figure below on Elo scores)[15].

Most of the performant open foundation models, however, are developed by digital incumbents, namely the Llama (Meta), Phi (Microsoft) or Nemotron (Nvidia) models. Mistral AI seems to be the most significant entrant developing performant open foundation models, though with a cloud partnership with Microsoft.[16]

Some AI developers, such as Google or Mistral, launch both open and closed models, but their closed models typically have better performance.

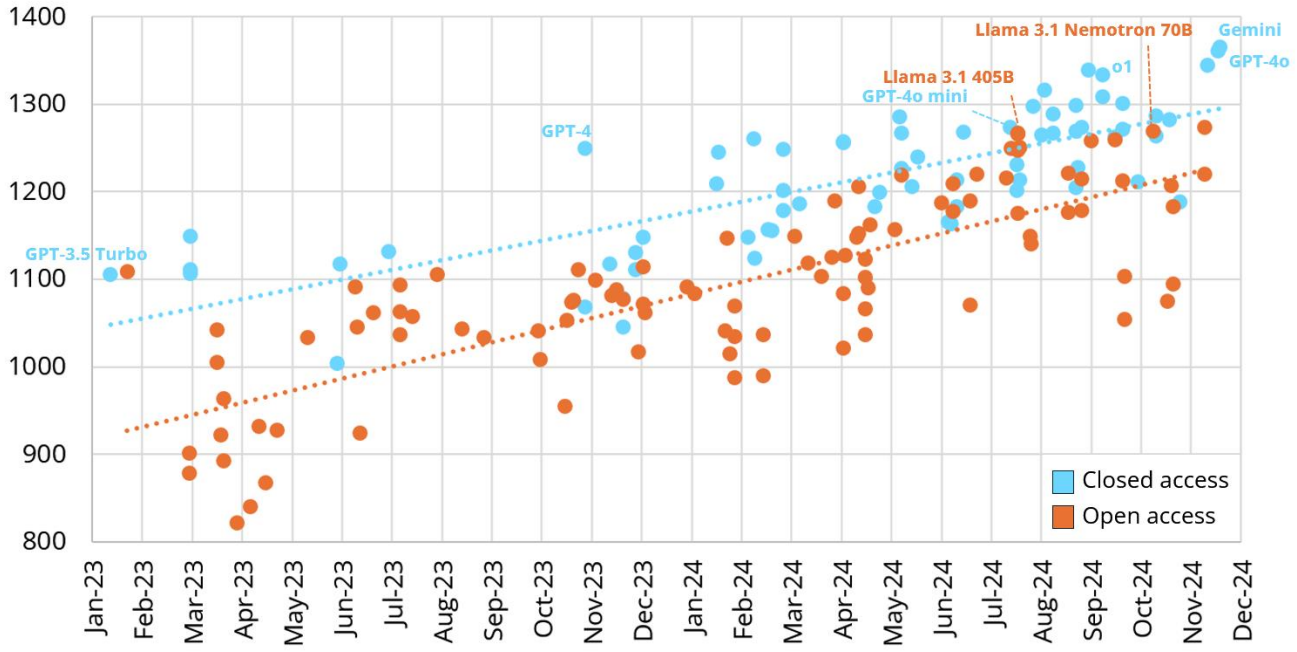
Access to foundation models and ecosystems of open models create choice

There are currently different modes of access to AI models and a relatively large ecosystem of open AI models, some of which well performant. This creates choice and flexibility for downstream AI developers. As such, it is key to ensure access points to AI models are not unnecessarily restricted.

[15] A report by Epoch AI presents similar findings, based on benchmark performance (lag of 5-22 months) and training compute (lag of 15 months), concluding that once-frontier AI capabilities are reached by open models within one year. See "How Far Behind Are Open Models?" by Epoch AI, available here.

[16] In the agreement, Mistral commits to use Azure for cloud computing, Microsoft distributes Mistral's models in Azure, Microsoft invests €15 million in Mistral and both collaborate on R&D. See more in the CMA's decision on the partnership, here.

Chatbot Arena Leaderboard (LLMs) – Elo scores



Source: Chatbot Arena (adapted by the AdC).

Notes: Models are compared by calculating an Elo score for each in the Chatbot Arena, where users may compare two LLMs by sending both the same query. After receiving the responses, users cast a vote for the model they believe provided the better answer, with the identities of the models only being disclosed after the vote is submitted. These votes are then used to calculate Elo scores, a rating system originally used in chess. Data as of 28 November 2024. The x-axis refers to the date of release, based either on the name of the model in Chatbot Arena, publications by the developer or the initial commit in Hugging Face. See the website and the paper presenting the Chatbot Arena, here and here.

III. The contribution of AI model openness to innovation

Promoting openness and a level playing field in access to foundation models is key to ensure competition and innovation in downstream generative AI markets[17], namely if third-party developers and researchers are able to experiment, adapt foundation models to different tasks and create new products.

The degree of access to foundation models can significantly affect the ability of third-party developers to innovate and adapt them to specific use cases. Gated forms of access are purposefully built features by foundation model developers (or the cloud providers hosting the foundation model) to allow third-party developers and researchers to interact with the model in specific ways (e.g., to do fine-tuning). Third parties will have greater flexibility to build custom downstream AI models and systems, and to experiment, if there are more of these features and if they can interact with and modify more components of the foundation model.

The degree of access to foundation models may also have an impact on the quality of the specialised models downstream. For example, third parties may fine-tune a foundation model but, if they are not able to access the inner components of the foundation model or their specialised model, they may have limited ability to experiment different configurations, to

[17] The contribution of AI model openness to innovation is underlined by the Commission in its Competition Brief on generative AI and virtual worlds (available here), based on the responses provided in its call for contributions on generative AI and virtual worlds, available here.

evaluate the specialised models or to customise their specialised models.

In gated forms of access, if third parties wish to use, interact, modify or adapt the foundation model in an unanticipated way, they may not be able to do so, as they do not have direct access to the foundation model.

Open foundation models may offer significant cost advantages relative to closed models and additional flexibility to third-party developers and researchers, facilitating innovation. Whenever a specific use case arises, third parties may adapt the foundation model to their needs on their own terms, provided they have the necessary know-how and access to compute.

The specialised AI models and systems by third parties may be shared with other third parties, fostering the creation of ecosystems of open AI development, which increase the diversity, quality and availability of AI models and systems, as well as the number of use cases[18]. This may be particularly relevant in the case of small AI models and systems[19] that require less compute, as the costs of experimentation, developing, sharing and using them are much lower.

Ecosystems of open AI development may significantly decrease know-how related barriers to the generative AI sector, as they create knowledge transmission channels in the industry. As such, it makes it easier for new and existing professionals in the sector to learn new tools, share knowledge and best practices.

Still, ecosystems of AI development around a specific foundation model may create advantages for the original developer and reinforce concentration, both in the case of open and gated access models. First, there may be network effects from having many third-party developers creating new applications and tools and, in the case of open models, experimenting and providing valuable ideas on how to improve the foundation model[20]. Second, opening foundation models may create a talent pipeline for the original developers, giving them easier access to know-how. Third, open foundation models may set de facto standards in the sector[21], such that the costs of using and developing from the foundation model are lower relative to the competition. This can create disproportionate advantages for the original developer, in terms of integration and accessibility, reinforcing concentration upstream.

---

[18] It has been reported that the leak in March 2023 of the first Llama family of models by Meta has cause a spurt of innovation and created an open-source ecosystem around Llama. See, for example, here.

[19] These refer to AI model and systems that can run locally even on low-powered devices, such as smartphones (e.g., small language models). The ability to run models offline may be a key feature for users wishing to protect proprietary, confidential or other sensitive information.

[20] Open-source libraries are illustrative of this potential benefit. Many of the coding libraries used to build AI models were initially made by major AI developers and then released in open source. As a result, they have been widely adopted and have benefited from the contributions and extensions by third-party developers, such that there is a development ecosystem around these tools. Examples of this include PyTorch, initially developed by Meta, or Tensorflow, initially developed by Google. It must be noted, however, that these libraries are open-source software, meaning that everyone can directly contribute to improve them. This does not happen in most open AI models.

[21] See, for example, Meta's publication "Open Source AI is the Path Forward", here, where Meta expresses the wish to make Llama the industry standard.

In addition, the degree of access to AI models and the existence of ecosystems of open AI may be key in fostering the development of plugins for AI models. These expand on the features and tasks AI models can perform, by connecting with external APIs and datasets.[22]

Access to AI models can boost innovation

The degree of access to AI models can significantly affect the ability of third parties to innovate and adapt them to specific use cases. Open AI models decrease barriers to entry and expansion and increase innovation.

Ecosystems around a specific AI model may reinforce concentration

Ecosystems around a specific AI model can foster experimentation, create standards in the sector and foster the development of plugins for AI. This can create advantages for the original developer, in terms of integration and accessibility, reinforcing concentration.

IV. Possible limitations of AI model openness

Despite its potential, AI model openness may have several limitations that restrict its ability to foster competition and drive innovation in downstream AI markets.

The fact that an AI model is open or open source does not preclude potential anticompetitive practices involving the model. Some of the limitations of AI model openness may be endogenous to the strategic decisions made by players in the generative AI sector. If such strategies are adopted by a dominant undertaking and involve, for example, self-preferencing, discriminatory access, refusal to supply, tying, bundling or lock-in strategies, they may be found to infringe the Portuguese Competition Act and the TFEU.

In gated forms of access, foundation model developers are in a privileged position to enter and expand in downstream generative AI markets. Even if foundation model developers give some degree of access to third parties, they retain full access to their own foundation models. As such, they may have the ability and the incentives to give their own downstream AI models privileged access to their foundation models, hampering the ability of other downstream models to compete.

The integration between cloud computing and AI may also raise competition concerns if cloud providers compete in the markets for specialised AI models. Cloud providers are often key intermediaries between foundation model developers and third parties and may have privileged access to foundation models. For this reason, they may have the ability and the incentive to integrate their cloud services with specialised AI models that compete with the offerings of third parties, while giving them limited access to the foundation models. This may create a scenario akin to the vertical integration between foundation models and downstream AI services and hamper the ability of third parties to compete in downstream AI markets.

[22] For further information on plugins for AI models, see section V.2. of the AdC's Issues Paper on Generative AI, available here.

AI developers may leverage open models to strengthen their market power in adjacent markets, even if they may be freely used, modified and shared by third parties. For example, complementary proprietary services could benefit from a privileged integration with the open AI models, hampering competition. These may include services such as search engines, social networks, operating systems, productivity software, cloud services, digital advertising, among others. Therefore, openness is not itself a guarantee of competition, nor does it preclude strategies that may raise concerns under the competition legal framework.

Openness may be part of a strategy to lock-in users and third-party developers, after which the original AI developer closes off future versions of the model[23]. AI developers may initially reap the benefits of openness by encouraging widespread use of AI, the development of an ecosystem around their AI model and third-party contributions. However, at some point, the AI developer may have the ability and the incentive to fully or partially close future versions of the model. This may create lock-in effects to the extent that the open model is deeply embedded into the products and tools of the third-party developer, there are barriers to data portability, or the ecosystem around the model built many specialised tools and resources.

It is nonetheless relevant to bear in mind that open AI model developers may face challenges in monetising open models, **especially foundation models**, which may reduce their size, scope and quality. Developing foundation models requires substantial financial resources, given the need for computing power, data and know-how. Making foundation models freely available may restrict access to capital and revenue necessary to invest in the development of large and more competitive models vis-à-vis closed models.

AI developers may impose limitations on the uses of their open models by third parties. Even if models are marketed as "open" or "open source", AI developers may prohibit the creation of competing models based on the open model (e.g., third parties cannot generate synthetic data or do transfer learning) or impose restrictions on their large-scale commercial use[24].

The most open AI models may also raise issues of safety. As the inner workings of the models are more transparent, they may be more exposed to adversarial attacks by malicious actors. This openness may, for example, increase the likelihood and frequency that models are "jailbroken", allowing users to bypass safety mechanisms of the models and use the model for harmful purposes, such as producing illegal or otherwise problematic content (e.g., disinformation). Once an open model has been compromised, the AI developer has little control

[23] As highlighted by the US Federal Trade Commission (FTC), in 29.06.2023: "*Experience has also shown how firms can use "open first, closed later" tactics in ways that undermine long-term competition. Firms that initially use open-source to draw business, establish steady streams of data, and accrue scale advantages can later close off their ecosystem to lock-in customers and lock-out competition*." (see here).

[24] For example, the Llama 3.2 community license agreement indicates that services using Llama 3.2 with more than 700 million active users must specifically request a license from Meta, without which Llama 3.2 cannot be used. Likewise, Mistral Large 2 is available as an open model under a research license, which only allows usage and modification of the model for research and non-commercial purposes. Commercial usage requires a commercial license that must be acquired by third parties.

over how it is used, as the model can be easily distributed and modified by others. Nonetheless, it is important to bear in mind the risk of safety claims being strategically (mis)used by firms to prevent access to third parties.

Finally, model openness is not a guarantee of effective access to AI models by third-party developers. First, using foundation models requires access to significant computing power and specialised know-how[25]. Without these inputs, third-party developers cannot use the most performant open models. Second, the quality of access to AI models is also a crucial factor. This encompasses factors such as the speed, rate or stability of access, or the time delay with which access is granted.

Given the opportunities offered by generative AI, it is crucial that the sector develops in a competitive way, to the benefit of consumers. To this end, competition policy and enforcement may play a key role in fostering contestability and preventing bottlenecks from materialising into market entrenchment.

A level playing field in third-party access to foundation models is key for competition in downstream AI services

Foundation model developers and cloud providers may have privileged access to foundation models, and may have the ability and the incentives to hamper the capacity of third parties to compete in downstream AI markets.

AI model openness may be instrumental in leveraging strategies

Openness may be part of a strategy to lock-in users, after which the model is closed. It may also be instrumental to gain market power and leverage it in adjacent markets, such as search engines, social networks, operating systems, productivity software, cloud services or digital advertising.

[25] See more in section III of the AdC's Issues Paper on Generative AI, available here.

Autoridade da Concorrência

COMPETITION AND GENERATIVE AI: OPENING AI MODELS

KEY HIGHLIGHTS

Access to foundation models and ecosystems of open models create choice

There are currently different modes of access to AI models and a relatively large ecosystem of open AI models, some of which well performant. This creates choice and flexibility for downstream AI developers. As such, it is key to ensure access points to AI models are not unnecessarily restricted.

Access to AI models can boost innovation

The degree of access to AI models can significantly affect the ability of third parties to innovate and adapt them to specific use cases. Open AI models decrease barriers to entry and expansion and increase innovation.

Ecosystems around a specific AI model may reinforce concentration

Ecosystems around a specific AI model can foster experimentation, create standards in the sector and foster the development of plugins for AI. This can create advantages for the original developer, in terms of integration and accessibility, reinforcing concentration.

A level playing field in third-party access to foundation models is key for competition in downstream AI services

Foundation model developers and cloud providers may have privileged access to foundation models, and may have the ability and the incentives to hamper the capacity of third parties to compete in downstream AI markets.

AI model openness may be instrumental in leveraging strategies

Openness may be part of a strategy to lock-in users, after which the model is closed. It may also be instrumental to gain market power and leverage it in adjacent markets, such as search engines, social networks, operating systems, productivity software, cloud services or digital advertising.