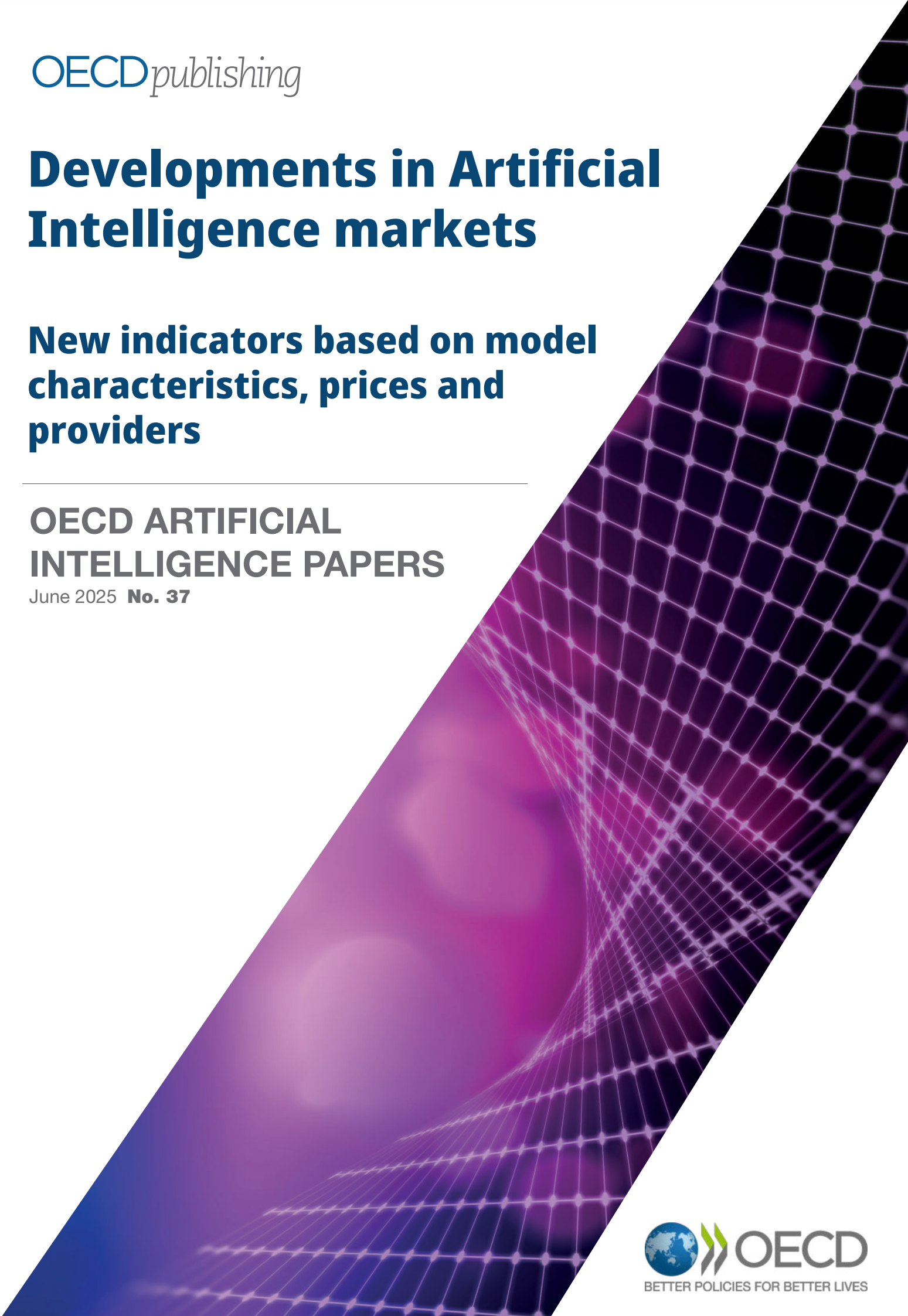
OECD publishing

Developments in Artificial Intelligence markets

New indicators based on model characteristics, prices and providers

OECD ARTIFICIAL INTELLIGENCE PAPERS
June 2025 **No. 37**



OECD
BETTER POLICIES FOR BETTER LIVES

OECD Artificial Intelligence Papers

Developments in Artificial Intelligence markets: New indicators based on model characteristics, prices and providers

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the author(s).

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcome, and may be sent to the OECD Economics Department.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Cover image: © Kjpargeter/Shutterstock.com

Attribution 4.0 International (CC BY 4.0)

This work is made available under the Creative Commons Attribution 4.0 International licence. By using this work, you accept to be bound by the terms of this licence (https://creativecommons.org/licenses/by/4.0/).

Attribution – you must cite the work.

Translations – you must cite the original work, identify changes to the original and add the following text: *In the event of any discrepancy between the original work and the translation, only the text of original work should be considered valid.*

Adaptations – you must cite the original work and add the following text: *This is an adaptation of an original work by the OECD. The opinions expressed and arguments employed in this adaptation should not be reported as representing the official views of the OECD or of its Member countries.*

Third-party material – the licence does not apply to third-party material in the work. If using such material, you are responsible for obtaining permission from the third party and for any claims of infringement.

You must not use the OECD logo, visual identity or cover image without express permission or suggest the OECD endorses your use of the work.

Any dispute arising under this licence shall be settled by arbitration in accordance with the Permanent Court of Arbitration (PCA) Arbitration Rules 2012. The seat of arbitration shall be Paris (France). The number of arbitrators shall be one.

ABSTRACT/ RÉSUMÉ

Developments in Artificial Intelligence markets: New indicators based on model characteristics, prices and providers

Given AI's potential to generate productivity and welfare gains, the paper provides new empirical evidence about AI markets to assess whether potential AI users benefit from favourable market developments regarding prices, quality and variety. It leverages an extensive data collection covering Generative AI model characteristics, including their performance and price, developers, cloud providers, and downstream AI-powered applications globally over the past two years. It finds several trends that are indicative of dynamism for the time being – including declining quality-adjusted prices and a growing number of market players and model offerings – but several risks remain, related to bottlenecks in the key inputs to AI, notably data, computing power and skills.

Keywords: Artificial intelligence (AI); generative AI; AI market developments; economics of AI; AI models, performance and price of AI; AI economic frontier.

JEL codes: D43, L12, L13, L4, L17, L86,O31,033.

******

Développements sur les marchés de l'Intelligence Artificielle : Nouveaux indicateurs basés sur les caractéristiques des modèles, leurs prix et leurs fournisseurs.

Compte tenu du potentiel de l'intelligence artificielle (IA) à générer des gains de productivité et de bien-être, cet article présente de nouveaux éléments empiriques sur les marchés de l'IA afin de déterminer si les utilisateurs potentiels de l'IA bénéficient de développements favorables du marché en termes de prix, de qualité et de variété. Il s'appuie sur une collecte de données exhaustive couvrant les caractéristiques des modèles d'IA générative, y compris leurs performances et leurs prix, les développeurs, les fournisseurs de cloud et les applications en aval alimentées par l'IA à l'échelle mondiale au cours des deux dernières années. Les résultats révèlent plusieurs tendances indicatives d'un dynamisme actuel, notamment une baisse des prix ajustés pour la qualité et une augmentation du nombre d'acteurs et de l'offre de modèles d'IA disponibles sur le marché. Cependant, plusieurs risques persistent, liés aux goulots d'étranglement dans les intrants clés de l'IA, notamment les données, la puissance de calcul et les compétences.

Mots clés : Intelligence artificielle (IA) ; IA générative ; développements sur les marchés de l'IA ; économie de l'IA ; modèles d'IA, performance et prix de l'IA ; frontière économique de l'IA.

Codes JEL : D43, L12, L13, L4, L17, L86,O31,033.

Table of contents

Developments in Artificial Intelligence markets: New indicators based on model characteristics, prices and providers 6
1. Introduction 6
2. A framework and data sources for assessing AI market developments 9
2.1. A simple value chain of Generative AI and potential bottlenecks in AI inputs 9
2.2. Indicators to track developments on AI markets 18
2.3. Data collection and measurement 20
3. Results 22
3.1. AI development 22
3.2. AI provision through the cloud 34
3.3. AI-powered digital services and popularity across sectors 39
4. Concluding discussion on risks, policies and future analysis 42

References 44

Annex A. Glossary 51

Annex B. Definition of AI models and key concepts 55

Annex C. Indicators to measure competition 61
Environment: 61
Definition of AI models 61
Definition of the AI Economic Frontier 62
Definition of AI market segments 62
A simple model of AI revenues in a competition environment 63
Number of AI users 63
Competition components 64
AI Market segments 64
Utilisation of AI 65
Intensity of utilisation 65
Usage of AI 65
Revenues from AI for each market segment 65
Revenues from AI for each company 65
Revenue for each company 65
Market shares for each company per modality 66
Calibrations of demand scenarios 66
Calibrations of competition scenarios 66

Annex D. Data sources 68

Annex E. Sensitivity analysis of key results 71
Details of the AI Economic Frontier 71
Details on the AI price index 73
Alternative method to estimate quality adjusted prices: Hedonic regression method 74
Market share under alternative scenarios and calibrations 75
Alternative method to simulate market shares using downloads of Open source models 77

Annex F. A simple model of the price of AI inference 80

Tables

Table 1. The majority of AI models come from the US, but less so than a year ago 27
Table 2. Testing for the pricing power of major AI providers and developers 38
Table 3. Subscription prices of selected popular AI-powered services 41

Table B.1. Definition, concepts and measurement in AI models 55
Table B.2. Contribution of Open source to the AI value chain 57
Table C.1. Assumptions on the determinants of market structure in AI 61
Table C.2. Calibration of scenarios 67
Table C.3. Key indicators for monitoring competition in AI markets 67
Table D.1. Data sources on AI prices 68
Table D.2. AI developers and providers by country 69

Figures

Figure 1. A simple value chain of AI 10
Figure 2. A comprehensive data collection of AI offer from the cloud 21
Figure 3. The exponential rise of the supply of AI 22
Figure 4. The *AI Economic Frontier*: increasing model performance at much lower prices 23
Figure 5. The *AI Economic Frontier* is oligopolistic with more and more players and models 25
Figure 6. Several followers are lagging behind the AI Economic Frontier by only a few months 26
Figure 7. Apparent US leadership in AI development persists but has been challenged over the past two years 28
Figure 8. The leading AI developers appear strongly challenged by large tech incumbents and by smaller players 30
Figure 9. Quality-adjusted AI prices have been falling rapidly for all modalities 31
Figure 10. The churning rate of models at the AI economic frontier is strong 32
Figure 11. US leadership is becoming less clear when jointly considering all modalities, especially when switching costs are low 34
Figure 12. A few AI developers are particularly popular among cloud providers 35
Figure 13. Cloud providers increasingly serve models from several developers 36
Figure 14. Several cloud providers offer the most capable models 36
Figure 15. The growing number of AI-powered services 39
Figure 16. There is diverse supply of AI services across several sectors 40

Figure B.1. From specialized AI models to AI agents 56
Figure B.2. Popularity of open-source AI models per license family 56
Figure B.3. Correlation across AI quality indicators 58
Figure B.4. Evolution of AI model Performance on industry benchmarks 58
Figure B.5. Price of selected AI models 59
Figure B.6. The share of multimodal models among all Text-to-Text models 59
Figure B.7. The share of open-weight models among all models 60
Figure D.1. Evolution of the number of cloud providers 68
Figure E.1. Alternative specification of the AI Economic Frontier, Text-to-Text 71

Figure E.2. AI Economic Frontier, Text-to-Image 71
Figure E.3. AI Economic Frontier, Audio-to-Text 72
Figure E.4. Evolution of the slope of the AI Economic frontier 72
Figure E.5. Quality-adjusted price index Text-to-Text, alternative price variable 73
Figure E.6. Quality adjusted price index by AI model segment and modality 73
Figure E.7. AI price index - hedonic methodology 74
Figure E.8. Simulated market shares by region, baseline scenario under low switching costs 75
Figure E.9. Simulated market shares by company status, baseline demand scenario under high switching costs 75
Figure E.10. Simulated market shares by company status, AGI demand scenario under high switching costs 76
Figure E.11. Simulated market shares by company status, AGI demand scenario under low switching costs 76
Figure E.12. Simulated market shares by company status, Edge demand scenario under high switching costs 77
Figure E.13. Simulated market shares by company status, Edge demand scenario under low switching costs 77
Figure E.14. The rapid rise of open-source enables AI local deployment 78
Figure E.15. Evolution of the market shares of AI models for local deployment 79

Boxes

Box 1. The role of open source in AI markets 15

Developments in Artificial Intelligence markets: New indicators based on model characteristics, prices and providers [1]

Christophe André, Manuel Bétin, Peter Gal and Paul Peltier

1. Introduction

Rapid advances in artificial intelligence (AI) have fuelled hopes of productivity and well-being gains, as well as fears of job losses and concerns about privacy, disinformation, and loss of human control (Filippucci et al., 2024; Ben-Ishai et al., 2024). Generative AI's ability to create new content (e.g. text, image, audio, video and software code) offers transformative potential across a wide range of sectors, including IT, entertainment, education, healthcare and scientific research among others (Lorenz, Perset and Berryhill, 2023; Stanford University, 2024). It is often considered as a new general-purpose technology, on par with previous major innovations in the digital realm, such as computers or the internet (Filippucci, Gal and Schief, 2024).

However, broad-based adoption of highly capable AI models that can assist with a wide range of tasks is a critical pre-requisite of macroeconomic and welfare gains. Widespread AI adoption depends on competitive markets that drive lower prices and more capable, better-quality AI models. The concentration of AI capabilities and data in a few firms could hamper innovation and competition in AI development and deployment (OECD, 2024; Coeuré, 2024; Korinek and Vipra, 2025; Business at OECD, 2024) and thus limit AI's broader economic and societal benefits. AI threatens to further entrench existing leading positions in digital markets (EC-CMA-DOJ-FTC, 2024; Kowalski, Volpin and Zombori, 2024), which has created a challenge for competition policy over the past decade.[2] In the more distant future, the possible arrival of

[1] Disclaimer: *The results presented in this paper are of analytical nature and preliminary, and do not intend to take a position as to any ongoing or future formal competition assessments or enforcement actions carried out by competition authorities.*

Corresponding authors: Christophe André (Christophe.Andre@oecd.org), Manuel Bétin (Manuel.Betin@oecd.org), Peter Gal (Peter.Gal@oecd.org), and Paul Peltier (Paul.Peltier@oecd.org), all from the OECD Economics Department. The authors thank Åsa Johansson, Alvaro Pereira, Alain de Serres, and Filiz Unsal (all from the OECD Economics Department) for their guidance and Sebastian Barnes, Filippo Maria D'Arcangelo, Dennis Dlugosch, Sean Dougherty, Vincent Koen, Tomasz Kozluk, Ruben Maximiano, Cristiana Vitale, Sebastien Turban (all from the OECD Economics Department), Richard May (OECD Directorate for Financial and Enterprise Affairs), Luis Aranda, Sarah Berube and Lucia Russo (OECD Directorate for Science, Technology and Innovation), as well as delegates to the Working Party 1 (WP1) of the OECD Economic Policy Committee (EPC) and the Directorate-General for Competition of the European Commission for comments and suggestions. The authors also thank Sarah Michelson-Sarfati for excellent editorial support.

[2] Indeed, Calligaris et al. (2024) document that mark-ups have been rising faster in digital intensive industries. On the role of competition policy in digital markets, see Nicoletti, Vitale and Abate (2023), Jullien and Sand-Zantman (2021), CEPR (2023) and Schrepel and Pentland (2024) for a more specific focus on AI.

Artificial General Intelligence (AGI), which could match human-level performance across nearly all cognitive tasks, raises concerns about further concentration in demand and even monopoly power if AGI is developed and controlled by a single player (Korinek and Vipra, 2024).

The literature on competition in AI markets tends to emphasise three main risks (OECD, 2024). First, firms controlling access to critical inputs for foundation model development may limit competition through vertical integration (Chardon-Boucaud, Dozias and Gallezot, 2025). Digital incumbents currently enjoy an advantage in access to the main ingredients of AI systems, such as computing capacity (Cottier et al., 2024; Vipra and Myers West, 2023), data (Azoulay, Krieger and Nagaraj, 2024; Bergemann and Bonatti, 2024; Cottier, Besiroglu and Owen, 2023) and talent, and benefit from network effects and positive adoption externalities for AI thanks to their large user base in traditional digital services (Gans, 2024; Bajari et al., 2019).[3] Second, powerful incumbents could exploit their positions in consumer or business-oriented markets to distort choice in foundation model services and restrict competition in foundation model deployment by leveraging gatekeeping positions (vertical foreclosure or product bundling).[4] This raises concerns about potential consumer lock-in, which could lead to higher prices and slower innovation in the medium term (CMA, 2024; CEA, 2024).[5] Third, partnerships involving key players could reinforce or extend existing positions of market power through the value chain (CMA, 2024; OECD, 2024; FTC, 2025).

Existing research on these issues mostly focuses on AI *input* bottlenecks and remains generally qualitative, but empirical evidence on the *outputs* and *outcomes* of AI markets is still scarce. This is partly due to the rapid evolution of the market with the appearance of Generative AI And difficulties in accessing appropriate data. This paper addresses this gap through a comprehensive data collection on AI markets over the past two years – approximately since the launch of ChatGPT by OpenAI – in three segments of the AI value chain: model development, model provision through the cloud and downstream applications. It draws on data from cloud providers as well as secondary sources, including open-source platforms and data partnerships with market players that track the releases and use cases of AI products.[6] The focus of the analysis is on companies that sell AI services on the market or release AI models for use by other companies, in three AI modalities: text (Large Language Models, LLMs), image generation, and audio transcription.[7] The paper considers a broad set of indicators and concepts to capture the complexities of measuring competition in digital markets in general (OECD, 2022a, b) as well as in AI in particular.

First, the paper presents new evidence on the evolution of supply in AI markets, focusing on the number of market players (AI developers and cloud service providers) and on the number and diversity of foundation models as well as AI-powered digital applications. We find that the market appears dynamic,

[3] See also Autorité de la concurrence (2024).

[4] For example large digital companies capture a large segment of the cloud market, which plays a critical role in AI systems, by providing compute and data storage capacity (Carugati, 2023; Pilz and Heim, 2023).

[5] In markets where innovation is the key determinant of market dominance, the relationship between innovation and competition may be ambiguous (OECD, 2023a; Calvano and Polo, 2021; Aghion et al., 2005).

[6] Data sources cover publicly available data from cloud providers' websites, the OECD.AI observatory (2024), EpochAI (Sevilla and Roldán, 2024), Center for Research on Foundation models (Stanford University, 2024), Arena Elo (Hugginface), *Microsoft Azure AI studio*, *Ollama* and data partnerships with *There is An AI for That* (TAAFT, 2024) and *Artificial Analysis* (2024). See more details in Section 2.3.

[7] These models are all built on the current technological paradigm of foundation models based on the Transformer (for text and audio) and Diffusion (for images) architectures (Bommasani et al., 2021). See more details in Section 2.1.

with a steady influx of new entrants across all segments, leading to more than 1,000 foundation models currently available from nearly 100 different developers across different AI modalities.[8] [9]

Second, by simultaneously tracking the quality and price of AI models, we construct an *AI Economic Frontier*, capturing the most competitive AI models – those offering the best performance at a given price. We find that around ten leading companies make it to the Economic Frontier, including both AI-specialised startups such as OpenAI, Anthropic, Mistral or, more recently, DeepSeek, and large digital incumbents such as Google or Meta. Given that the Frontier continuously shifts towards higher quality models at similar or declining prices, we document a significant, 80% decline in *quality-adjusted* AI model prices over the past two years. The emergence of smaller, more efficient models, often from new developers, and the increasing availability of open-source models have contributed to this tendency, combined with technological innovations leading to falling hardware prices and improvements in model building.[10] Despite the speed of innovation and the leading position of the main AI developers at the frontier, other AI developers from various countries (digital incumbents and startups alike) remain within one year or less from the frontier.

Third, to estimate the evolution of AI model market shares, given the lack of publicly available reliable data, we rely on simulations by combining information on supply with assumptions on demand and market characteristics (e.g. the importance of switching costs across models). Many key AI developers are not publicly listed or do not report AI-specific revenues. Hence the calculations are illustrative and are shown under various scenarios that capture different assumptions about demand and market characteristics – in terms of user preferences and the importance of switching costs across models.[11] These simulated markets shares suggest that the markets of AI models are dynamic for the time being under most scenarios – in particular in image- and audio-focused models.

Fourth, we find that the large cloud providers (hyper-scalers: Amazon, Google and Microsoft) have been offering models at similar prices to other cloud providers, as suggested by a series of regressions that control for AI model and provider characteristics. We also document a rapidly rising number of models offered by most cloud providers; and *vice versa*, we show that most developers' models are accessible through several cloud providers.

Overall, these results are consistent with several technological features of Generative AI markets that contribute to their current dynamism. These include the apparent relative substitutability across models

[8] Further downstream among AI powered service apps, we identify more than 12 000 applications and a rapidly rising trend, thanks to a new platform on AI applications called *There's An AI For That* (TAAFT, 2024).

[9] We do not observe, however, the demand side of this market, i.e. the actual use of these models by potential AI adopters.

[10] Strategic pricing considerations by large incumbents (exclusionary pricing) could have also played a role in falling prices, which would be indicative of competition issues. However, the fact that model quality has been rising in parallel – a sign of technical innovation –, and that new smaller players have contributed to these pro-competitive tendencies suggests that such strategic pricing considerations are unlikely to be the sole explanation. Also, while the ability to lower prices may seem surprising given the increasing cost of model development, it could be explained by an increasing use of models, hence larger total revenues despite lower prices – which helps to cover a larger part of development costs. In addition, even though the costs of the largest, top models kept rising, there have been trends towards smaller, more cost-effective models with lower development and inference costs (Martens, 2025).

[11] Other sources rely mostly on the *inputs* to AI systems – financial or physical investments into AI firms as well as imputations of revenue-based market power in the digital sector – as proxies for market shares (OECD.AI, 2024; Korinek and Vipra, 2025). We instead focus on the output and compute simulated shares of revenues from AI under different assumptions about demand and market characteristics. For overviews and discussions of measuring the degree of competition through concentration metrics, see Calligaris et al. (2024), Syverson (2019) and OECD (2022a).

and providers from a technical perspective in several contexts of AI use[12], combined with an active open-source ecosystem (Hagiu and Wright, 2023; Gans, 2024). In addition, the standardisation of industry benchmarks for evaluating and comparing model capabilities and performance further promotes transparency, allowing AI-using companies to more easily adopt solutions that best fit their needs. Alongside the oligopolistic business model of vertically integrated conglomerates (covering model development, cloud services and consumer end-products), there exists an active ecosystem with specialised companies operating only in selected segments of AI markets. While leading platforms have the power to leverage their existing user base, no "*killer apps*" – which would successfully attract and lock-in the majority of users – appear to have emerged yet. The findings of the paper are in line with Hagiu and Wright (2025), Korinek and Vipra (2025) and Martens (2025), who suggest that while there are risks of market dominance by large integrated players and hardware producers, thus far they do not appear to have materialised.

However, many of the developments in AI are still uncertain and prone to rapid evolution, similarly to what was observed around the early days of the internet. This paper provides initial empirical evidence about the evolution of the market over the past two years, without precluding the possibility that the situation changes. It proposes a measurement framework and a set of key indicators to follow going forward. A continuous update and monitoring of these indicators could help to assess in a timely manner whether there are signs that risks to competition materialise. However, several characteristics of the market are left out from the paper, primarily due to limited access to relevant data or other information. As such, this analysis does not substitute for detailed competition assessment.

The paper is organised as follows. Section 2 presents the framework including key concepts about AI markets and concerns to competition, measurement and data. Section 3 presents empirical evidence on market developments in three segments of the AI value chain: Section 3.1 focuses on the market for the development of AI foundation models, Section 3.2 assesses the provision of AI from the cloud, and Section 3.3 discusses developments in the downstream segment describing the offer for AI-powered services and discusses the risks posed by existing digital incumbents. The concluding section discusses risks to competition, the potential role for policies and next steps of the analysis.

2. A framework and data sources for assessing AI market developments

2.1. A simple value chain of Generative AI and potential bottlenecks in AI inputs

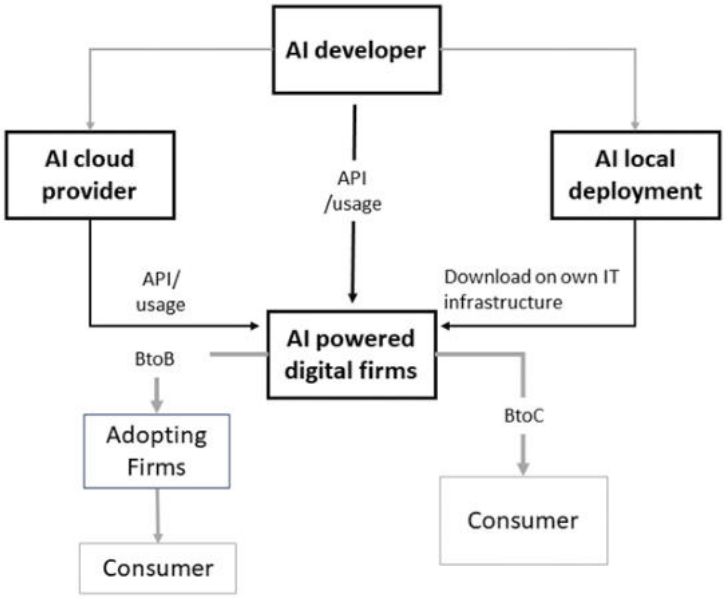
AI systems are "a machine-based system that for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment" (OECD, 2023b). For a brief overview of further basic terms, see the glossary in Annex A.

Recent Generative AI models are a subset of all AI systems that create text, media and software code. They are provided by a complex web of market layers, constituting the AI value chain or "AI stack" (Figure 1). This paper considers three layers. First, most upstream, *AI developer* companies refer to firms or labs that train and finetune Generative AI models, which are then sold (licenced) or rented (per usage) on the

[12] This is not the case for all applications, for instance, AI-powered software bundled with digital devices, such as voice assistants and operating systems. However, Hagiu and Wright (2025) highlight that competition concerns for core AI services at this point are not specific to the AI technologies but rather related to traditional concerns that are already on the radar of competition authorities, such as leveraging model power to adjacent markets through exclusive deals, data-driven acquisitions of unique or proprietary data, hyperscalers transferring their dominance from traditional cloud (e.g., NVIDIA leveraging the network effect of CUDA), and strategic hiring of key workers.

market for the use by other companies or individual developers. Developers include AI labs from digital technology incumbents like Meta, Google and Alibaba and specialised AI companies such as OpenAI (USA), Anthropic (USA), X.AI (USA), MistralAI (FRA), DeepSeek (CHN), Cohere (CAN), AI21 (ISR), BlackforestLabs (DEU), StabilityAI (GBR), or RecraftAI (GBR), some of which having partnerships with incumbents.

Figure 1. A simple value chain of AI



Note: "AI-adopting firms" are companies that integrate artificial intelligence into their workflows. "AI-powered digital firms" are companies specifically built around AI technologies. "AI cloud providers" are companies that host and provide access to one or more model endpoints (hosted instances of a model) via an API (Application Programming Interface). "AI local deployment" refers to the installation and execution of AI models directly on the AI-adopting company's computing infrastructure rather than through cloud services. "AI Developing companies" are companies that train and finetune AI models that are sold on a market for the use of other companies. See the glossary in Annex A.
Source: Authors' elaboration, inspired by (CMA, 2024) and (OECD, 2024).

Certain types of AI are not covered in this paper. Notably, previous generations of AI models (e.g. non-Generative AI) are typically based on traditional machine learning algorithms that are different from those used by foundational models, tend to be internally developed and used – sometimes by firms outside the digital sector. They lack a market and thus fall outside of our scope. Other models that are not publicly available on the market (such as AI models from Apple) or are very narrowly specialised (such as AlphaFold for predicting protein folding) or are only on preview or in "waiting list" (such as several AI video models), are also not considered.

In the second layer, further downstream in the AI value chain, we consider "*AI cloud providers*" (Figure 1). Access to large modern AI models typically occurs through cloud services as Models-as-a-Service (La Malfa et al., 2024; FTC, 2025; Bergemann, Bonatti and Smolin, 2025) or requires users to possess a high-performance IT infrastructure for "*AI local deployment*". [13] AI cloud providers are companies that offer computing services specific to AI and a platform to access models through an Application Programming Interface, API. There is some overlap between providers and developers, since many developers are also AI cloud providers (e.g. Google, Amazon, Microsoft but also OpenAI, Anthropic, Deepseek or Mistral that

[13] Yet another alternative is renting specialised hardware from cloud providers and directly hosting open-sourced or licenced models.

also offer their models on their own platform, through what is often referred to as *first-party APIs*). However, some developers do not provide models directly via an API (such as Meta), and some cloud providers do not develop foundation models but provide access to models from other companies (e.g. Perplexity, TogetherAI, Nebius or Infomaniak). Local deployment, on the other hand, is an alternative way to access AI models – mostly open-source variants – and may be preferred to avoid dependence on cloud providers due to data privacy concerns, or to exert more control over the model. However, this often comes with high investment needs in terms of upgrading the IT infrastructure and skills.[14] Potential competition concerns even more upstream, around the production of hardware (e.g. specific AI chips) are not discussed in the paper (see Korinek and Vipra (2025).

Most downstream in our analysis, "*AI-powered digital firms*" provide AI services to other firms and consumers directly (third layer in Figure 1). These firms usually wrap AI foundation models into user-friendly interfaces (e.g. chatbot interfaces) and by finetuning general-purpose foundation models for more specific tasks, using context- or business-specific data.

To a varying degree, all segments of the value chain combine three key inputs – compute, data and skills – to develop, provide or integrate AI models into digital applications. These inputs are often the focus of competitive assessments given that access to them may be restricted and, hence, may constitute "bottlenecks" to competition. Indeed, AI requires large amounts of computing capacity ("compute") both for training and, increasingly so for recent advanced "reasoning" models, the usage of the models ("inference" phase) (Ho et al., 2024; Erdil, 2024; Sevilla and Roldán, 2024; Vipra and Myers West, 2023). These are concentrated among a few AI labs and big tech companies (Cottier, Besiroglu and Owen, 2023), which are, in turn, highly dependent on very few hardware producers. The development of these technologies requires high upfront fixed costs, some of which may also be considered sunk costs (e.g. electricity used during model training). To the extent that only a few incumbents are able to afford the build-up of the necessary infrastructure to train and serve large AI models, these technical characteristics imply rising risks of natural monopoly.

Data concentration is another potential concern for competition (OECD, 2024). It is an input at every stage of the AI lifecycle. In digital markets, data give incumbent leaders a competitive advantage through data-driven economies of scale and scope due to network effects and data feedback loops. Data feedback loops – where more users improve the product, which in turn attracts more users – can operate through different channels in AI. *Customer data* feedback loops among AI-powered digital applications are common to non-GenAI digital products and are based on tailoring the final product to the preferences of the user using statistical predictions calibrated on the entire user base (e.g. pricing and advertising models used by online retailers). AI-*training data* feedback loop is the collection of data through user-AI interactions, and then relying on them to train or finetune future versions of the models. *Synthetic data* feedback loop refers to the idea that companies producing the best AI model today can generate better AI-produced data that is later used to increase the size of the database for training the model. Finally, an additional potential source for increasing returns to AI is the *intelligence* feedback loop (Korinek and Vipra, 2025) where AI labs generate high-quality code and software that help to improve the efficiency of the company itself. If such

[14] Local deployment includes open and close weight models, but no public information is available for close models. The possibility of locally deploying AI is a critical pro-competitive force that also has a central role in business adoption as it gives more control to AI-adopting firms over the technology and allows for greater specialization and lower dependence from developers and cloud providers. AI for local deployment is not covered specifically in the core of the paper but many dimensions are addressed in Annex E using data on open weights models and showing the trends of local adoption, the dynamism of supply of open-source models and their relative popularity. Overall, the trends and results found for AI from the cloud match those observed on the subset of open weight models and are considered as a robustness check for the core results (see Box 1 and Table B2 in Annex B for more details on the contribution of open source to AI markets).

data feedback loops indeed materialise, that could lead to a winner-takes-all phenomenon or first-mover advantage, both detrimental to competition.

Skills, in particular specialised expertise to develop and operate advanced AI models, are also a key input in AI markets. R&D staff costs are estimated to account for between 29% and 49% of total amortised model development costs (Cottier et al., 2024). Given the scarcity of talents and their complementarity with access to data and computing resources, AI research is highly concentrated within a few leading US tech firms (Cottier, Besiroglu and Owen, 2023). The risk that innovative AI startups fail to hire and retain such talents to the advantage of digital incumbents is an additional risk for long-term competition.

In our framework that focuses on the *outcomes* of AI markets (see next subsections), we do not assess of the importance of these input bottlenecks and their evolution over time. Nonetheless, existing literature suggests that some of these input-related bottlenecks may currently be less binding than initially feared:

- While compute remains a central bottleneck in AI (You and Owen, 2024; Frymire and Owen, 2025), the results achieved by newcomers like DeepSeek (China) and MistralAI (France) show that software improvements, rigorous data curation, parsimonious use of compute and declining cost of compute currently allows the building of high-performance models at lower training costs despite limitations in access to the most advanced hardware (Erdil, 2024; Rahman, 2024).
- Furthermore, several sources suggest that data feedback loop-driven advantages do not appear to be the main drivers of AI model quality improvements so far. Even when benefits from feedback loops exist, they can usually be obtained with relatively limited data and, in most cases, with replicable and publicly available data (Hagiu and Wright, 2025). At this stage, only the customer data feedback loop has been robustly documented, in particular in search (Klein et al., 2023; Schaefer and Sapi, 2023; Allcott et al., 2024), while other sources of *AI-specific* data feedback loops remain uncertain (Korinek and Vipra, 2025; Gans, 2024). Following this argument, Hagiu and Wright (2025) argue that the analogy with data feedback loops in search should be nuanced and conclude that even if data feedback loops exist, they do not always provide a lasting advantage.[15]
- Regarding the concentration of talent, incumbents may not be able to capture talent as much as previously thought (The Economist, 2024), even if risks of talent retention via non-compete agreements by large incumbents cannot be excluded. The large number of companies founded by former US tech employees (in the US and in countries in Europe and Asia) and the example of companies like DeepSeek's, which is reported to rely mostly on engineers trained in China, suggests that talent may currently be less concentrated in large incumbents than initially feared. While the US remains a major pool of potential AI talent, along with China and India (CEA, 2025)[16], other regions including Europe, also benefit from highly trained AI engineers (Macropolo, 2023) – even though attracting those talents is very costly and usually possible

[15] Although data for training and finetuning is highly concentrated, it may not always be a critical bottleneck and depends on specificities along three dimensions: first, data substitutability, which captures the extent to which the relevant information in the data can be obtained at low cost from alternative data sources (for instance, in the training of a language model, accessing proprietary sources may add little value compared to using already publicly available online data sources); second, data complementarity, which captures the degree to which a greater variety of information generates better predictions (for instance, linking different types of information about a person from different datasets); and third, data returns to scale, which captures the extent to which more information generates better predictions in general (Hagiu and Wright, 2023; Hagiu and Wright; 2025).

[16] Measured by Science and Engineering degrees, as well as PhD's specialised on AI. In the latter metric, the US has a clear lead according to CEA (2025).

only for digital incumbents or a few very well-funded start-ups (Cottier, Besiroglu and Owen, 2023).

2.1.1. Model characteristics and market segments

Generative AI models that are created by AI developers (first layer in Figure 1) can be characterised along four major dimensions (see more details in Annex B):

- The *Performance* (or quality) of an AI model indicates its capability, as measured by an index based on the success rates of giving correct responses in benchmark tests (scores on standardised tests for reasoning, knowledge or task completion). Our AI quality index *q* is a composite index of key benchmarks assessing different aspects of model performance, following the practice in the AI industry (Artificial Analysis, 2024).[17] It represents the success rate of the AI model in completing a representative cognitive task. In turn, the failure rate (1-q) captures the probability that the AI gives a wrong answer (including so-called hallucinations where the model is unaware of its mistake) or is unable to carry out the task.[18] Model performance and other dimensions of model capability (number of modalities, the maximum length of input and output, etc.) are usually an increasing function of the size of the model, the quantity and quality of the training data, and the amount of computing for pretraining (usually seen as an implication of the so-called "scaling law"; see Hoffmann et al., 2022; Kaplan et al., 2020).
- The *Modality* of an AI model defines how the user interacts (input to the model) with the AI and what type of results (model output) the model returns to the user. Single-modality models are pure *Text-to-Text* language models (Large Language Models, LLMs), *Text-to-Image* image generators or *Audio-to-Text* language transcription models. In addition, there are Multimodal models that can perceive the environment via voice, text or vision, and they are increasingly common. Moreover, various models be integrated into *AI Agents*, which are digital or physical

[17] See the following formula, weighting 4 different benchmark indices capturing different performance aspects of AI models: $q = 0.35 * q^{MMLU} + 0.35 * q^{ELO} + 0.15 * q^{GPQA} + 0.15 * q^{livebench}$

where the MMLU stands for the *Massive Multitask Language Understanding* benchmark which evaluates models on a wide range of academic subjects, focusing on comprehension and reasoning tasks of undergraduate complexity; ELO stands for the *Arena Elo* benchmark which is a relative measure that involves community voting to determine the best model responses to a set of questions; GPQA stands for *Graduate-Level Google-Proof Q&A*, which, similarly to the Livebench index, provide further evaluations of model performance on complex and advanced tasks with particular attention to avoid dataset contamination (models being trained on the questions and answers of the benchmark). See Annex A for a presentation of key benchmarks and Annex B for more details, in particular Figure B.3 that shows high pairwise correlations across them as well as Figure B.4 that show the quality improvements over time on the main benchmarks.

[18] A broader concept of quality from the user's perspective also includes the *speed* of inference, which defines how fast the AI responds to users and is a decreasing function of the performance of the model and the number of tasks and modalities it can perform. It mainly depends on the quality of the software (i.e. architecture of the model and training data), and the quality of the hardware used as the endpoint for inference (i.e. AI chips), and it is critical for smooth user interaction and/or the use of AI at scale in industrial processes. For example, a small ("edge") task-specialised AI model can run (fast) on less advanced (older or non-AI specialised computers) while top quality generalist models (that are considered by some as precursors to AGI) may still be slower on the latest cutting-edge specialised AI hardware.

tools designed to perform complex tasks or a series of tasks *autonomously*.[19] The paper covers all of these modalities, including Multimodal variants under the *Text-to-Text* category, given that language models provide the "brain" of Multimodal variants.

- The *Openness* of an AI model refers to the accessibility of model weights to the public (which in turn allows for downloading and hosting the model by anyone with adequate hardware capabilities) or the availability of the model through commercial licencing. Openness in AI is a highly debated concept that has important implications for market structures, innovation, transparency, and regulation (Box 1). AI developers have been undertaking a variety of open-source strategies: either by only releasing the weights of AI models ("open-weights"), by contributing to the development of critical parts of AI software, or by giving access to training data or platforms to run and access AI models (see Annex B2).
- Finally, the *Price* of AI models is the cost of accessing the model (that is, providing inputs to it and receiving outputs from it) (see Annex C for details). More specifically, this price is paid when accessing foundation models developed upstream, most often for building on top of these models to create digital applications, either integrated directly in business operations or to be sold further downstream.[20] It is an important element of the total costs related to AI adoption – which also include complementary investments in intangible assets such as data and specialised skills. Hence, it is a key determinant of current and future adoption rates. The price is expressed per million units of text ("tokens"), in minutes of sound and per number of images. It represents the amount charged to the user for the interaction (questions and answers) with the AI model. One token represents approximately four characters, although this correspondence can vary depending on the model and the language used. This token-based pricing approach to evaluate AI models follows industry standards (Artificial Analysis, 2024).[21]

[19] For instance, the culmination of this "agentic" integration is achieved in *AI Robotics*, where physical tools (humanoid robots, industrial machines, as well as self-driving cars) are equipped with advanced AI capabilities to interact with and manipulate the physical world.

[20] It is not the price that final consumers pay for user-friendly applications such as chatbots. For a summary on those subscription prices, see Section 2.3, Table 3.

[21] One million tokens are roughly equivalent to a book of around 800-1000 pages. The input price is defined as the cost in dollars of one million tokens of prompting the AI model, while the output price is the cost in dollars of one million tokens of answer from the AI model. To provide a comprehensive price metric on the usage of AI from the API, a blended price is calculated as a weighted average of the input and output prices in order to proxy a representative interaction with an AI model (see Annex C for more details). For images, the price only includes the price of output (number of images generated by the AI), and for sound, the number of minutes of audio input to be transcribed into text (see Annex C for more details). While common across the industry the use of tokens as the unit for text can have a few limitations. First, the variability in tokenisation methods makes standardisation difficult, as the same text input can result in different token counts and costs depending on the model or language used. Second, the arbitrary weighting of input and output tokens may not accurately reflect usage patterns or underlying technologies. Lastly, focusing on price per token ignores other user costs, such as computational resources or data storage.

Box 1. The role of open source in AI markets

Why do for-profit companies contribute to open source?

The concept of open-source software was first introduced to refer to computer software that is freely available, provides the source code, and imposes no restriction on derivative work and usage – even though not all of them apply to several AI models that are labelled "open-source" (see below). For-profit companies undertake several strategies to capitalise on open-source software (Lerner and Tirole, 2002). First, companies can provide complementary services and products that are not supplied efficiently by open-source solutions. For example, RedHat Linux provides support targeted at corporations for Linux-based operating systems. AI cloud providers also fall in this category. The second strategy is to directly subsidise the development of open-source solutions, expecting to boost benefits in a complementary market segment. In the current AI landscape, the release by Meta of frontier AI models illustrates this strategy. The literature also suggests that the temptation to bet on an open-source strategy is particularly strong if the company is too small to compete commercially in the primary segment or when it is lagging behind the leader (Lerner and Tirole, 2002).

In addition, when innovations in the technology cannot be protected (for instance, no patenting is available) and no highly profitable use case has emerged yet, the innovator may decide it is better to release the details of the technology openly so as to benefit from feedback in the user and competitor community. When companies may not directly benefit from a clear complementary segment, they may still be willing to consider releasing in open source to benefit from the feedback and contributions of the community leveraging distributed peer reviewing and cumulative innovation. Companies can also benefit from a signalling mechanism of transparency and leverage open source as part of their public relations and marketing tools.

The benefits of open source in AI markets

Today, most actors in AI contribute to and benefit from open source – for instance, Meta by providing highly capable open-weight models, Nvidia with CUDA, a specialised software running its AI hardware, or IBM supporting the Linux operative systems that run most of the world's data centres. Open-source AI in a broader sense (including open research, open data and open tools and platforms) has been one of the major contributors to the AI technological and diffusion boom in every segment of the AI value chain (see details in Annex B, Table B2). Furthermore, the existence of this open-source ecosystem enables research and regulatory scrutiny of the technology.

Open sourcing software in particular generates positive externalities with beneficial effects on competition and innovation (Blind et al., 2021; Hoffmann, Nagle and Zhou, 2024). Open licensing enables easier access to technology, especially for small companies. It fosters the cumulative build-up of efficient innovation ecosystems by allowing output reproduction and extension, and by generating innovation feedback loops for the sector: the company developing open-source benefits from the improvements provided by members of the open-source community at large. It promotes transparency and helps mitigate anti-competitive behaviour by reducing switching costs and by limiting rent-seeking (White et al., 2024; Nagle, 2019; Chesbrough, 2023).

Open source has contributed to enabling new entrants as well. Indeed, the open-source community plays a critical role in ensuring efficient operability and maintaining low switching costs to ensure that downstream adopting firms can smoothly arbitrage across models and providers. The existence of a solid and dynamic open-source AI ecosystem has led to the creation of numerous new companies building end-consumer products relying on a combination of AI models (close and open) and open-source platforms, contributing, in turn, to competition in AI markets and across sectors.

Nevertheless, there is a risk that open-source ecosystems around specific foundation models could lead to market concentration by generating advantages for model developers that are able to create standards, integrate AI with cloud services, strengthen their power in adjacent markets or eventually decide to lock-in users by moving to closed models (Portuguese Competition Authority, 2024). Besides, open source raises issues around licensing.

Open source AI has many faces and can also pose challenges

Open source does not only mean free access to the source code or to the weights of the models – the latter being the case for Meta's offerings (Open Source Initiative, 2024; White et al., 2024; Solaiman, 2023; Bommasani et al., 2023; Liesenfeld and Dingemanse, 2024). To qualify as open-source, freely distributed software also needs to be licensed under recognised license terms that specify the terms of use under open-source standards and avoid license fees or usage restrictions for the commercial diffusion of models (White et al., 2024). In the context of AI, licensing has not been following the open-source conventions. Based on data from the Ollama platform, as of October 2024 (Figure B.2), only 25% of open-source demand (measured by the downloads of models for local deployment) qualify as true open source in the sense of conforming to the original ten principles of open source (Open Source Initiative, 2024). Often, only the weights of the models are released.

The reasons behind licensing restrictions may be related to litigation risks regarding the copyrighted nature of some data used for training the models and potential harmful downstream usage, as well as setting limits to the appropriation of the benefits by competitors. Restricted licensing could limit the benefits of open-source AI for diffusion and could constrain model extensions and cumulative innovation. This may be particularly the case for large scale usage in critical, high-risk applications or systemically important services (e.g. finance), where an uncertain dependence on a third party is not possible – especially under the terms of foreign jurisdictions, which is a relevant concern for most countries without local AI model providers. Clarifying licensing terms and providing a robust legal framework for open-source AI will be a key facilitator of the widespread adoption of AI (White et al., 2024).

Based on the model characteristics, several market segments within AI markets can be defined. Differentiating across them is essential to correctly assess market dynamics and highlight relative market power that could lead to incumbency advantage and abuse of market dominance in some of the (critical) segments. For example, for many tasks that need to be performed at scale, such as text and image classification or extraction, single-modality, smaller, faster, and cheaper models perform equally well as larger multimodal models. They are much less intensive in compute and data, allowing smaller and specialised AI labs to compete and users to interact with AI even without specialised cutting-edge AI hardware.

Conversely, *Agentic interaction*, which allows AI systems to autonomously make decisions and pursue complex goals with limited human supervision, involves large and highly polyvalent models. They are also extremely demanding in terms of high-quality and massive corpora of text, trained in highly-specialised and large data centres, generating potentially natural monopoly-type dynamics and gatekeeping positions. Therefore, large models for agentic interactions are orders of magnitude more expensive than smaller specialised models. This is true not only in terms of the initial training and finetuning phases, but also regarding the actual use, or *inference,* phase (Pope et al., 2022).

Large price differences across model segments lead to non-negligible marginal cost variations for AI model users, which rise steeply for more capable and multimodal models (Anthony Quentin, Biderman and Schoelkopf, 2023; Sardana et al., 2024). This is a key difference from many other digital services where the marginal cost is very small or zero (e.g. search platforms such as Google; social networks such as Meta) and explains why the inherent trade-offs of adopting AI entail several strategies that successfully coexist and address different needs and constraints of AI-adopting firms. Given their preference for quality,

intended usage, and budget constraints, AI-adopting firms will optimise across models and providers to power their services cost-efficiently and gain a competitive advantage in serving their downstream customers. Hence, it is necessary to focus on market developments within different market segments.

To capture differences both in usage (reflecting demand) and in technical characteristics (reflecting supply), we define three AI model segments in the offer of AI developers (see Annex C for a formal definition):[22]

- *Tier 1* model segment refers to "large models" that are the most capable (and expensive). For instance, as of February 2025 according to our definition based on model quality and price, OpenAI's *o1* and DeepSeek's *r1* models appear to be the current technological leaders in this segment among language models, but other companies are releasing similar, "reasoning"-type variants these months.
- *Tier 2* model segment refers to "medium models" that usually correspond to the main generalist models that handle most tasks in an AI chatbot, for example. Models like GPT4o (OpenAI), Claude Sonnet 3.5 (Anthropic) or Gemini 2.0 flash (Google), Deepseek v3, Mistral Large (MistralAI) or GrokV2.5 (XAI) fall in this category.
- *Tier 3* model segment refers to "small models", which are the cheapest and excel at specific tasks. They are much less constrained by compute capacity, both for training and for inference (usage). Models like GPT 4o-mini from OpenAI and Mistral Nemo from MistralAI fall in this category (in general, models of less than 10 billion parameters).

The ability of AI users to switch between models is a key enabler of competition, particularly as technology continues to evolve. Indeed, high switching costs, including both technical or other (financial) barriers that AI users incur when changing from one product, service, or supplier to another, can create barriers to entry, reduce the price elasticity of demand, and reduce the incentives to innovate if suppliers have successfully locked-in their user base. In a broader sense, switching costs are also influenced by the time and effort required to learn a new system, transfer data, or adapt to new processes.

Switching costs in Generative AI markets have several determinants, with some of them hard to detect given limited available data on demand and switching patterns. Technical factors include the architecture of AI systems, whose modularity currently allows for swapping or upgrading of components.[23] However, this may not be the case in all applications and is less true for deeply integrated AI systems in digital devices, such as voice assistants and operating systems. Regarding the required data for training or finetuning models, it can often be reused from one model to another and from one company to another with minimal modification.[24] As they stick to widely accepted technical standards, any data curation is portable with no additional fixed costs. Moreover, the open-source availability of many AI models and frameworks lowers the financial barrier to switching by eliminating licensing fees. Cloud-based solutions offer a variety of AI models as services, allowing users to switch across models (within and between AI

[22] The paper applies a similar market segmentation by model performance to each AI modality (for example, distinguishing between low-quality (Tier 3) and high-quality (Tier 1) images and textual models, separately). Those segments are formally defined in Annex C and will be used when defining assumptions about the composition of AI demand, as explained in the following section.

[23] Standardised interfaces and APIs enable different AI models to be integrated with minimal adjustments, reducing the effort and cost associated with switching. Additionally, the availability of pre-trained models and the compatibility and interoperability between different models within the same open-source framework (e.g., TensorFlow, PyTorch) further simplify the switching process. Rapid prototyping tools and automated coding pipelines streamline model selection and deployment, reducing manual effort and costs.

[24] As of the time of writing, more intensive use of a specific Generative AI model does not generally lead to substantially better performance, since models improve little from direct user feedback. This may change in the future and generate de facto switching costs.

developers) that are available from the same cloud providers with relative ease, at least from a technical point of view. They also reduce upfront costs and the need for extensive infrastructure investments. Strong community and ecosystem support, along with extensive documentation, lower the learning costs and reduce the time and resources needed to switch models.

These aspects notwithstanding, data feedback loops related to training or finetuning of models could be seen as *de facto* switching costs as they involve initial adjustment costs and learning on the side of users (Hagiu and Wright; 2025). However, attempts to exploit past user interactions, such as chat history (with the aim of personalisation which can increase the cost of switching to a competitor's model) have so far not appeared as a critical determinant of user choice in particular given the speed of innovation with new, more capable models rapidly replacing previous ones. Whether this strategy will provide a successful competitive advantage and effectively increase switching costs in the future by locking in early adopters remains to be seen. Switching costs can also include consumer or user habits or reputational aspects, related to the perceived risk or inconvenience associated with changing to a new service.

To account for the uncertainty around switching costs, most of the results are presented and discussed under high and low switching costs alike (Section 2.2 and in Annex C).

2.2. Indicators to track developments on AI markets

Several indicators are derived to capture AI market developments (Table C.3). First, by combining information on model quality (performance) and price, we trace the *AI Economic Frontier*: it is the subset of AI models available on the market that minimise prices for a given capability (Annex C). By incorporating prices to capture the cost of accessing AI, this paper evaluates competition based on a broader approach than those relying exclusively on capabilities (Korinek and Vipra, 2025) or on models' technical characteristics (Sevilla and Roldán, 2024). From the point of view of potential AI users, it captures the trade-off between price and quality, which in turn guides their choice of AI models in a way that matches their preferences and business needs. As more models become available and they evolve along the price and quality dimension, the shape and position of the AI Economic Frontier change rapidly over time, requiring frequent and near real time updates for effective monitoring (in our data collection, we followed a monthly frequency).

Second, from this frontier concept and underlying observations on model prices and quality, we develop further indicators both regarding estimates for market outcomes (the evolution of quality-adjusted prices and the churning rate of companies) and market structure (number of players and models, composition by country, and simulated market shares).

In the case of market outcomes, indicators of the movement and shape of the frontier are used to inform about the degree to which the benefits of innovation in AI have been passed on to AI-adopting firms in terms of lower prices and better quality. We derive two indicators to track this evolution:

- First, a quality-adjusted *AI price index* which measures the evolution of the price of AI models at the Economic Frontier for each AI modality (Text-to-Text, Text-to-Image and Audio-to-Text) and each segment (Tier 1, Tier 2, Tier 3) while accounting for improvements in quality (measured by model performance, see Section 2.1). An aggregate price index is also calculated to give a picture of the general dynamic of the AI market, by using assumptions about the composition of demand among the various AI segments.
- Second, the *churning rate* at the frontier, which is informative about the rate of innovation in AI and is measured as the proportion of models that are at the frontier in a month (at time t) but no longer at the frontier in the following month (at time t+1).

In the case of market structure, a common approach, which consists of looking at the evolution of market shares, is used to assess the intensity of competition. However, in our case we need to rely on estimates for the revenues generated from supplying AI models because not publicly available, comparable and

verifiable data is available, neither by listed companies (i.e. Google or Microsoft) nor from other leading AI companies who are not public and do not release official figures. Due to the lack of data on revenues and number of effective users, the literature has mostly focused on measures of inputs, costs, or popularity. This includes the size of investment in AI by Venture Capital firms, the size of investment (CAPEX) in AI-related infrastructures from incumbents, the ranking of AI capabilities of AI labs, the technical characteristics in terms of compute for training (EpochAI), the numbers of model downloads from open source AI platforms (like Huggingface) or estimations based on the web traffic of downstream applications (e.g. Generative AI Market Report 2023–2030 by IoT Analytics , discussed by Korinek and Vipra (2025) and Liu and Wang (2024)).

The approach in this paper instead focuses on the revenues from supplying AI. In particular, we provide simulations of the *relative* revenues accruing to foundation models, by all developers that provide models on the market (ie available from at least one cloud provider).[25] The revenues R_{jt} for each model *j* in each month *t* are calculated by combining three components: the supply component (S_{jt}) corresponds to the price and quality of models offered by developer *j* at time (month) *t,* the demand component (D_t) corresponds to the proportion of total AI demand addressed to each AI market segment, and the competition component (C_{jtN}) captures the capacity of each developer *j* to attract demand given the competition pressure at time *t* from *N* competing firms:

$$R_{jt} = f(Supply_{jt}, Demand_t, Competition_{jt,N}) \quad (1)$$

where *f*(.) is a function that sums the revenues generated by each model offered by developer *j* and given its competitiveness position with respect to the AI Economic Frontier (see Annex C for more details). Intuitively, an AI model generates revenues only if it is the AI Economic frontier, and this revenue is proportional to the price of the model and the proportion of the demand that this segment attracts.

Note that the supply component (price, quality) and some elements of the competition components (position with respect to the AI Economic Frontier) are observed and collected at a monthly frequency for the period January 2023 - January 2025. However, the (relative) demand component is not directly observed but calibrated, based on different scenarios to reflect global AI adoption during the period and the demand addressed to each model segment. The baseline demand scenario for the analysis corresponds to a scenario where total AI demand (number of users *x* intensity of utilisation) is distributed across the different market segments (use cases) and calibrated to reflect observed usage, such that it is 5% for the Tier 1 segment (Larger and most capable models), 70% for the Tier 2 (Medium) segment, and 25% for the Tier 3 (Small) segment.[26]

The third component that needs to be calibrated is the competitive environment. While this is a complex phenomenon in practice, driven by several aspects such as user preferences for branding, interfaces, bundling, we focus on a single aspect of it: the degree to which AI-adopting firms can switch across models and companies (switching costs). Since it is challenging to access data on actual use and switching, we consider two extreme cases to provide illustrative results that represent weakly and highly competitive environments:

[25] This includes open and closed-source solutions.

[26] Alternative demand scenarios are also presented in Annex D: an “AGI” scenario which stands for Artificial General Intelligence and captures high demand for the most capable models; and an “Edge scenario”, which captures high demand for smaller models, which are often deployed on local, lower-performance hardware such as personal computers or less cutting-edge cloud infrastructures, on the “edge” of networks. Their results can also be seen as competitive assessments that are more specific to the demand segments, in particular to the Tier 1 segment in the “AGI” scenario and to the Tier 3 segment in the “Edge” scenario.

- "*High switching*" costs refer to a scenario where AI users switch to the model at the frontier only if it costs significantly less (half as expensive) than the currently used model, and reputation (focality) driven concerns are strong (AI-users remain with their initial company of choice – *inertia* – at the frontier persistently, although the model does not belong to the frontier anymore).
- "*Low switching*" costs reflect a scenario where each month, all AI-users switch to the models at the frontier that corresponds to their preferred segment, regardless of the company of origin (absence of focality).

For further discussion, see Section 2.1 and the technical details in Annex C, and robustness checks under alternative scenarios in Annex E.

Combining these components lead to simulated market shares as follows:

$$S_{jt} = \frac{f(S_{jt}, D_t, C_{jtN})}{\sum_{n=1}^{N} f(S_{nt}, D_t, C_{ntN})} \quad (2)$$

where S_{jt} is the simulated market share of developer j generated by each of its active models and considering its quality, price, the comparable offer from competitors (position with respect to the AI Economic Frontier), the relative demand for this type of model and the competition environment (number of companies at the frontier, intensity of switching costs and reputation).

In conjunction with several other indicators of competition (Table C.3), the market share simulations provide a model-based benchmark combined with technical AI characteristics. Actual data on effective demand and more precise information on the supply of AI models and switching costs could further enrich this framework.

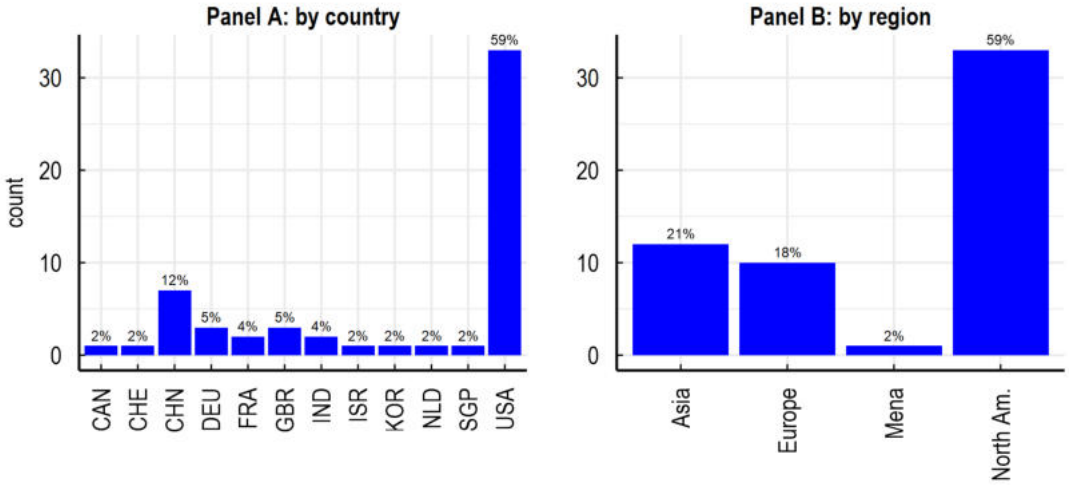
2.3. Data collection and measurement

A comprehensive data collection was carried out by screening the websites of cloud providers to track the characteristics of AI models on offer, including the price of access.[27] The database stands out for the global scope of the data collection, the relatively long-time dimension, the monthly update, and the comprehensive coverage of the supply of AI models available. The left panel of Figure 2 shows the distribution of AI cloud providers by country, with the United States accounting for 60% of the sample, followed by China at 12%, and other countries such as the United Kingdom, Germany, France, and India each contributing smaller percentages. The right panel aggregates this data by region, highlighting that North America accounts for 59% of AI providers, while Asia and Europe contribute 21% and 18%, respectively. The high degree of dynamism in AI markets is reflected in the fact that each month, additional providers enter the sample, particularly in Asia and Europe, which is taken into account by regularly updating the database.

[27] See Annex D for the full list of providers. We aim to maximise the representativity of the sample across countries and regions. However, given access restrictions from foreign sources in some countries (for example, in China) and the fast dynamics of the AI cloud provision markets (new players every month), recent and fast-growing countries or providers with no comparable offer may be underrepresented. Furthermore, a few cloud providers have announced the launch of AI services or implemented free preview plans (beta versions), but they are not included. For new providers, we retroactively compile historical vintages by extracting the price and model information from the internet archives using the Internet Wayback machine.

Figure 2. A comprehensive data collection of AI offer from the cloud

The distribution of AI cloud providers included in the sample, January 2025



Note: Country of origin denotes the country of the cloud provider. The figures show the number of cloud providers (y-axis) and the proportion (number on top of the bars) for cloud providers with publicly available pricing webpages of AI-as-a-Services (AI accessible from the cloud) identified during the period January 2023 and January 2025 and accessible from the United States and Europe. Offers by these cloud providers constitute our database in terms of model price, quality and other characteristics.
Source: authors' calculations.

To complement this information on AI prices, the data is merged with additional information from several additional sources (Annex D). First, the Large Model Systems Organization (LMSYS) project, hosted on Hugging Face, allows users to compare Large Language Models (LLMs) through the ArenaElo platform providing the source for the Elo rating. Second, Artificial Analysis offers comparative insights into AI models on the market, including performance indicators and inference speed metrics. Third, LiveBench provides dynamic evaluation metrics by testing models monthly on new prompts across six categories, mitigating potential contamination from repeated data. Lastly, EpochAI, a research institute, offers complementary information on training parameters and general characteristics of AI models. After combining all these sources, our sample includes around 500 models with information on both price and quality and many more with partial information available, totalling around 1,100 active models. Finally, we identify each model's developer and provider and their respective country of origin. The final sample is composed of models trained by AI developers from 14 countries and available from 51 cloud providers from 11 countries. This data is the building block of the analysis in Sections 3.1 and 3.2.

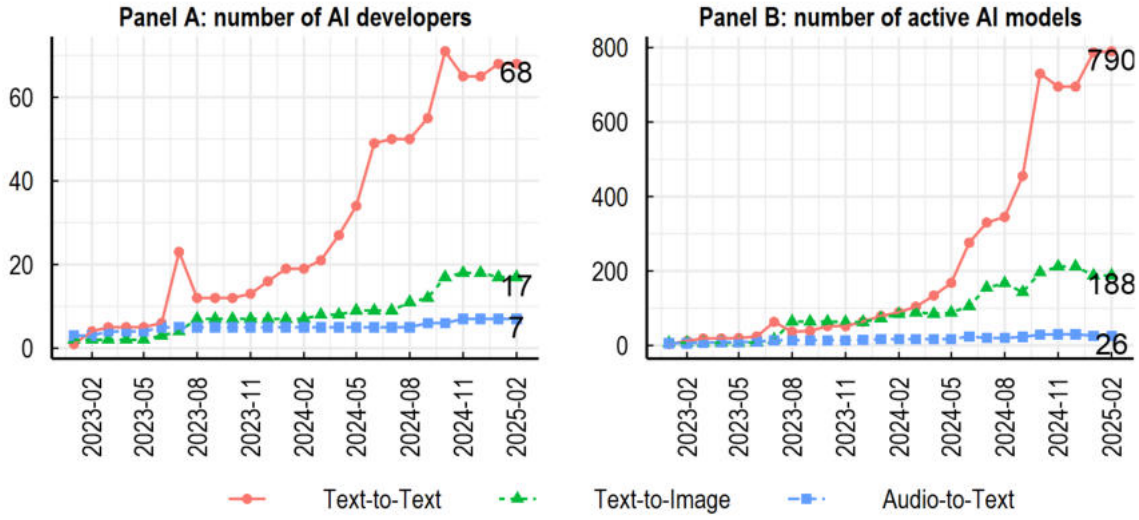
Finally, to track the downstream diffusion of AI accessible to the final user this paper tracks AI-powered digital services available in the market. The main source of data for this analysis is collected from There's an AI for That (TAAFT). TAAFT is a comprehensive AI discovery platform that gives users the capacity to find, explore and leverage the best AI tools. As of September 2024, the platform maps more than 12 182 AI services, classified across 2 608 different tasks and 123 ISIC economic sectors. This mapping provides a granular overview of the sectorial supply of AI-powered services and the sectorial popularity of those services on the platform discussed in Section 3.3.

3. Results

3.1. AI development

Since the launch of GPT 3.5 by OpenAI (ChatGPT) in November 2022, the number of AI models on the market has risen rapidly, driven by new developers entering the market and incumbent developers releasing new models as the technology advanced (Figure 3). The number of active models has surged more than 100-fold over the past year and a half, surpassing 1,000 as of January 2025. This rapid increase and the diversity of available models suggests developers are keen to address the diversity of AI user needs.[28]

Figure 3. The exponential rise of the supply of AI



Note: Developers refer to companies that train and optimise the AI models (e.g. OpenAI, Anthropic, Google, etc.), and Country refers to the country of origin of the AI Developer. Some companies both develop and provide models (e.g. OpenAI, Google, Microsoft, Amazon), some only develop models (e.g. Meta), and some others only provide models from other developing companies (e.g. ReplicateAI, PerplexityAI, Deepinfra). One developer can offer several models of different modalities through several cloud providers.
Source: authors' calculations.

Supply is largest and fastest rising for Text-to-Text models (large language models – LLMs – including assistants for coding and multi-modal features), at 78% of all model offerings, totalling 790 models in our last recorder data point. This is followed by Text-to-Image (image generation, at 18%, or 188 models) and Audio-to-Text (2.5% or 26 models) (Figure 3, Panel B). Within the Text-to-Text modality, multimodal variants have become the dominant paradigm for the more capable models like GPT4 in 2023, then later for smaller models in June 2024 (Annex B, Figure B.6).

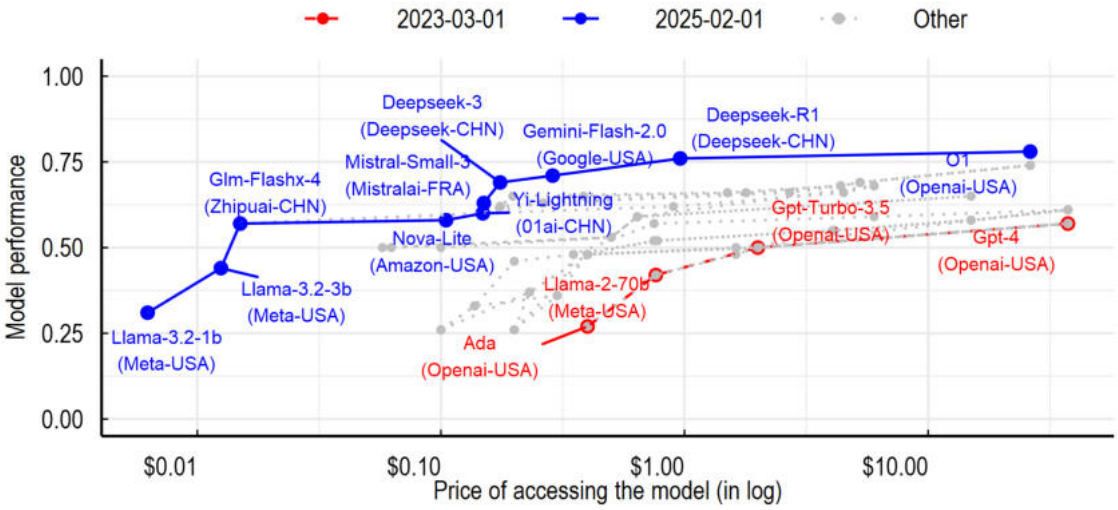
[28] Between March 2023 and November 2024, the estimated growth of the AI supply is an increase of 14 active models, 2.6 new AI developers and 0.67 new countries per month for the *Text-to-Text* modality, with a fast acceleration since the second quarter of 2024. This increase is half as fast for the image modality but almost null for the audio modality suggesting that the market for LLMs follows a particularly rapid expansion that is not as marked for other modalities or usages.

There has been substantial churn during this short period across models: 64% of models that were available at some point over the past two years are no longer active. Nevertheless, models of different generations coexist[29], with a succession of models and updates complementing the existing offers.

Capturing the joint evolution of model qualities and prices, Figure 4 depicts the *AI Economic Frontier* at the beginning of the sample period (March 2023) and in February 2025 at the latest data point retrieved, with each monthly frontier in-between depicted in grey.[30] Several conclusions stand out. First, AI models have continuously improved in quality (frontier moving to the top) and drastically declined in price (frontier moving to the left). The quality of the best model two years ago (GPT-4, in March 2023) is accessible in February 2025 at less than *one-hundredth* of the price. Second, in early 2023, the frontier was composed of four models from two companies; by now, users can choose across 10 Text-to-Text models at the economic frontier from 4 companies. This larger choice significantly broadens the possibilities for potential AI adopters to optimise along the price-performance trade-off.

Figure 4. The *AI Economic Frontier*: increasing model performance at much lower prices

Evolution of the AI Economic Frontier of AI language models (*Text-to-Text modality*)



Note: Performance is defined by a normalised weighted index of performance based on common benchmarks of the industry. Each dot represents the model with the best price-performance trade-off over the full sample of *Text-to-Text* (including multimodal) active models available each month. The solid (dashed) line represents the AI economic frontier in February 2025 (March 2023). For more details about the methodology, including performance and price measurement, see Section 2.2 and Annex B and C.
Source: authors' calculations.

Finally, the logarithmic shape of the *AI Economic frontier* suggests strong diminishing returns, where incremental improvements in quality at the top are accessible only when paying an order of magnitude higher price for them. For example, the model DeepSeek V3 is available for a price of USD 0.17 per one million tokens, while the ChatGPT o1 model (the most capable model at the time of completing the analysis) costs USD 26.23 for only a little higher quality when measured by model performance on industry

[29] This is likely explained by the fact that providers allow users to stick to previous model generations in their applications, and in this way avoid switching costs altogether (even if they are usually negligible).

[30] Figure 4 depicts the Economic Frontier for *Text-to-Text* and multimodal models, and Figures E.2 and E.3 illustrates show similar frontiers for *Text-to-Image* and *Audio-to-Text* models.

benchmark tests.[31] Such steeply rising model prices at the high end of performance create strong incentives for AI adopters to carefully optimise across model characteristics and not necessarily to aim at the best performer and most capable model. This fast increasing price near the top end of performance could reflect recoupment strategies for very high fixed costs related to model development. However, given the strong presence of open-source models at the frontier - where the cloud provider has free access to the model – they can also be indicative of the role of inference costs (the marginal cost component of AI model provision) and their rise as performance increases. This is due to the much higher computing capacity needed to serve more capable models (inference costs incurred by providers), which are also usually orders of magnitude bigger (in parameter size), and much more complex and energy intensive than smaller models.[32]

3.1.1. AI market structure

Several key features of AI model supply emerge when examining the *composition* of the *AI* Economic Frontier. While the number of LLMs on offer is large, at more than 700 models (Figure 3), the AI Economic Frontier consists of a small subset, around 10 (Figure 5). Of these, 6 are from the United States (*OpenAI*, *Google*, *Anthropic*, *Meta*, *Microsoft, Amazon*), 4 from China (*Alibaba*, *Deepseek, 01.AI, ZhipuAI*) and 1 from France (*MistralAI*) (see Annex D for the list of companies, their country of origin and their estimated distance to the frontier expressed in months; see the formal definition in Annex C).[33] The fact that a variety of models with similar characteristics (quality and price) coexist is indicative of heterogeneous preferences and several adoption frictions across consumer groups in different countries or regions, for instance, due to language and cultural specificities or national security concerns.

The second observation is that the Economic Frontier includes not only long-term digital incumbents (Google, Meta, Alibaba) but also new AI model developers, such as OpenAI (founded in 2015), Anthropic (2021), MistralAI (2023) and DeepSeek (2023). The latter group consists of not vertically integrated digital companies, although these companies are in most cases linked to incumbents via various partnerships including contractual relationships (on compute or distribution) or ownership. Over the past two years, they all have advanced the frontier, either by expanding capabilities or lowering prices. Notably, OpenAI has been leading innovation by consistently being the first to introduce the most advanced features, newer modalities and enhanced user experience. Google has introduced AI models and downstream applications within a much more vertically integrated system, which also produces hardware and provides cloud services. Meta has disrupted the market with highly capable open-source models, which are available through the cloud at lower prices[34], allowing a dynamic open-source ecosystem (Box 1) and an aggressive

[31] Based on the estimated slope of the frontier in January 2025, switching from a small, lower-performance *Tier 3* model (bottom left of the frontier) to a *Tier* 2 model (median of the frontier) improves the model quality by 10 p.p. (on a 0 to 100% scale) for less than 5 additional dollars. However, upgrading from a Tier 2 to Tier 1 model (top right of the frontier), a similar 10p.p. jump in quality – currently a theoretical possibility – would entail an additional USD 130 (Annex E4).

[32] Given that open-source models are present at the frontier, the prices of AI models on cloud offerings mostly reflect marginal (inference) costs and there is little room for markups, under the assumption that users can easily switch across models (see discussion in Section 2.1). Model prices may still vary across offerings, reflecting different architectures, quality of hardware and market conditions.

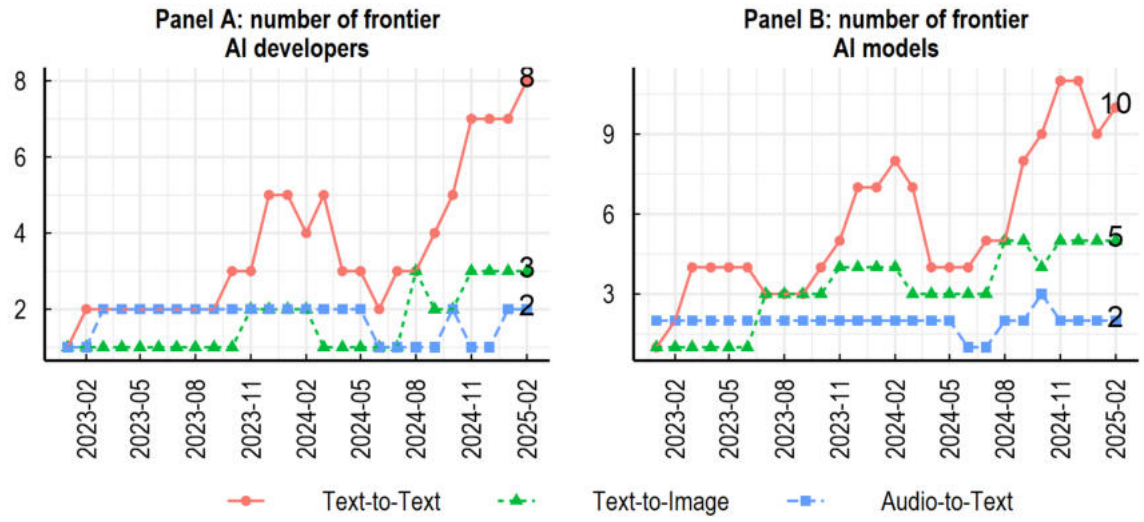
[33] While OpenAI and Meta were the only players at the frontier in the first semester of 2023, the number of frontier companies has fluctuated around four since 2023-Q3. The cyclicality of the number of players reflects a standard innovation pattern, with one leader moving the frontier for a few months (leading to a drop in the number of companies at the economic frontier) before competitors catch up (leading to a rise in the number of players at the economic frontier).

[34] Even though Meta's Llama models can be freely downloaded and installed for local deployment, accessing them through the cloud—even at a non-zero inference cost—can still be appealing for AI adopters, since they can avoid the

price war. MistralAI – another major open-source contributor – has pushed the economic frontier by providing small, efficient, and capable models with an order of magnitude fewer resources and hence provided at significantly lower prices (also in open-source with minimal licensing restrictions). More recently, Chinese models from Alibaba, ZhipuAI, AI.01, and DeepSeek have challenged existing leaders by proposing highly capable models with aggressive pricing, a comprehensive open-source strategy, and some cultural specificities, such as high performance in Asian languages (Artificial Analysis, 2025).

Figure 5. The *AI Economic Frontier* is oligopolistic with more and more players and models

Evolution of the number of AI-developer companies and AI models at the AI Economic Frontier



Note: The figure displays the number of AI developers and models at the AI Economic Frontier for each month during the period as defined in Annex B and C.
Source: authors' calculations.

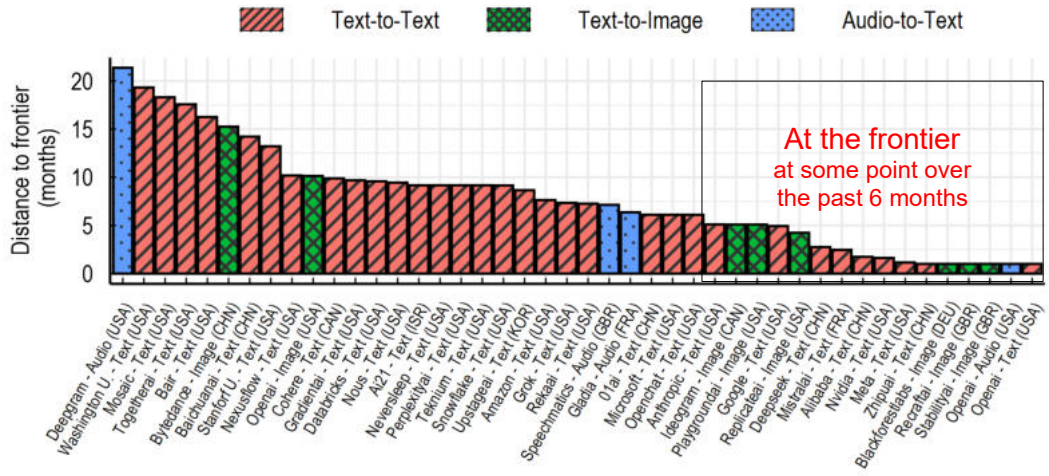
Finally, while the main tech incumbents have a leading role at the frontier among LLMs, this is less the case in the other modalities. The market for text models seems more vibrant with more models on offer, although also more models from the large digital companies (or new entrants financed directly by some of them, Microsoft – for OpenAI – or Amazon and Google – for Anthropic), which heightens risks of exercising market power. This is critical given that LLM models are often considered to be the "brains" of AI agents hence they are a crucial subsegment of AI markets.

When zooming on the country of origin of the developer company and looking at the *distance to the frontier in January* 2025 (Figure 6), we see that the frontier for AI models is composed of 10 companies from five countries: the United States, China, the United Kingdom, Germany and France. Closely following the frontier is, with less than a year delay, a group of 15 companies again from a diverse set of countries (the United States, France, China, Israel, Canada and the United Kingdom).

IT expenses associated with local deployment (Section 2.1). To illustrate the interest in using various open source models, Figure E.15 (in Annex E) presents the distribution downloads for local deployment. This shows a leading position for Meta, followed by Alibaba, Mistral, Google and Microsoft, at the end of 2024.

Figure 6. Several followers are lagging behind the AI Economic Frontier by only a few months

Average distance to the AI Economic Frontier of the best model by company and modality in the last 6 months (in months)



Note: The figure shows the distance to the frontier of the best model of each of the AI-developing companies by modality in January 2025. The distance to the frontier corresponds to the number of months necessary for an active model to be at the AI Economic Frontier. For example, a distance to the frontier of 3 months in February 2025 means that the current best model would have been at the frontier in November 2024.
Source: authors' calculations.

To illustrate how much the performance of models offered by followers have caught up over the past years, it is interesting to note that 17 companies or AI Labs (10 from the United States, 4 from China, and 1 from Canada, France and Israel) propose at least one model that surpasses the performance of GPT-4, the flagship model of OpenAI, which was the top performer in March 2023. Also, more than half (53%) of currently active models have better performance than GPT3.5, the version that powered ChatGPT at the moment of its initial launch in November 2022.[35]

Estimated market shares, based on the simulation based approach described in Section 2.2 are shown in Table 1, by region and modality under the baseline demand scenario and for two assumptions of the AI market structure (high and low switching cost).[36] In *Text-to-Text* AI models, the simulated market shares reveal a strong dominance of the United States under all scenarios: between 59% when assuming low switching costs and 87% under high switching costs. The combined share of France, Germany, the UK and Canada – which accounts for the rest of the OECD on the *AI Economic Frontier* – captures 10% (mostly from France and Canada) under high switching costs while China captures between 5% (high switching costs) and 36% (low switching costs) of the market. When comparing January 2025 to January 2024, the leading position of the United States in *Text-to-Text* models has remained roughly similar; however, the challenger position has moved from France and Canada to China, which is now the leading competitor to the United States, especially in the low switching costs scenarios. Among *Audio-to-Text* and

[35] See (Rahman et al., 2025[73]) on the number of AI labs and models trained at the scale of GPT-4.

[36] This computation does not include the revenues from subscriptions to end-user platforms like *ChatGPT, Claude* or *LeChat*, but only for the revenues generated from the providing access of AI to adopting firms through API (Application Programming Interface). For a brief overview of final user-oriented, subscription-based models, see Table 3.

especially among *Text-to-Image* models, the leadership is generally more contested and becoming more so over time.[37]

Table 1. The majority of AI models come from the US, but less so than a year ago

Simulated market shares under different assumptions about switching costs

(January 2024 and 2025, Baseline demand scenario)

Type of AI models	Country/Region	January 2024		January 2025	
		High switching costs	Low switching costs	High switching costs	Low switching costs
Text-to-Text	USA	86%	75%	86%	59%
	Rest of OECD (FRA+DEU+GBR+CAN)	12%	22%	10%	5%
	CHN	2%	4%	5%	36%
Text-to-Image	Rest of OECD (FRA+DEU+GBR+CAN)	38%	3%	57%	100%
	USA	62%	97%	43%	0%
	CHN	0%	0%	0%	0%
Audio-to-Text	USA	75%	32%	78%	58%
	Rest of OECD (FRA+DEU+GBR+CAN)	25%	69%	22%	42%
	CHN	0%	0%	0%	0%

Note "Rest of OECD" only includes countries that have at least one company at the frontier in at least one modality during the period. Simulated market share of AI model revenues per region of the AI developing company for different calibrations reflecting several assumptions on the determinants of market structure as detailed in Annex C. For the definitions of the scenarios and switching costs, see Section 2.2. The table reads as follows. In January 2025, in the baseline scenario under high switching costs the simulated market share of US companies developing AI is 91% of the AI revenues generated via API calls for AI models as a Service. Countries and country groups are ranked from highest to lowest simulated market shares under the central scenario (High switching cost) in January 2025 within each modality.
Source: authors' calculations.

Given the speed of innovation in AI and the market environment where competitors (and new entrants) can seemingly produce similarly high-quality models as existing players with a relatively short delay (Figure 6), the snapshot in January 2025 does not reflect the full dynamics of the market over the entire period and may overstate the position of the leaders with the latest model updates. Figure 7 shows the evolution of the simulated market shares during the entire period, assuming high switching costs among models to remain conservative regarding the contestability of the market (and under the *baseline demand* scenario).[38] The simulated market share of the United States in *Text-to-Text* models in January 2025 has come down from nearly 100% to around 80-85% since early 2024 when assuming high switching costs. Under low switching costs, when users are more likely to move to the best price-quality offer every time a new attractive offer becomes available, positions are more volatile and change drastically from one month to another. From September to December 2024, while the United States captured close to 98% of the

[37] Under the *"Edge"* scenario (not reported), where a large part of demand goes for *small* models, the non-US players, mainly from China and from France and Canada, capture a somewhat larger portion of the market.

[38] High switching costs characterise the strong friction environment modelled as a tolerance threshold that allows models twice as expensive as the frontier to be in demand by AI users and allows companies to retain their consumers despite a better offer available. Furthermore, one additional friction is introduced by including a focality parameter that smooths the market share over the entire period to capture the stickiness of consumers who may stick to a suboptimal model for a few months before switching to the best new model in the segment.

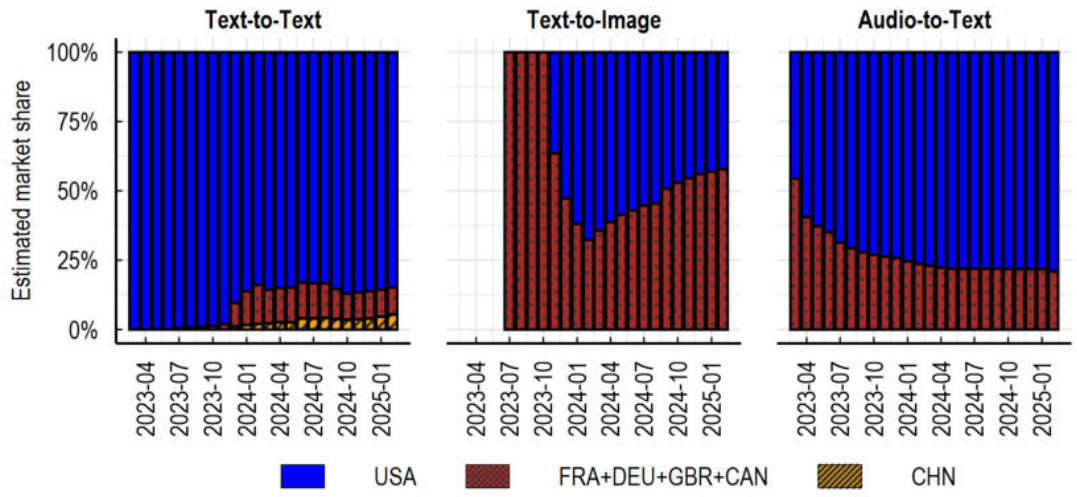
market, it only reached 60% in January and February 2025 (Annex E8) after the release of frontier models from Chinese companies (DeepSeek, Alibaba and 01.AI).

The leadership of the United States is much lower in AI vision models (Figure 7) than in the text modality where European companies like StabilityAI (GBR) and BlackforestLabs (DEU) have been proposing several models at the AI image Economic Frontier (Figure E.2). This result is even stronger in the scenario with low switching costs where the best offer immediately attracts all the demand (Figure E.8). In audio, the market was split between Europe (mostly the United Kingdom) and the United States in early 2023, but the United States (with *OpenAI* and *RecraftAI*) has gradually consolidated its position at around 75% of the market under the high switching costs scenario. This result holds under the low switching cost scenario, but the picture is much more volatile, with the United States moving from 100% of the market between September and December 2024 to around 60% in January and February 2025.

Overall, the estimates suggest that the market has been mostly captured by the United States, but companies from at least five countries have been able to get part of the AI market at some point in time and in some segments of the market and AI modalities. As such, at the country or regional level, simulated market shares in AI development have shown important swings under a range of assumptions, including high switching costs.

Figure 7. Apparent US leadership in AI development persists but has been challenged over the past two years

Simulated market share, by region of origin of the developer (Baseline demand scenario with high switching costs)



Note: Simulated market share of AI model revenues per region of the AI developing company for different calibrations reflecting several assumptions on the determinants of market structure as detailed in Annex C.
Source: authors' calculations.

Figure 8 shows simulated market shares with a view to highlight the market share of the largest developers, the leader (OpenAI) and other AI companies (including Anthropic, Cohere, Mistral, Alibaba or DeepSeek). To reflect the idea that changing models *across* companies (and not *across* countries) is less costly for users, these estimates are shown under the low switching costs assumptions and the baseline demand scenario (results for high switching costs are shown in Figure E.9 and Figure E10 to E13, also under alternative demand scenarios). In *Text-to-Text* the market leadership of the United States is mapped one-to-one with four companies (*Google*, *Meta*, *OpenAI* and *Anthropic*) while in other regions and across all modalities startups are leading - *Cohere* (Canada), *MistralAI* (France), *Blackforest Labs* (Germany), *StabilityAI* (United Kingdom), *Speechmatics* (United Kingdom) or *DeepSeek* (China) among a few others.

While this is indicative of a vibrant supply underpinned by startups who develop highly capable models, it is a question whether this will remain the case or these newcomers would be absorbed by incumbents or risk running out of financial resources to stay in the race.

The simulations where the market leadership of OpenAI is the strongest (above 90%) is under the "*AGI"* demand scenario, where users prefer more strongly the most capable models. Under high switching costs, OpenAI's leading position has been virtually uncontested (Figure E.10). However, under low switching costs, simulations suggest that OpenAI lost ground temporarily between June and July 2024 mostly to the benefit of Meta, Google and Anthropic – but it managed to regain the leadership between September and December with the release of the *o1* models series (Figure E.11). Results for the leader OpenAI under the "*Edge demand"* scenario, when users more strongly prefer smaller models are fairly similar to the baseline scenario (Figure E.12 and Figure E.13). However, the position of tech incumbents (Microsoft, Meta, Amazon and Google) is stronger (around 30% in the high switching cost scenario (Figure E.12)), consistent with the intuition that companies with the existing user base and already profitable non-AI digital products have a stronger comparative advantage in cost-effective smaller models that they can integrate to existing products and deploy at scale on existing IT infrastructures and users.

These results suggest that even under higher assumed switching costs, the simulated AI revenues have not been primarily concentrated among the leading tech incumbents and that a large amount of entry has been coming from emerging start-ups, in particular in modalities other than text. The current leader in AI, OpenAI has been vulnerable to important market share losses according to our simulations, virtually losing its leadership position between May and July 2024 to companies like Anthropic, Google and Meta in Text modalities and StabilityAI and BlackforestLabs in image models. These results generally hold true under most of the scenarios, with the more concentrated results (under the AGI demand scenario and high switching costs) obtained in the scenario where the assumptions are the farthest from what could be observed during the period under consideration.[39]

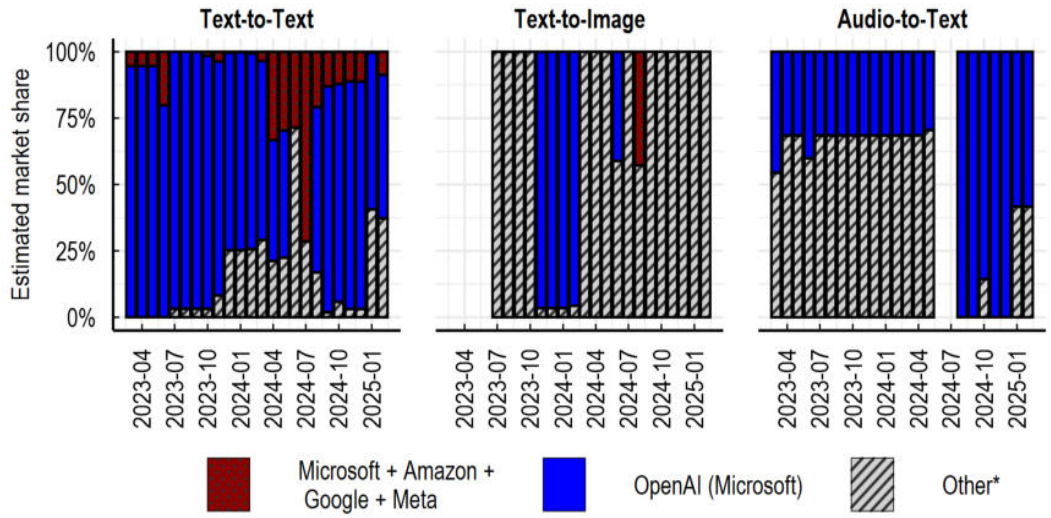
Overall, as far as model development is concerned and during the period under consideration, our estimates support the idea that the concentration of inputs (data, compute and talent) has neither blocked new entries nor prevented quality from rising and prices from falling. Leaders have been exposed over the past two years to new offers by followers within two to three months of the release of their models. This could have played a role in their innovative efforts on several fronts (performance, user experience, speed, modalities, etc.) to maintain their competitive advantage. So far, the market for AI development appears to be a market which new players can enter with attractive offers in terms of quality and price, even if it is unclear how successful these players have been in attracting users in practice and several risks persist.[40]

[39] See section 2.2 for a discussion on switching costs. The AGI demand scenario assumes 70% of the demand for the most capable model available at every period. This assumption is unlikely to be possible due to the infrastructure constraints of the inference of these models. During this period, quotas on the usage of the most capable model have been the norm from all companies. Even if the optimal choice for most consumers would have been the more capable model on offer, supply constraints have forced companies to rely on tier 2 models for most of their usage. This is consistent with the communication of the companies (API documentation), from the choice of models on chatbot interface like ChatGPT and from the token demand reported by OpenRouter. For both demand and supply arguments, the baseline demand scenario under low switching cost is our preferred scenario and what we believe is the most likely scenario.

[40] This section does not consider the capacity of incumbents with existing user bases to leverage their "captive" consumers in adjacent markets. This is discussed in Section 3.3 when talking about AI-powered services to end consumers. This section assumes that AI-adopting firms are rational profit maximisation companies with perfect information and full model substitutability. Extensions of the model could relax these assumptions to include, for example, "within user externalities" (Hagiu and Wright, 2023) that would capture the capacity of platform companies to leverage the vast amount of personal data collected on their users. In practice, the focality parameters set to 22 months in the high friction environment capture part of this effect but assume that the value of personal data is zero

Figure 8. The leading AI developers appear strongly challenged by large tech incumbents and by smaller players

Simulated market share, by selected companies (Baseline demand scenario with low switching costs)



Note: Simulated market share of AI model revenues per tech incumbent status of the AI developing company under the baseline demand scenario and in a low switching cost environment (as detailed in Annex C). Microsoft appear twice due to its close partnership with OpenAI (FTC, 2025) and their own AI labs that also develop AI models. Other include Cohere (CAN), MistralAI (CHN), 01.AI (CHN), Anthropic (CHN), Alibaba (CHN), and Deepseek (CHN) in the Text-to-Text modality; StabilityAI (GBR), Blackforestlabs (DEU), RecraftAI (USA) in Text-to-Image, and Speechmatics (GBR) and Gladia (FRA) in Audio-to-Text.
Source: author's calculations.

The evolution of simulated market shares up to the beginning of 2025 suggests volatile market shares: one leap from the leader is being caught up a few months later by several competitors. Those results are in line with the findings by Hagiu and Wright (2025), Korinek and Vipra (2025), Artificial Analysis (2024) and Artificial Analysis (2024). Nevertheless, the diminished advantage of the leaders, observed in the first half of 2024, has been partly reversed by more recent updates from some of them (*OpenAI* and *Google*) in the *Text-to-Text* modality. Whether the early quarters of 2025 will allow followers to catch up like in the first months of 2024 or whether (US) leaders will succeed in consolidating their oligopolistic position is an open question that should be closely monitored. Similarly, the acquisition by leaders of leading startups focussing on other AI modalities (horizontal acquisition) could raise concerns, insofar as it may stifle innovation. As more data become available a more precise calibration of demand should help to provide an accurate and real-time picture of the AI market structure.

3.1.2. AI market outcomes

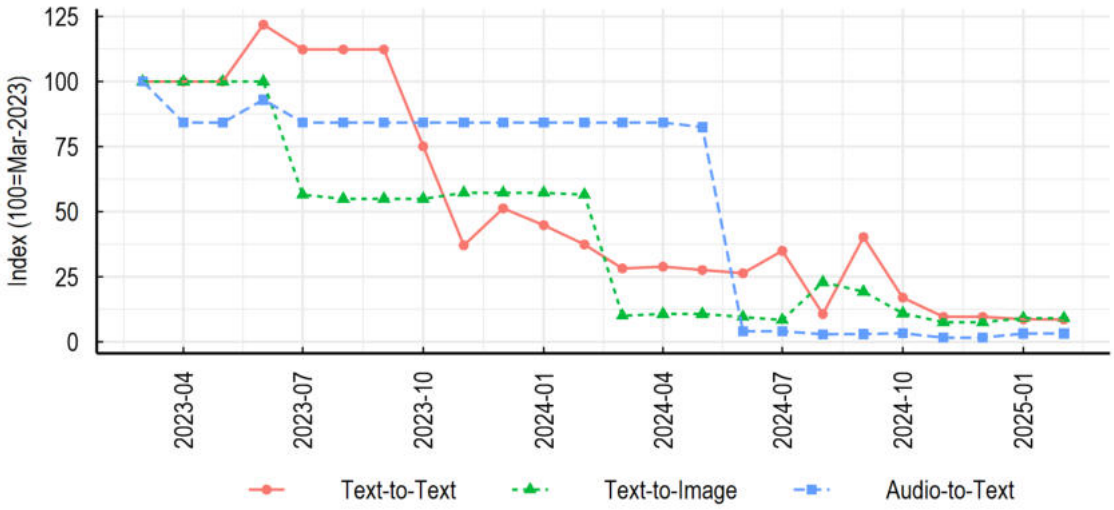
This section discusses several market outcomes: the evolution of prices, quality, and the pace of innovation. The *AI Economic Frontier* has been shifting to the left and upwards, indicating that models have, in all segments, improved in quality and decreased in price – as is typically the case for new technologies and in markets with network externalities, which in turn enables and fosters their broad-based adoption. Figure 9 shows the evolution of the quality-adjusted price of AI models. While the timing and amplitude are different across modalities (and across AI quality segments) the consistent trend is a massive price decline that has largely benefited AI adopters. A combination of several factors is likely to

when older than 22 months, ruling out the advantage of long-term incumbents in adjacent markets but still allowing for some incumbency advantage.

explain the rapid declines in quality-adjusted prices: improvements in hardware (Erdil, 2024; Sevilla and Roldán, 2024) and in training data quality (better data "curation" and use of synthetic data; Hunt et al, 2023) and software (Ho et al, 2024); increased mode use intensity (and competition pressures, in particular, due to the availability of open-weight models that moderate prices by facilitating market entry and allow cloud providers to offer models close to marginal cost (see Table B.2 for a summary of the role of open source in AI markets).

Figure 9. Quality-adjusted AI prices have been falling rapidly for all modalities

The quality adjusted price of AI models (March 2023-January 2025)



Note: The index represents the evolution (starting in March 2023) of the quality-adjusted price of using AI models from the cloud for each modality. The index for each modality is a weighted sum of the index for each model segment (Tier 1, 2 and 3) under the baseline demand scenario (respectively 5%, 70%, and 25% of total demand). Each model segment is represented by the model with the best price /quality trade-off available in a given month at the AI Economic Frontier.
Source: authors' calculations.

Figure 9 shows the aggregated price index by modality with a weight for each quality segment (Tier 1, Tier 2, Tier 3) calibrated to the *baseline demand scenario (consistent with the central scenario in section 3.1)*. Because the price decline has been stronger for cheaper models, the aggregate index is sensitive to the demand scenario, with a decline of 45% in the *AGI scenario* (demand mostly for the more capable models) and 98% in the *Edge scenario* (demand mostly for the cheaper models)[41] but the general trends apply to all scenarios and modalities (see Figure E.6). Using the information on the number of AI-powered services from the platform *TAAFT*, we can infer the relative importance of text, image and audio-based tasks. This *"Agentic scenario"* (aggregating all modalities to match the demand for AI-powered services) generates an average price decline of about 80% under the *Baseline demand* scenario.

However, several risks stand out regarding the sustainability of price declines going forward. First, hidden price increases could come in the form of opaque changes in the pricing and usage quotas rather than via nominal changes in the price, making it more difficult for adopting firms to notice price increases ex-ante.

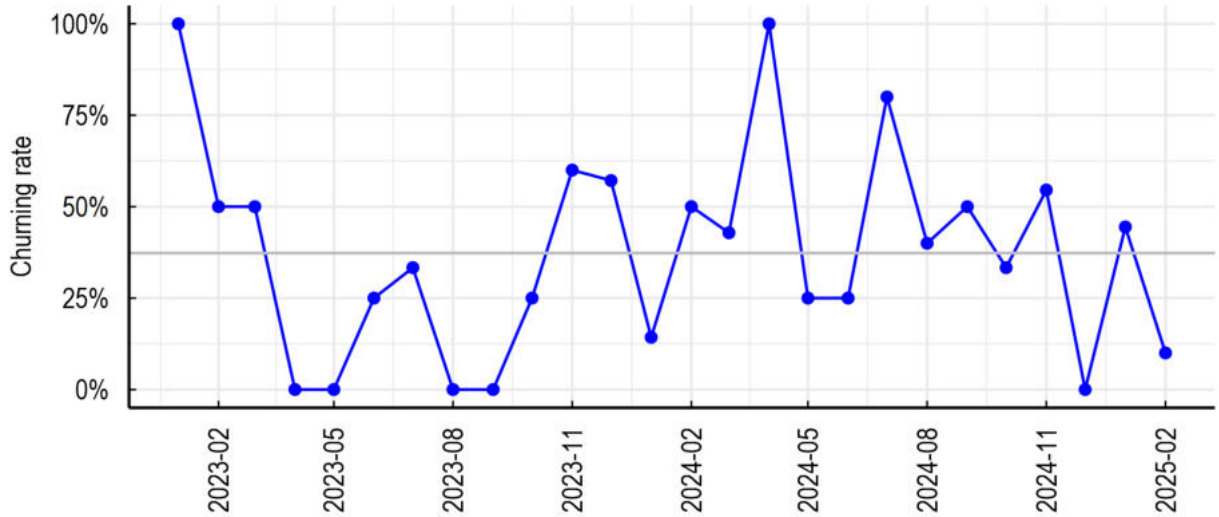
[41] The weights of segments have no reason to be uniform, with the assumption of the best model attracting all the demand *(AGI scenario)* probably being less plausible. The relative demand for different segments of LLMS access via the cloud provider *OpenRouter* provides metrics on the usage of the most active models available. According to the data, demand for the most capable o1-preview model represents less than 1% of the total demanded tokens of the platforms between the 10th of November 2024 and the 10th of December 2024, with 70% going for models in the tiers 2 segments like Claude sonnet 3.5, GPT4o, Gemini 1.5 and 30% for small models in the Tier 3 segment.

This could also allow companies present in all segments to apply excessive markups in the less competitive segments (Tier 1 segment dominated by OpenAI and Anthropic) to subsidise the segments where competition is the highest, and "exhaust" competition by intense dumping. Similarly, the release and pricing of *"reasoning models"* (*o1* et o3 models from *OpenAI or Deepseek r1*) poses a measurement challenge, as it leads to underestimating the price of the more capable models.[42] Depending on the complexity of the reasoning (size of the intermediary output), the effective cost would be multiplied by five if we assume a five-step reasoning process. Figure E.5 shows the evolution of the price index when factoring in the cost of "hidden" tokens. In this scenario, the most capable models are around 50% more expensive than two years ago, even after adjusting for increasing quality. Moreover, in the longer run, it is likely that this pattern of price reductions will only be sustained by tech incumbents with the capacity to subsidise AI to the detriment of non-AI services (increasing the price of services where consumers are already locked in and where their market power is high).

Turning to the speed of innovation in AI, it has been incredibly fast with new models released every month, leading to increasing performances on benchmarks, better user experience through smooth integration with digital tools like web search or programming languages, and extensions of the perception capabilities with more and more vision and speech interactions. Indeed, Figure 10 shows the churning rate of AI models at the *AI Economic Frontier,* providing evidence that the competitive advantage given by the release of a new frontier model has never provided definitive leadership and has required an acceleration of innovation from the leaders themselves to maintain their advantage and sustain the product momentum of AI, critical for long term adoption. For example, in *the Text-to-Text* modality, 100% of the models at the frontier in March 2024 were no longer at the frontier in April 2024, and on average 30% of the frontier models across all modalities get depreciated every month (Figure 10), similar results holds for the other modalities (not reported). Despite the important fixed cost of training, models at the economic frontier are, on average, displaced in less than three months. The comparable figure is around six months to one year for the most capable, top-quality (Tier 1) models.

Figure 10. The churning rate of models at the AI economic frontier is strong

Share of new models that appear at the Economic Frontier from one month to the next, Text-to-Text



[42] The underestimation of the price of the more most capable models may arise due to the lack of transparency of the reasoning ("*Chain of Thought*") process, which include intermediary steps of output that are often hidden but are still billed.

Note: The churning rate at the frontier is computed as the proportion of models at the frontier at time t-1 that are not at the frontier at time t. The grey line represents the average during the period for each modality. The figure reads as following. 60% of models that where at the frontier of the text modality in august 2024 were no longer at the frontier in September,
Source: authors' calculations.

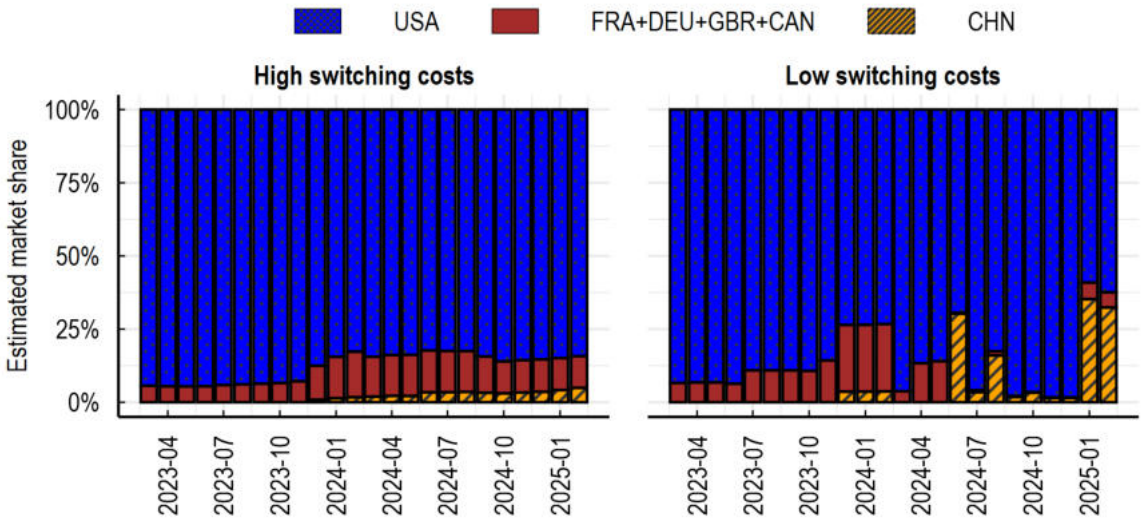
3.1.3. Horizontal economies of scale vs economies of specialisation

Many of the initial concerns about competition in AI were rooted in the assumption that economies of scale and scope in compute and data would pose challenges to new and smaller players. Indeed, leading companies like *OpenAI* are very horizontally integrated across AI modalities (text, image, audio, videos) around LLM architectures and *Google Deepmind* is even more broadly integrated across architectures (LLMs, AlphaFold) and domains (entertainment, biology, health, etc...). While those companies have taken the lead in research and marketing, leveraging economies of scale and scope (including customer base), they are challenged by startups (mostly founded by former tech incumbent employees) explicitly betting on economies of *specialisation* like Anthropic (USA), *Deepseek* (CHN), *MistralAI* (FRA), *RekaAI* (USA), or *Cohere* (CAN) in *Text-to-Text* modalities; *StabilityAI* (GBR), *Midjourney* (USA), *RecraftAI* (GBR) and *Blackforest Lab* (DEU) for *Text-to-Image* or *Speechmatics* (GBR), *Gladia* (FRA), *ElevenLabs* (POL), *Cartesia* (USA), *LMNT* (USA) in audio. The ability of several companies to reach or approach the AI Economic frontier with significantly less funding suggests that current competitive advantage in model development extends beyond economies of scale and scope.

Figure 11 shows the simulated market shares in a scenario of horizontal integration (aggregating all modalities), assuming the *Baseline demand scenario* (as in Section 3.1.1 and 3.1.2). Results suggest that US leadership is strong but slightly declining over time, starting around 2024. This is especially the case when switching costs are low, and users adopt the best price/quality models more rapidly and easily. In this scenario, the US simulated market share during the past months is reduced to around 60% (right panel).

Figure 11. US leadership is becoming less clear when jointly considering all modalities, especially when switching costs are low

Simulated market shares when combining the three modalities (text, image, audio)



Note: Simulated market share of AI model revenues per country of origin of the AI developing company under the baseline demand scenario and aggregating all modalities according to Annex C. In this scenario, 40% of AI demand is addressed to Text-to-Text models, 10% to Audio-to-Text and 50% to Text-to-Image and is calibrated to match the relative popularity of each modality in downstream AI apps using data from Taaft (TAAFT, 2024).
Source: author's calculations.

3.2. AI provision through the cloud

The simplest approach for AI adopting firms to access generative AI models is through cloud providers (via an API), since this approach requires no additional internal IT resources but is based on on-demand access and per-usage payment. This dependence on the cloud infrastructure generates incentives for vertical integration. Three companies (*hyper-scalers*) *Amazon* (AWS), *Microsoft* (Azure) and *Google* (GCS) are far ahead in AI-specialised investments (Draghi, 2024). The vertical (de facto) integration of leading AI labs with hyper-scalers raises concerns about the risk of cloud gatekeeping, defined as the exclusive position of a few actors that control access to the AI models of other AI developer companies (FTC, 2025).

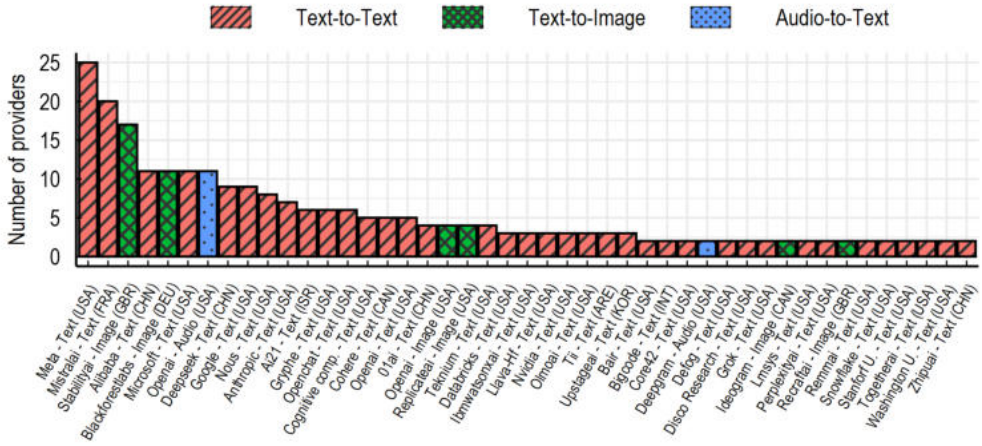
The following subsections discuss two potential initial signs of cloud gatekeeping in AI inference. First, whether *AI-developer* companies are locked in within a single cloud provider and are at risk of monopsony (i.e. relying on a single cloud provider to serve their models), creating risks of asymmetric market power between AI developers and cloud incumbents. Second, whether models accessible from the largest cloud providers by AI-adopting firms are priced above market prices and whether their pricing strategies have changed during the period. The results are not a substitute for a comprehensive and definitive competition assessment of the entire AI cloud market but provide initial empirical evidence on a subset of two potential channels that incumbents could have leveraged.

3.2.1. The links between AI developers and cloud providers

We find that AI models are available from 46 different cloud providers (through APIs). Most of them offer a large array of models with diverse capabilities and prices.[43] Models from a few companies betting on open-source solutions are available from more than ten providers (Figure 12). Among the companies that stand out, we find a few non-US start-ups like *MistralAI* (France), *StabilityAI* (The United Kingdom) and *Blackforestlabs* (Germany), as well as open-source models from tech incumbents like *Meta* (US), *Alibaba* (China), *Microsoft* (US) and *Google* (US). This broad availability within providers allows for switching across models and their developers while continuing to use the same cloud provider, exploiting arbitrage possibilities on the market of AI models. However, this could be less the case in the future if cloud providers favour their own models or if they successfully personalise their model (using user-specific data for context) and effectively lock in their existing user base. This may be particularly critical as AI-adopting firms integrate AI more deeply in their production process and their existing IT and data infrastructure, effectively leading to a rise in *switching costs* as is the case in traditional cloud services (such as storage, computing capacity, etc.).

Figure 12. A few AI developers are particularly popular among cloud providers

Number of cloud providers serving AI models, by developer of model (and their country of origin), January 2025



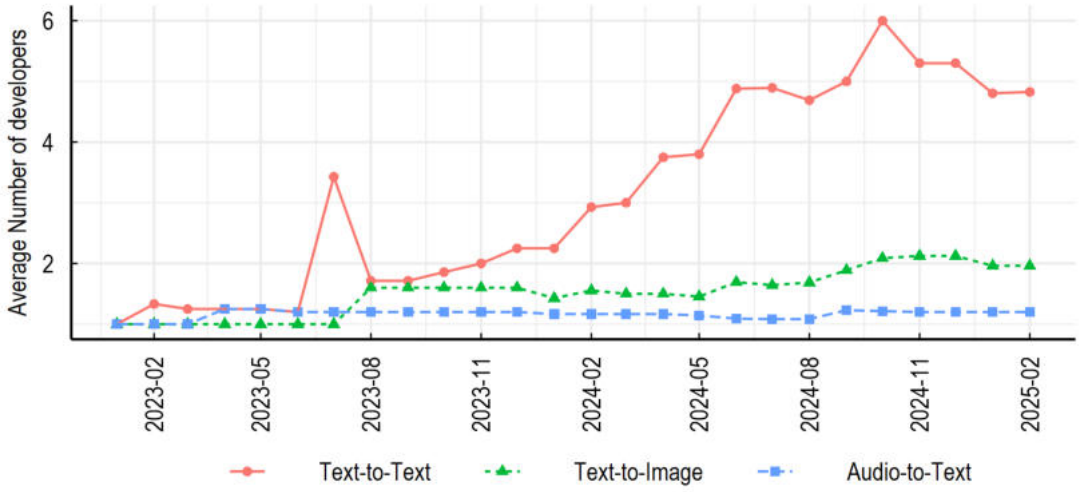
Note: Only companies developing models served by at least two cloud providers are shown.
Source: authors' calculations.

Larger cloud providers increasingly provide models from several developers, particularly for Text-*to*-*Text*, since early 2024, reaching, on average, seven developers per provider by the end of our sample (Figure 13). Cloud providers compete both for the provision of frontier models and for "good enough" models, suggesting that their common strategy is to provide tools for different tasks, and various model qualities (best and worst model range in Figure 14) and prices, allowing customers to arbitrage and select among models and developers at no additional cost.

[43] The current data only considers text, image, audio or multimodal generative AI models accessible from the US or Europe. Some Chinese models are not accessible outside China. Some cloud providers like OVH propose a non-standard pricing of the model, which makes the comparison difficult. We excluded these cases from the discussion in this paper.

Figure 13. Cloud providers increasingly serve models from several developers

Average number of AI developers per cloud provider

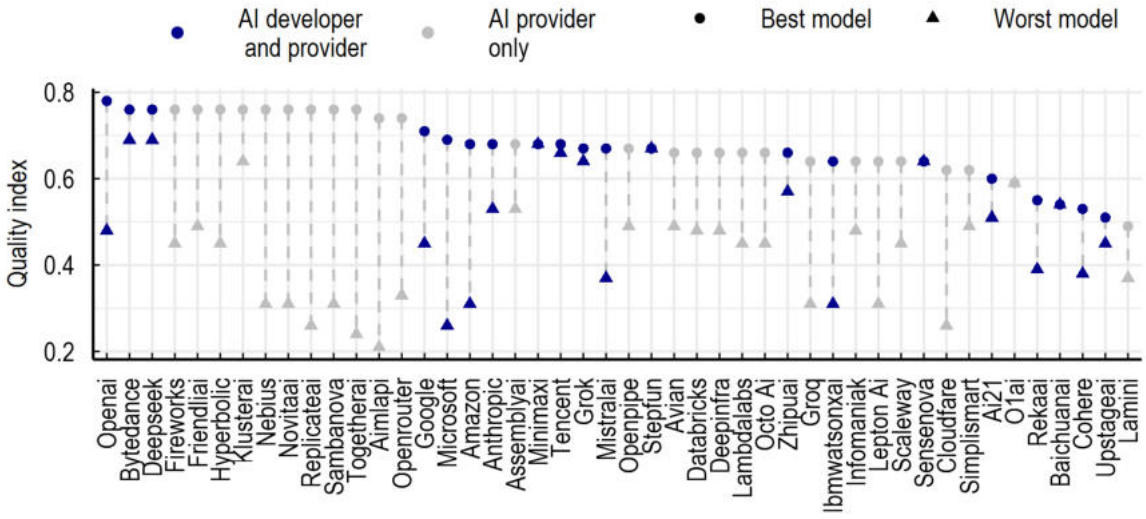


Source: author's calculations.

Specifically, we find that around 15 providers offer highly capable models, with performances very close to those of OpenAI's *best* model (at least 0.75 on the quality index reported in Figure 14). Moreover, almost all providers' models are at least as capable as the *median* model from OpenAI (around 0.6 on the quality index reported in Figure 14). Cloud providers that do not develop AI models in-house (grey dots in Figure 14, indicating *AI provider only*) also offer highly capable models. This broad access to the most capable models is possible because open-weight models have largely bridged the capability gap with top closed models and because partnerships between AI developers and the main cloud providers for distributing models are frequent (*AI provider only* on Figure 14).

Figure 14. Several cloud providers offer the most capable models

Quality index of the models offered per cloud provider for Text-to-Text models, February 2025



Note: The *Quality index* on the vertical axis measures the quality (or performance) of AI models, by combining various industry benchmark test scores of foundational AI model performance, as described in Section 1.2.
Source: author's calculations.

3.2.2. Testing for the pricing power of large cloud providers and the leading developer

This subsection undertakes a simple econometric analysis to investigate whether prices offered by hyperscalers, namely Amazon (AWS), Microsoft (Azure) and Google (GCS), and that of the leader in AI development (OpenAI), differ from the rest of the market, which would be one indication of market power.[44]

We run the following regression based on the structural equation presented in Annex F that links the price of inference P_{ticp} (indexed by t for time, i for model, c for the model's developer company and p for the cloud provider of the model) to a software component (quality of algorithms and training data), a hardware component (marginal cost of running the AI model) and a demand component that reflects the capacity utilisation (extensive margin of adoption) and intensity utilisation (intensive margin of usage of AI):

$$P_{ticp} = \alpha_1 + \beta_1 Q_i + \beta_2 \mathrm{Q}_i^2 + \beta_3 O_i + \beta_4 M_i + \beta_5 H_p + \beta_6 L_c + \left[P_P + T_t + L_c * T_t + H_p * T_t\right] \quad (3)$$

where Q_i is a quality index measuring the performances of models on industry benchmarks (Section 2.2 and in Annex C), O_i a dummy variable denoting whether the model is open-source, M_i is a dummy variable if the model is multimodal, H_p a dummy variable for models served by *hyper-scalers* (Amazon, Google, Microsoft), L_c is a dummy for the leading developer in terms of model quality (*OpenAI*), P_p is the provider fixed effect that controls for unobservable, time-invariant hardware characteristics to serve the model (including provider specific compute infrastructure) and its user base and usage intensity (demand component), and T_t is a time fixed effect that control for unobserved evolving characteristics common to all providers such as the trend price of compute and the evolution of the user base size and usage intensity.

The results reported in Table 2 first relate prices only to model capabilities (quality and multimodality; column 1) and show that they are significantly and positively related to model prices. Column 2 then identifies a clear non-linear relationship between quality and price: accessing marginally higher quality models costs disproportionately more at high-quality levels, implying diminishing returns to AI adopters from paying higher AI prices, consistent with Figure 4 on the AI Economic Frontier. This result remains when including a full set of provider and time fixed effects (column 3).

Second, we find that open-source models that require no licence fees are cheaper than the equivalent closed models (column 2), reflecting that cloud providers serving open-source models do not need to recoup the fixed costs of training models and can offer them for cheaper, more closely reflecting the marginal cost of providing model access (inference cost).

Third, we find significantly declining time fixed effects, which may reflect a combination of improving technology (cheaper input costs, in particular hardware) and declining pricing power (column 4). Fourth, a sign of pricing power or "premium" with respect to the market from the leader in model quality, OpenAI, was observed only in the first semester of 2023 (column 5). This premium disappeared during 2024 (interaction terms) when models offered by OpenAI were not significantly more costly than models from other companies, after controlling for quality and other model characteristics. Finally, the coefficients for hyperscalers in columns 5 and 6 are not statistically significant, indicating that these leading cloud providers do not stand out among other providers for charging either higher or lower prices than other competitors.

[44] See (Bergemann, Bonnatti and Smolin, 2025) for an economic framework to analyse the optimal pricing and product design of Large Language Models (LLM).

Table 2. Testing for the pricing power of major AI providers and developers

Estimation results from regressing the price of AI models on model and cloud provider characteristics

	Dependent variable:					
	AI price of inference					
	(1)	(2)	(3)	(4)	(5)	(6)
AI Quality	**0.133*****	-1.253***	-0.958***	-0.965***	-1.039***	-1.084***
	(0.020)	(0.134)	(0.144)	(0.144)	(0.141)	(0.136)
(AI Quality)^2		**0.015*****	0.013***	0.013***	0.013***	0.014***
		(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
Multimodal	5.417***	2.584***	2.476***	2.479***	2.837***	2.738***
	(0.653)	(0.687)	(0.731)	(0.729)	(0.684)	(0.492)
Open source		**-2.988*****	**-2.126*****	-2.133***	-2.183***	-2.628***
		(0.480)	(0.572)	(0.571)	(0.524)	(0.492)
Hyperscaler					-0.749	0.878
(Amazon, Google, Microsoft)					(0.657)	(0.653)
OpenAI					**15.017*****	
					(3.443)	
2023-S2				**-6.085*****	2.395	
				(1.953)	(3.033)	
2024-S1				**-8.739*****	0.546	
				(1.804)	(2.827)	
2024-S2				**-10.528*****	-0.451	
				(1.778)	(2.784)	
2025-S1				**-10.152*****	-0.180	
				(1.840)	(2.823)	
OpenAI * 2023-S2					-11.043***	
					(4.097)	
OpenAI * 2024-S1					-13.687***	
					(3.945)	
OpenAI * 2024-S2					**-15.493*****	
					(3.713)	
OpenAI * 2025-S1					**-13.061*****	
					(4.625)	
Constant	-3.577***	29.187***			22.693***	
	(0.967)	(3.034)			(4.170)	
Provider FE	No	No	Yes	Yes	No	No
Year-month FE	No	No	Yes	No	No	Yes
Observations	3,614	3,614	3,614	3,614	3,614	3,614
R^2	0.049	0.093	0.078	0.082	0.108	0.094
Adjusted R^2	0.049	0.086	0.062	0.072	0.105	0.086
F Statistic	93.997*** (df = 2; 3611)	92.040*** (df = 4; 3585)	75.162*** (df = 4; 3552)	40.052*** (df = 8; 3572)	31.183*** (df = 14; 3599)	74.010*** (df = 5; 3584)

Note: *p<0.1; **p<0.05; ***p<0.01. The unit of observations is AI models, available at different cloud providers (Provider) and observed during every month between January 2023 and January 2025. The quality index is rescaled between 0 and 100.

Overall, several tendencies documented in this section shows that a large variety of models are accessible from a growing number of cloud providers; model quality has risen and quality adjusted prices have fallen substantially, to the benefit of potential AI users; the leading developers' models have become available at ever lower prices over time, after an initial phase when they were offered at higher price than those of

competitors.[45] However, given the speed of evolution in AI markets, the potential market power of cloud incumbents in adjacent markets and the scale of their investments, risks for competition in this segment persist and prices and quality should continue to be monitored.

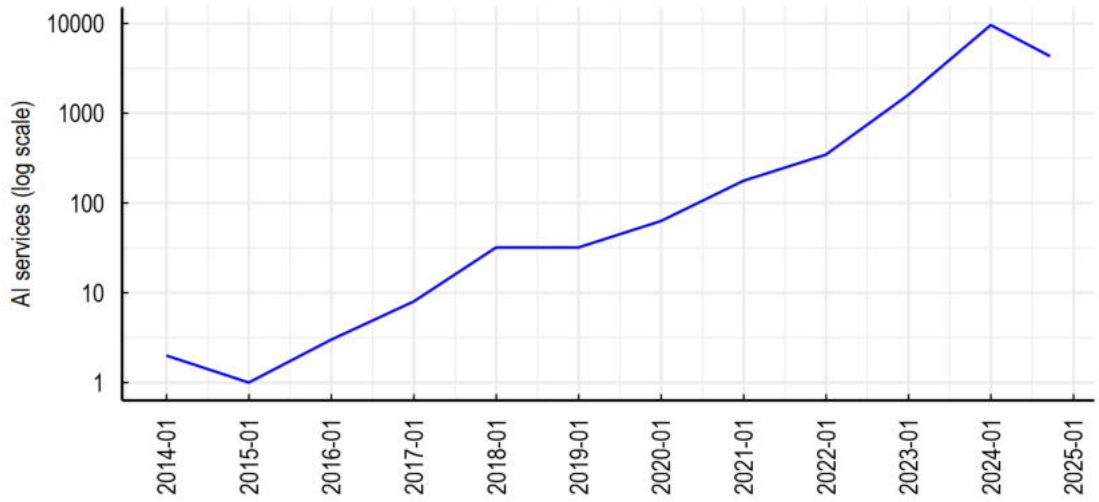
3.3. AI-powered digital services and popularity across sectors

Downstream in the AI stack (Figure 1) consumers and workers do not interact with AI but with digital service applications. The current dominant AI paradigm features AI-powered services built around "AI agents" that combine several AI models and software wrapped in a user interface accessible from the web browser or mobile device applications (i.e. Chatbots). To track the downstream diffusion of AI accessible to the final user this paper tracks AI-powered digital services available in the market.[46] The main source of data for this analysis is collected from *There's an AI for That* (TAAFT).

The number of downstream AI-powered services has increased exponentially since 2014 (Figure 15) with the number and popularity of AI tasks evolving as the technology improved. In 2020, AI tools were concentrated on tasks like music creation (119 AI tools), customer support (94), image generation (58), text-to-speech (60), image editing (56) and data analysis (47). In March 2024, image generation (6810), content generation (5135), writing (3515), chatting (3259), customer support (2579) and data analysis (2507) were the most widely supplied AI tasks, reflecting the boom of LLMs and AI vision models.

Figure 15. The growing number of AI-powered services

Number of AI-powered digital services (applications) released per year, log scale.



Note: Number of AI-powered services per release date available on TAAFT platform.
Source: Authors' calculations based on data from There's an AI For That (TAAFT, 2024).

In line with AI's general-purpose technology nature, Figure 16 confirms that downstream applications indeed can impact a broad variety of sectors, through assisting or carrying out specific tasks. They range from chatbots and other digital assistants, object detection, music search, text translations, to the

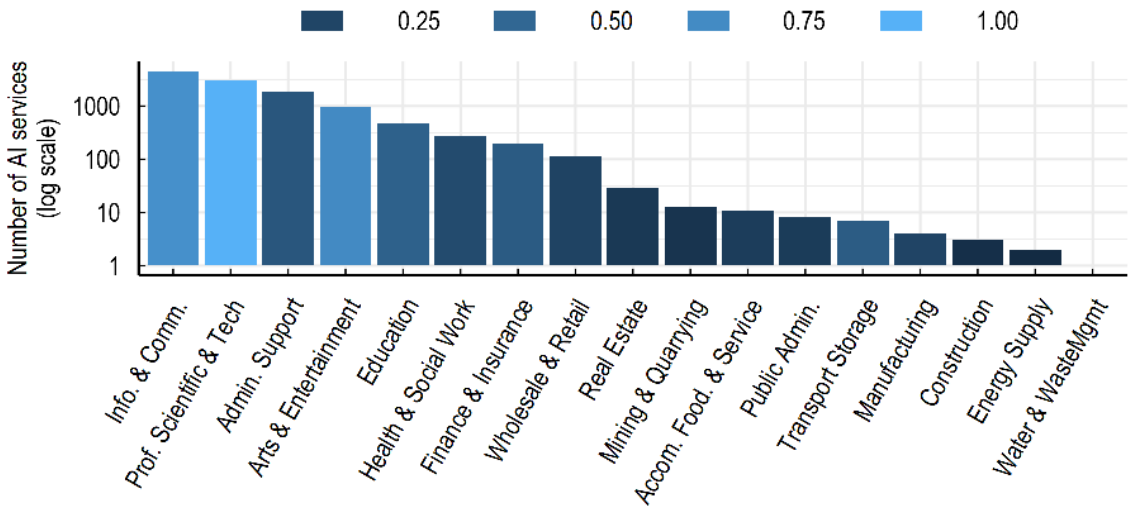
[45] Recent discussions on the role of test time compute underscore the stark technical and economic difference between computing power used for *pretraining* models or for the *inference* (using) of the models. In this paper, we do not discuss the market power of cloud providers in pretraining, but focus on the inference (usage). The latter is often perceived to be the main source of computing power needs in a scenario of broad base adoption and use of AI.

[46] AI-powered services are digital services based on any AI (text, vision, audio). It is dominated by AIs based on the *Transformer* and *Diffusion* architecture but also includes other AI technologies.

presentation of slides, essay grading, education mentoring, or academic search. However, most of these tasks are among knowledge intensive services (professional, technical and ICT services), in line with the studies that focus on AI exposure from a task-based perspective (Eloundou et al., 2024).

Figure 16. There is diverse supply of AI services across several sectors

The number of available AI-powered digital services and their popularity, by detailed economic sectors



Note: The figure shows the number of AI services (x-axis) and their popularity (normalized number of clicks in colour) for AI tasks available on the platform (x-axis) and classified by economic sectors (ISIC level 4). Popularity is normalised to range between 0 and 1.
Source: Authors' calculations based on data from There is an AI For That (TAAFT).

As mentioned in the previous section, the market for AI-powered services is still emerging, with many unknowns surrounding final applications, business models and long-term adoption. AI is potentially a singular technology that has the potential to shake existing competitive positions, particularly where network effects and data feedback loops are strong. Early evidence regarding the business models of AI-powered services and their data collection policy (La Malfa et al., 2024) may alleviate concerns about a race to collect personal customer data.[47]

Across our sample, on average, around 30% of AI services have free (entry-level) subscription plans, and less than 10% of services are free of charge (6% in *Professional. scientific and technical activities*, but 16% in *Art, Entertainment and Recreation*). Table 3 presents a few examples for these among a few well-known applications, such as ChatGPT of Open AI, Claude from Anthropic or Mistral AI among multimodal chatbots, Perplexity AI for web search or MidJourney for image generation. Beyond free options, there is a strong presence of paid ones too, suggestive of business strategies other than counting on two-sided market features to generate revenue (where the incentives to subsidise the side of the market that generates the strongest adoption externalities should drive the price to zero; Hagiu and Wright, 2023). Tracking the evolution of this indicator for the more popular services across sectors could provide a valuable measure of the relative value of accumulating user data.

As of September 2024, the market for downstream AI-powered services appears dynamic and evolves rapidly, with more than 12 182 services available, with the highest popularity for services from tech

[47] For instance, a data feedback loop from user data may not be the determinant driver of higher quality models; similarly, while data collection from the user interface (i.e. ChatGPT) is explicit when using the web browser, it provides an opt-out option. For clients using the API, providers usually mention explicitly that they do not collect user data for further training of the model, but those policies may evolve rapidly.

incumbents[48] (Table 3) or (very) small new entrants. The market for downstream applications is far from being settled, given that the search for a "killer app" is still ongoing. So far, clear business models and uses have yet to be established. AI has also acted as a potential disruptor in tech markets, allowing new players like OpenAI to challenge platform incumbents and other players in the value chain to capture a greater part of the value-added of the digital sector (i.e. Nvidia and other chipmakers; Varas et al., 2021).

Nonetheless, concerns that new entrants today could become powerful incumbents tomorrow are still important and evolutions should be monitored closely. The early phases of a technological revolution are always characterised by a myriad of new start-ups, applications and business models. However, only a few of those services and companies will reach a sustainable market share and will be able to consolidate their positions (Draghi, 2024). In addition, potential acquisitions of new players by incumbents in some specific markets (Gautier and Lamesch, 2021; Carugati, 2023; Katz, 2021) could be an important threat to competition in the coming years, all the more as the many successful new entrants are largely financed by the same large tech companies that compete with them.

Table 3. Subscription prices of selected popular AI-powered services

Company	Is free access available?	"Individual" Subscription with most basic functions ($/month)	"Business" subscription with more expanded capabilities ($/month)	"Company" Subscription with most capabilities and large-scale use ($/month)	Type of Service
Anthropic Claude	Yes	$18	$25	-	Multimodal Chatbot
OpenAI ChatGPT	Yes	$20	$200	-	Multimodal Chatbot
Mistral AI - Lechat	Yes	$15.5	-	Custom	Multimodal Chatbot
Google One AI Premium	1 month	$20	-	-	Multiple AI Products
Clarifai	Yes	$30	-	$300	Image/Multimodal
QuillBot	Yes	$8.33	-	-	Text Generation
Adobe Creative Cloud	No	$6.2	-	-	Text Editing
Perplexity AI	Yes	$20	-	-	Websearch
Browse AI	Yes	$19	$99	$249	Web Scraping
MidJourney	No	$10	$30	$60	Image Generation
Otter AI	Yes	$8.33	$20	-	Audio Generation
Synthesia	Yes	$16	$58	-	Video Generation
GitHub Copilot	No	$10	$19	$39	Coding Assistant
Cursor	Yes	$20	$40	-	Coding Assistant
Hugging Face	Yes	$9	$25	Custom	AI Models and APIs
Microsoft Copilot Pro	No	$22	-	-	Productivity Assistant
Salesforce Sales Cloud	No	-	$80	$165	CRM with AI Capabilities
Smartsheet	No	$8	$17	-	Project Management

Note: Data collected on the 12 February 2025.
Source: Author's calculations.

[48] Among the popular AI tools, specific GPTs built on top of ChatGPT from OpenAI attract a significant part of the traffic on the TAAFT (TAAFT, 2024).

4. Concluding discussion on risks, policies and future analysis

Competitive AI markets, which drive lower prices and improved model quality, are essential for widespread AI adoption across diverse tasks. This, in turn, is critical for realising significant macroeconomic and welfare gains. To shed more light on the evolution of AI markets, this paper proposes a preliminary and partial assessment – focusing mostly on the supply side of markets – by collecting and analysing novel data regarding three key segments of the AI value chain over the past two years until February 2025.

During this period, markets have shown signs of dynamism in these market segments. Upstream, in *AI model development*, about ten companies are forming the AI Economic Frontier (capturing the best price-quality offerings) in large language models, with around ten additional AI labs competing in niche segments of the markets (audio transcription, image generation). This diversity leverages the specific know-how of developers and allows for serving heterogeneous user preferences. As a further sign of dynamic and innovative markets, quality-adjusted prices have dropped by about 80% during the past two years, due to a combination of continuously rising model capabilities and falling model prices. Nevertheless, price wars in the early days of technological revolutions are common, and can be detrimental to competition in the medium term if only the actors with the most resources can sustain the shrinking of revenues.

The contribution of open-source AI models has been significant in creating such a vibrant market environment and advancing AI research, fostering transparency and leading to faster innovation and diffusion. Indeed, open weight models have now largely closed the gap with closed solutions in terms of performance and have been a major pro-competitive force in the AI landscape. A continued lively open-source AI community involving a diversity of companies is likely to be an important factor in enabling and sustaining vibrant AI markets.

Further downstream, the *AI cloud provision* market shows signs of having several active players. Large cloud incumbents (hyperscalers) possess several competitive advantages, including their strong presence in traditional cloud services, the scale of their infrastructure dedicated to AI for training models and a large existing customer base. However, although there is a lack of data on their effectiveness in capturing users, our results show that smaller cloud providers also offer high quality open-source AI models at attractive prices, thanks to innovations in hardware and by providing specific and optimised offers. Optimising across models and providers with different offers seems relatively smooth today, also reflected in the possibility of multi-homing in many applications.

Going forward, several uncertainties remain, related to technical innovations, energy use intensity, and AI adoption preferences and most importantly the critical gatekeeping position of a few players (Draghi, 2024; Carugati, 2023; Pilz and Heim, 2023; Kowalski, Volpin and Zombori, 2024). Gatekeeping may yet be a hurdle to competition in AI, similarly to other digital markets. Business strategies by AI suppliers to lock in users, through product bundling and self-preferencing, could limit the ability of AI adopters to rely on the best offerings for their production processes. Signs for gatekeeping behaviour by leading cloud providers should be monitored, especially in the light of evolving partnerships with some leading AI developers. Many companies in the AI supply chain operate in several connected segments (hardware producers like Nvidia, cloud providers like Amazon and platform services like Meta or Google, see Box 1) where losses in one segment (in this case, AI) combined with strong profits in another segment (in non-AI segments where they have more market power) can contribute to sustained overall market power and result in *de facto* entry barriers. The potential (horizontal) consolidation of specialised companies will also contribute to the evolution of the competition landscape, with acquisitions from already large players carrying larger risks. Ensuring the availability of transparent information on the price and performance of model offerings, and ensuring low barriers to switch across developers and providers will also be critical.

Downstream, among AI-powered applications, AI may have allowed entrants to challenge incumbents in new markets (i.e. *OpenAI* or *Anthropic* in AI chatbots) AI tools are provided by a variety of companies with many startups and incumbents offering diverse applications assisting in a range of tasks. Nevertheless,

most of the apps are concentrated in a few activities within the sectors *Information and Communication* and *Professional. scientific and technical activities*, echoing previous evidence on the exposure of occupations and sectors to AI (Eloundou et al, 2024). Moreover, important risks for competition persist. First, the capacity for new entrants to displace incumbents appears highly uncertain. Second, large incumbents are strongly present in many layers of the AI value chain and can leverage their market power in other markets (Nicoletti, Vitale and Abate, 2023), for example by building *AI app platforms* generating indirect network effects and squeezing the margins of other, smaller app developers. Third, AI embedded in hardware could give a particular advantage to companies with large existing user bases, such as Apple, Google or Microsoft. Monitoring the downstream segment is important going forward, especially in critical sectors like defense, finance, education or health.

Continued monitoring of the market evolution (number of players, prices, quality and market shares) will provide early indicators of the state of competition, although they should be complemented by broader assessments (e.g. strategic partnerships between market players; CMA, 2024; OECD, 2022). Several avenues for future research could help to better track developments in areas that may affect how competitive the market for AI development remains. First, the future dynamism of the open-source ecosystem, given its beneficial role in nurturing competition and innovation in several market segments (Blind et al., 2021); second, the future of the "scaling law" (i.e. that model quality rises predictably as compute and data increase (Hoffmann et al., 2022; Sevilla et al., 2024) and the need for ever larger computing capacity (Cottier et al., 2024, Sastry et al., 2024), especially given the tensions arising from intensive energy use (Sasha Luccioni, Yacine and Strubell, 2024); third, the question of access, usefulness, and regulation of human-generated data versus AI-generated (synthetic) data (Villalobos et al., 2024; Longpre et al., 2024; Samuelson, 2023; Martens, 2024); fourth, the evolution of financial conditions to support the industry and the long-term sustainability of the current boom in infrastructure investments in data centres.

Finally, risks for competition in specific sectors could well be amplified by AI adoption, but the aggregate impact is yet unknown. The current pattern of AI diffusion suggests that the risks for competition for AI largely concern sectors where concentration is already high (Babina et al., 2024). Whether AI will contribute to the global trend of concentration observed in the last decades (Calligaris et al., 2024), particularly in the United States (Covarrubias, Gutiérrez and Philippon, 2020), remains an open question. On the one hand, AI could raise concentration in currently competitive sectors, if it allows early adopters to increase and consolidate their market shares. On the other hand, AI could lower concentration by bringing new competition in currently highly concentrated markets, particularly in the ICT sector. How AI adoption itself will impact competition and market power in the broader economy will be examined in upcoming planned work relying on micro-level data that combines information on the performance of firms with AI intensity metrics.

References

Artificial Analysis (2024), Independent analysis of AI models and API providers, https://artificialanalysis.ai/, https://artificialanalysis.ai/.

Artificial Analysis (2025 Q1), State of AI: China, https://artificialanalysis.ai/downloads/china-report/2025/Artificial-Analysis-State-of-AI-China-Q1-2025.pdf

Angeletos, G., G. Lorenzoni and A. Pavan (2022), "Wall Street and Silicon Valley: A Delicate Interaction", *The Review of Economic Studies*, Vol. 90/3, pp. 1041-1083, https://doi.org/10.1093/restud/rdac043.

Anthony Quentin, S. Biderman and H. Schoelkopf (2023), "Transformer Math 101", blog.eleuther.ai.A

Artificial Analysis (2024), Independent analysis of AI models and API providers, https://artificialanalysis.ai/, https://artificialanalysis.ai/.

Autorité de la Concurrence (2024), "Opinion 24-A-05 of 28 June 2024 on the competitive functioning of the generative artificial intelligence sector", Paris,

https://www.autoritedelaconcurrence.fr/en/opinion/competitive-functioning-generative-artificial-intelligence-sector

Azoulay, P., J. Krieger and A. Nagaraj (2024), Old Moats for New Models: Openness, Control, and Competition in Generative AI, *National Bureau of Economic Research*, Cambridge, MA, https://doi.org/10.3386/w32474.

Babina, T. et al. (2024), "Artificial intelligence, firm growth, and product innovation", *Journal of Financial Economics*, Vol. 151, p. 103745, https://doi.org/10.1016/j.jfineco.2023.103745.

Bai, G. and J. Liu (2024), "MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues", https://arxiv.org/abs/2402.14762.

Bajari, P. et al. (2019), "The Impact of Big Data on Firm Performance: An Empirical Investigation", *AEA Papers and Proceedings*, Vol. 109, pp. 33-37, https://doi.org/10.1257/pandp.20191000.

Ben-Ishai, G. et al. (2024), "AI and the Opportunity for Shared Prosperity: Lessons from the History of Technology and the Economy", arXiv preprint arXiv:2401.09718.

Bergemann, D., A. Bonnatti and A. Smolin (2025), "The economics of large language models: token allocation, finet-tuning, and optimal pricing", arxiv:250207736v1

Blind, K. et al. (2021), The impact of open source software and hardware on technological independence, competitiveness and innovation in the EU economy – Final study report, European commission, https://doi.org/doi/10.2759/430161.

Bommasani, R. et al. (2021), “On the opportunities and risks of foundation models”, arXiv preprint arXiv:2108.07258.

Bommasani, R. et al. (2023), “The foundation model transparency index”, arXiv preprint arXiv:2310.12941.

Business at OECD (2024), “Artificial Intelligence, Data and Competition – Note by BIAC”, Comments by the Business at OECD (BIAC) Competition Committee to the OECD Competition Committee, https://www.businessatoecd.org/hubfs/Artificial%20Intelligence%2C%20Data%2C%20and%20Competition.pdf?hsLang=en.

Calligaris, S. et al. (2024), Exploring the evolution and the state of competition in the EU, European Commission.

Calvano, E. and M. Polo (2021), “Market power, competition and innovation in digital markets: A survey”, *Information Economics and Policy*, Vol. 54, p. 100853, https://doi.org/10.1016/j.infoecopol.2020.100853.

Carugati, C. (2023), The competitive relationship between cloud computing and generative AI, https://www.bruegel.org/working-paper/competitive-relationship-between-cloud-computing-and-generative-ai (accessed on 23 August 2024).

CEA (2024), Economic Report to the President, https://www.whitehouse.gov/wp-content/uploads/2024/03/ERP-2024.pdf (accessed on 23 August 2024).

CEA (2025), AI talent report, https://bidenwhitehouse.archives.gov/cea/written-materials/2025/01/14/ai-talent-report/

CEPR (2023), “GEN AI & Market Power: What Role for Antitrust Regulators?”, CEPR Competition Policy RPN Webinar Transcrip, https://cepr.org/system/files/2023-07/Competition%20Policy%20RPN%20Webinar_Gen%20AI%20%26%20Market%20Power_12%20July%202023_Transcript.pdf (accessed on 6 September 2024).

Chardon-Boucaud, S., A. Dozias and C. Gallezot (2024), “The Artificial Intelligence Value Chain: What Economic Stakes and Role for France?”, *Trésor-Economics*, No. 354, Direction générale du Trésor, Paris, https://www.tresor.economie.gouv.fr/Articles/00be60a5-dea5-437f-8a49-b44b5663ded2/files/3a04af78-d4d3-4de4-b24b-d43653a55f9a

Chesbrough, H. (2023), Measuring the Economic Value of Open Source: A Survey and a Preliminary Analysi, The Linux Foundation.

Chiang, W. et al. (2024), “Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference”, https://arxiv.org/pdf/2403.04132.

CMA (2024), AI Foundation Models: technical update report, Competition and Market Authority.

Coeuré, B. (2024), Comments on “The simple macroeconomics of transformative AI” by Daron Acemoglu, https://www.autoritedelaconcurrence.fr/sites/default/files/2024-04/Economic%20Policy%20Panel%20Comments%20-%20BCoeur%C3%A9.pdf (accessed on 6 September 2024).

Cottier, B., T. Besiroglu and D. Owen (2023), “Who is leading in AI? An analysis of industry AI research”, arXiv preprint arXiv:2312.00043, https://doi.org/10.48550/arXiv.2312.00043 (accessed on 22 July 2024).

Cottier, B. et al. (2024), “The rising costs of training frontier AI models”, arXiv preprint arXiv:2405.21015, https://arxiv.org/abs/2405.21015 (accessed on 22 July 2024).

Covarrubias, M., G. Gutiérrez and T. Philippon (2020), “From Good to Bad Concentration? US Industries over the Past 30 Years”, *NBER Macroeconomics Annual*, Vol. 34, pp. 1-46, https://doi.org/10.1086/707169.

Draghi (2024), The future of European Competitiveness, Part B, In-depth analysis and recommendations.

EC-CMA-DOJ-FTC (2024), “Joint Statement on Competition in Generative AI Foundation Models and AI Products”, European Commission, U.K. Competition and Markets Authority, U.S. Department of Justice, U.S. Federal Trade Commission, https://competition-policy.ec.europa.eu/about/news/joint-statement-competition-generative-ai-foundation-models-and-ai-products-2024-07-23_en.

Economist, T. (2024), The war for AI talent is heating up, https://www.economist.com/business/2024/06/08/the-war-for-ai-talent-is-heating-up.

Eloundou, T. et al. (2024), “GPTs are GPTs: Labor market impact potential of LLMs”, Science, Vol. 384/6702, pp. 1306-1308, https://doi.org/10.1126/science.adj0998.

Erdil, E. (2024), Optimally Allocating Compute Between Inference and Training, https://epochai.org/blog/optimally-allocating-compute-between-inference-and-training (accessed on 22 July 2024).

Erdil, E. (2024), Frontier language models have become much smaller, https://epoch.ai/gradient-updates/frontier-language-models-have-become-much-smaller.

Filippucci, F. et al. (2024), “The impact of Artificial Intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges”, *OECD Artificial Intelligence Papers*, No. 15, OECD Publishing, Paris, https://doi.org/10.1787/8d900037-en.

Filippucci, F., P. Gal and M. Schief (2024), “Miracle or Myth? Assessing the macroeconomic productivity gains from Artificial Intelligence”, *OECD Artificial Intelligence Papers*, No. 29, OECD Publishing, Paris, https://doi.org/10.1787/b524a072-en.

Frymire, L. and D. Owen (2025), Training compute growth is driven by larger clusters, longer training, and better hardware, https://epoch.ai/data-insights/training-compute-decomposition.

FTC (2025), Partnerships between cloud services providers and AI developers.

Gans, J. (2024), Copyright Policy Options for Generative Artificial Intelligence, National Bureau of Economic Research, Cambridge, MA, https://doi.org/10.3386/w32106.

Gans, J. (2024), Market Power in Artificial Intelligence, National Bureau of Economic Research, Cambridge, MA, https://doi.org/10.3386/w32270.

Gautier, A. and J. Lamesch (2021), “Mergers in the digital economy”, Information Economics and Policy, Vol. 54, p. 100890, https://doi.org/10.1016/j.infoecopol.2020.100890.

Giovannetti, E. and P. Siciliani (2023), “Platform Competition and Incumbency Advantage under Heterogeneous Lock-in effects”, Information Economics and Policy, Vol. 63, p. 101031, https://doi.org/10.1016/j.infoecopol.2023.101031.

Hagiu, A. and J. Wright (2025), “Artificial intelligence and competition policy”, International Journal of Industrial Organization, https://doi.org/10.1016/j.ijindorg.2025.103134.

Hagiu, A. and J. Wright (2023), "Data-enabled learning, network effects, and competitive advantage", The RAND Journal of Economics, Vol. 54/4, pp. 638-667, https://doi.org/10.1111/1756-2171.12453.

Hendrycks, D. et al. (2020), "Measuring Massive Multitasks language understanding", arXiv prepreint arXiv:2009.03300.

Ho, A. et al. (2024), "Algorithmic progress in language models", arXiv preprint arXiv:2403.05812, https://doi.org/10.48550/arXiv.2403.05812 (accessed on 22 July 2024).

Hoffmann, J. et al. (2022), "Training compute-optimal large language models", arXiv preprint arXiv:2203.15556.

Hoffmann, M., F. Nagle and Y. Zhou (2024), "The Value of Open Source Software", SSRN Electronic Journal, https://doi.org/10.2139/ssrn.4693148.

Hunt, S. et al. (2023), You Are What You Eat: Nurturing Data Markets to Sustain Healthy Generative AI Innovation, CPI Tech Reg Chronicle, Vol 1, https://www.keystone.ai/wp-content/uploads/2023/11/NURTURING-DATA-MARKETS-TO-SUSTAIN-HEALTHY-GENAI-INNOVATION-Stefan-Hunt-Wen-Jian-Aman-Mawar-Bartley-Tablante-2.pdf (accessed on 23 August 2024).

Jullien, B. and W. Sand-Zantman (2021), "The Economics of Platforms: A Theory Guide for Competition Policy", Information Economics and Policy, Vol. 54, p. 100880, https://doi.org/10.1016/j.infoecopol.2020.100880.

Kaplan, J. et al. (2020), "Scaling laws for neural language models".

Katz, M. (2021), "Big Tech mergers: Innovation, competition for the market, and the acquisition of emerging competitors", *Information Economics and Policy*, Vol. 54, p. 100883, https://doi.org/10.1016/j.infoecopol.2020.100883.

Korinek, A. and J. Vipra (2025), "Concentrating intelligence: scaling and market structure in artificial intelligence", *Economic Policy*, Volume 40, Issue 121, January 2025, Pages 225–256, https://doi.org/10.1093/epolic/eiae057.

Kowalski, K., C. Volpin, and Z. Zombori (2024), "Competition in Generative AI and Virtual Worlds", Competition Policy Brief, European Commission, Brussels, https://competition-policy.ec.europa.eu/document/download/c86d461f-062e-4dde-a662-15228d6ca385_en.

La Malfa, E. et al. (2024), "Language-Models-as-a-Service: Overview of a New Paradigm and its Challenges", *Journal of Artificial Intelligence Research*, Vol. 80/2024, pp. 1497-1523.

Lerner, J. and J. Tirole (2002), "Some simple economics of open source", *The Journal of Industrial Economics*, Vol. 50/2, pp. 1997-234.

Liesenfeld, A. and M. Dingemanse (2024), "Rethinking open source generative AI: open washing and the EU AI Act", *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, https://doi.org/10.1145/3630106.3659005.

Liu, Y. and H. Wang (2024), "Who on earth is using generative AI?", *World Bank, Policy research Working paper*.

Longpre, S. et al. (2024), "Consent in Crisis: The Rapid Decline of the AI Data Commons", arXiv 2407.14933.

Lorenz, P., K. Perset and J. Berryhill (2023), "Initial policy considerations for generative artificial intelligence", *OECD Artificial Intelligence Papers*, No. 1, OECD Publishing, Paris, https://doi.org/10.1787/fae2d1e6-en.

Macropolo (2023), Global AI talent tracker, https://archivemacropolo.org/interactive/digital-projects/the-global-ai-talent-tracker/.

Martens, B. (2024), "Economic arguments in favour of reducing copyright protection for generative AI inputs and outputs", Bruegel, https://www.bruegel.org/working-paper/economic-arguments-favour-reducing-copyright-protection-generative-ai-inputs-and (accessed on 23 August 2024).

Martens, B. (2025), "How DeepSeek has changed artificial intelligence and what it means for Europe", Bruegel Policy Brief, https://www.bruegel.org/policy-brief/how-deepseek-has-changed-artificial-intelligence-and-what-it-means-europe (accessed on 17 March 2025).

Nagle, F. (2019), "Open Source Software and Firm Productivity", Management Science, Vol. 65/3, pp. 1191-1215, https://doi.org/10.1287/mnsc.2017.2977.

Nagle, F. (2018), "Learning by Contributing: Gaining Competitive Advantage Through Contribution to Crowdsourced Public Goods", *Organization science*, Vol. 29/4, pp. 569-587.

Nicoletti, G., C. Vitale and C. Abate (2023), "Competition, regulation and growth in a digitized world: Dealing with emerging competition issues in digital markets", OECD Economics Department Working Papers, No. 1752, OECD Publishing, Paris, https://doi.org/10.1787/1b143a37-en.

OECD.AI (2024), OECD.AI - Evolution of new AI models, https://oecd.ai/en/data?selectedArea=ai-models-and-datasets&selectedVisualization=evolution-of-new-ai-models.

OECD (2024), "Artificial intelligence, data and competition", *OECD Artificial Intelligence Papers*, No. 18, OECD Publishing, Paris, https://doi.org/10.1787/e7e88884-en.

OECD (2023a), "Competition and Innovation: A Theoretical Perspective", *OECD Roundtables on Competition Policy Papers*, No. 294, OECD Publishing, Paris, https://doi.org/10.1787/4632227c-en.

OECD (2023b), Recommendation of the Council on Artificial Intelligence, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

OECD (2022a), "The Evolving Concept of Market Power in the Digital Economy", *OECD Roundtables on Competition Policy Papers*, No. 278, OECD Publishing, Paris, https://doi.org/10.1787/2cfcb4a8-en.

OECD (2022b), "Measuring the value of data and data flows", *OECD Digital Economy Papers*, No. 345, OECD Publishing, Paris, https://doi.org/10.1787/923230a6-en.

OECD.AI (2024), OECD.AI - Evolution of new AI models, https://oecd.ai/en/data?selectedArea=ai-models-and-datasets&selectedVisualization=evolution-of-new-ai-models.

Open Source Initiative (2024), Open source AI, https://opensource.org/deepdive.

Open Source Initiative (2024), Open Source Definition, https://opensource.org/osd.

Open-Weights-definition (2024), https://github.com/Open-Weights/Definition.

Pilz, K. and L. Heim (2023), "Compute at Scale--A Broad Investigation into the Data Center Industry", arXiv preprint arXiv:2311.02651.

Pope, R. et al. (2022), "Efficiently scaling transformer inference".

Portuguese Competition Authority (2024), "Competition and generative AI: opening AI models", AdC

Short Papers, https://www.concorrencia.pt/sites/default/files/processos/epr/AI%20short%20paper%20-%20Opening%20AI%20models%20-%20EN.pdf.

Rahman, R. (2024), Performance per dollar improves around 30% each year, https://epoch.ai/data-insights/price-performance-hardware.

Rein, D. (2023), "GPQA: A Graduate-Level Google-Proof Q&A Benchmark", https://arxiv.org/abs/2311.12022.

Samuelson, P. (2023), "Generative AI meets copyright", Science, Vol. 381/6654, pp. 158-161, https://doi.org/10.1126/science.adi0656.

Sardana, N. et al. (2024), "Beyond Chinchilla-Optimal: Accounting for inference in Language Model Scaling Laws".

Sasha Luccioni, A., J. Yacine and E. Strubell (2024), "Power Hungry Processing: Watts Driving the Cost of AI Deployment?", arXiv preprint arXiv:2311.16863.

Sastry, G. et al. (2024), "Computing Power and the Governance of Artificial Intelligence", arXiv preprint arXiv:2402.08797.

Schaal, E. and M. Taschereau-Dumouchel (2023), "Herding through booms and busts", *Journal of Economic Theory*, Vol. 210, p. 105669, https://doi.org/10.1016/j.jet.2023.105669.

Sevilla, J. and E. Roldán (2024), Training Compute of Frontier AI Models Grows by 4-5x per Year, https://epochai.org/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year.

Solaiman, I. (2023), "The Gradient of Generative AI Release: Methods and Considerations", arXiv preprint arXiv:2302.04844.

Schrepel, T. and A. 'Sandy' Pentland, (2024), "Competition between AI foundation models: dynamics and policy recommendations", Industrial and Corporate Change, 2024; https://doi.org/10.1093/icc/dtae042.

Srivastava, A. et al. (2022), "Beyond the imitation game: quantifying and extrapolating the capabilities of language models", ArXiv preprint arXiv:2206.04615.

Stanford University (2024), Artificial Intelligence Index Report 2024.

Syverson, C. (2019), "Macroeconomics and Market Power: Context, Implications, and Open Questions", Journal of Economic Perspectives, Vol. 33/3, pp. 23-43, https://doi.org/10.1257/jep.33.3.23.

TAAFT (2024), There is an AI For That.

Varas, A. et al. (2021), Strengthening the Global Semiconductor Supply Chain in an Uncertain Era, BCG.

Varian, H. (2021), "Seven deadly sins of tech?", Information Economics and Policy, Vol. 54, p. 100893, https://doi.org/10.1016/j.infoecopol.2020.100893.

Villalobos, P. et al. (2024), "Will we run out of data? Limits of LLM scaling based on human-generated data", https://arxiv.org/abs/2211.04325 (accessed on 22 July 2024).

Vipra, J. and S. Myers West (2023), Computational power and AI, AI Now Institute.

White, C. (2024), "LiveBench: A Challenging, Contamination-Free LLM Benchmark",

https://arxiv.org/abs/2406.19314.

White, M. et al. (2024), “The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI”, arXiv preprint arXiv:2403.13784.

You, J. and D. Owen (2024), Leading AI companies have hundreds of thousands of cutting-edge AI chips, https://epoch.ai/data-insights/computing-capacity.

Zheng et al (2024), “Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference”, https://arxiv.org/html/2403.04132v1.

Annex A. Glossary

AI-adopting firms: Companies that integrate artificial intelligence into their operations to improve efficiency, innovation, and competitiveness. They may use AI models through cloud services or deploy them internally.

AI agent: refers to an autonomous entity designed to perceive its environment through sensors and act upon that environment through diverse modalities (text, image, audio) to achieve specific goals in multiple steps. Unlike traditional software, which follows predefined instructions, AI agents use machine learning algorithms to learn from data, adapt to new situations, and make decisions.

AI developing companies: companies that (pre)train and finetune AI models that are sold on a market for the use of other companies or individual developers. Only companies that offer the use of AI models for their use from their website, from a cloud provider or from an open-source platform are considered AI developing companies.

AI lifecycle: The set of stages in the development of artificial intelligence, ranging from data collection and model training to deployment, usage, and continuous improvement.

AI modality: The type of data or tasks a given AI model is designed to work with. Modalities can include text, images, audio, or a combination of these (multimodal).

AI-powered digital firms: Companies that use artificial intelligence as the main engine to deliver innovative digital services or products.

AI Provider: A company that hosts and provides access to one or more model endpoints via an API. An endpoint is a hosted instance of a model. Using AI through a cloud provider is considered serverless (for the consuming firm), where the consumer only pays per usage (generally in dollars per token). AI provision is a B2B intermediary that serves AI models (from companies developing AI) to companies that use AI as an intermediary input to power services for final consumption.

Application programming interface (API): A form of software interface that acts as a conduit for computer programs to interact and interoperate.

Benchmarks: Standardized tests used to evaluate and compare the performance of artificial intelligence models.

Among the main benchmarks are:

- **MMLU** (Massive Multitask Language Understanding): MMLU is a benchmark designed to evaluate the performance of language models across a wide range of academic subjects. It includes a diverse set of multiple-choice questions that test knowledge and reasoning abilities. The benchmark was introduced by Hendrycks et al. (2020) in the paper "Measuring Massive Multitask Language Understanding".
- **Arena Elo**: Arena Elo is a benchmark platform that uses the Elo rating system to evaluate the performance of large language models (LLMs) based on pairwise comparisons. It allows for dynamic and scalable evaluation of models in a crowdsourced manner. The Arena Elo system is discussed in the paper "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference" by Zheng et al (2024). See the live benchmarking of models on the Chatbot Arena GitHub page (https://github.com/lmsys/chatbot-arena).

- **MT-Bench** (Multi-Turn Benchmark): MT-Bench is designed to evaluate the capabilities of LLMs in multi-turn conversations, focusing on coherence, informativeness, and engagement over multiple exchanges. The MT-Bench was introduced in the paper "MT-Bench: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues" by (Bai and Liu, 2024).
- **Livebench:** is a dynamic benchmark that focuses on providing contamination-free evaluations for LLMs. It includes recently released questions and updates monthly to ensure the relevance and difficulty of the benchmark. It is detailed in the paper "LiveBench: A Challenging, Contamination-Free LLM Benchmark" by White et al. (2024).
- **GPQA** (Graduate-Level Google-Proof Q&A Benchmark): is a challenging dataset designed to evaluate the capabilities of LLMs in answering complex, graduate-level questions in biology, physics, and chemistry. The questions are crafted to be "Google-proof," meaning they cannot be easily answered with a web search. The GPQA benchmark was introduced by Rein et al. (2023).

In addition to MMLU, ArenaElo, MT-Bench, Livebench, and GPQA, other major AI benchmarks include Big Bench, HELM, AGIEval, BBH, HumanEval, MBPP, GLUE, SuperGLUE, SQuAD, and the Winograd Schema Challenge, each evaluating different aspects of AI model performance across various tasks and domains.

Big data: Data generated as a by-product of the mass consumption of digital products/services.

Context window: The maximum number of combined output and input tokens that can be queried to an AI model.

Cloud providers: Companies that provide computing services, offering users remote access to computing resources, including computing power, data storage and artificial intelligence models, via the Internet.

Computing power: The amount of computing power required to train and operate artificial intelligence models, generally measured in terms of computational capacity (in Flops*: Floating-Point Operations Per Second*).

Costs of computing: The costs associated with accessing and using cutting-edge computing technologies for training and executing the most advanced artificial intelligence models. This includes expenses for specialized hardware, such as GPUs, as well as energy costs and maintenance of the infrastructure required to support these computation-intensive operations.

Data clusters: Sets of interconnected servers that store and process large amounts of data. In the context of AI, these clusters are essential for managing the massive volumes of data required for training and inference of models, offering high computational capacity and redundancy to ensure the reliability and performance of operations.

Data feedback loops: A mechanism by which data generated by an artificial intelligence system or application is fed back into the system to improve its performance.

Diffusion: A specific architecture used in artificial intelligence models for image generation. This method is employed in some image generation models, where an image is progressively generated or refined from an initial noise by applying successive diffusion steps.

Digital gatekeepers: Companies or digital platforms that control access to digital markets or online services. These digital gatekeepers, often tech giants, have the ability to influence or restrict access to essential resources, such as data or AI services, which can give them significant power over the digital ecosystem and competition.

Embeddings: In artificial intelligence, embeddings are a technique used to represent complex data, such as words or images, in a simplified numerical format. These representations help AI systems understand

and process the relationships between different pieces of data. For example, in language models, embeddings help capture the meanings of words and how they relate to one another, improving the system's ability to understand context and meaning. Embeddings are intermediary steps between input and output of AI models that are not directly visible to the user (in a chatbot conversation for example) but can be used for AI search or data retrieval.

Finetuning: The process of adapting a pre-trained artificial intelligence model by retraining it on a specific dataset or for a particular task. A common method is LoRA finetuning (*Low-Rank Adaptation finetuning*), which involves adjusting a small number of specific model parameters while keeping the other parameters fixed. This approach reduces computational complexity and allows for efficient customization of the model for specific applications while using fewer resources.

Foundation models: Large artificial intelligence models, often pre-trained on massive volumes of data, capable of performing a wide range of tasks. These models serve as a base for developing specific applications through processes like finetuning. For example, *Claude 3* by Anthropic is a foundation model with derived versions such as Claude 3 Sonnet for a balance between performance and speed, Claude 3 Haiku for lightweight tasks, and Claude 3 Opus for complex and demanding needs.

Hyper-scalers: Large technology companies that operate extensive global networks of data centers, providing vast cloud computing and data management services. These companies, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP), support large-scale data processing and storage needs, essential for technologies like artificial intelligence (AI), machine learning, and big data analytics.

Inference: The process by which an artificial intelligence model applies its learned knowledge to make predictions, perform tasks, or make decisions based on new data.

Quantization: A technique that reduces the precision of the numbers used to represent an AI model's parameters (usually by going from 32 bits to 16 or 8 bits). This allows for reducing the model's size and speeding up computations, often with minimal loss of precision.

Multi-modal: Describes models that cover more than one form of input or output, for example, text, images, or audio.

Network effects: In the context of artificial intelligence (AI), network effects occur when an increase in the number of users improves the quality and efficiency of AI models. More users generate more data, allowing the models to train on larger and more diverse datasets, thereby increasing their accuracy.

Local deployment: Local deployment refers to the installation and execution of AI models directly on a company's computing infrastructure rather than through cloud services. This allows for greater data control, reduced recurring cloud costs, and better customization of models for specific needs, while requiring significant internal computing resources to manage the demands of processing power and storage.

Multi-homing: A practice where a user or company simultaneously uses multiple competing platforms or services. In the context of AI, this might involve using multiple AI models or cloud services to arbitrage across price, capabilities and user experience.

Open-source software: Software whose source code is made public, allowing anyone to study, modify, and redistribute it. In the context of AI, open source promotes collaboration, innovation, and transparency by enabling developers to build on existing foundations without licensing costs.

Open-weight models: Artificial intelligence models whose weights (internal parameters learned during training) are accessible to the public.

Parameters: These are the underlying components of trained models that determine how inputs are transformed into outputs. The number of parameter is usually a proxy for the computation needs of a model

(size of memory require to use it), and the size of the training dataset. In the paradigm of the scaling law bigger models are usually associated to better performances and so-called emergent properties.

Tokens: They are the standard unit of measurement for text in AI models. They are numerical representations of words and characters. Different LLMs use different tokenizers but one token is approximately four characters in English. The price of usage of AI models is usually expressed in million of tokens to provide a comparable statistic across models.

Training: During training, the model is exposed to a large volume of data, the size of which will influence the training compute, gradually adjusting its parameters to improve its performance on a specific task. For LLMs (Large Language Models), this involves analyzing vast text datasets to effectively understand and generate natural language.

Transformer: A machine learning (neural network) architecture introduced in 2017 by researchers at Google, which serves as the foundation for many natural language processing (NLP) models and other artificial intelligence applications. Transformers use attention mechanisms to process sequential data more efficiently than previous models like RNNs.

Scaling law: A principle that describes the relationship between the size of an artificial intelligence model (number of parameters), the amount of data used for training, and the model's performance. Generally, the scaling law suggests that the larger a model is and the more data it is trained on, the better its performance.

Annex B. Definition of AI models and key concepts

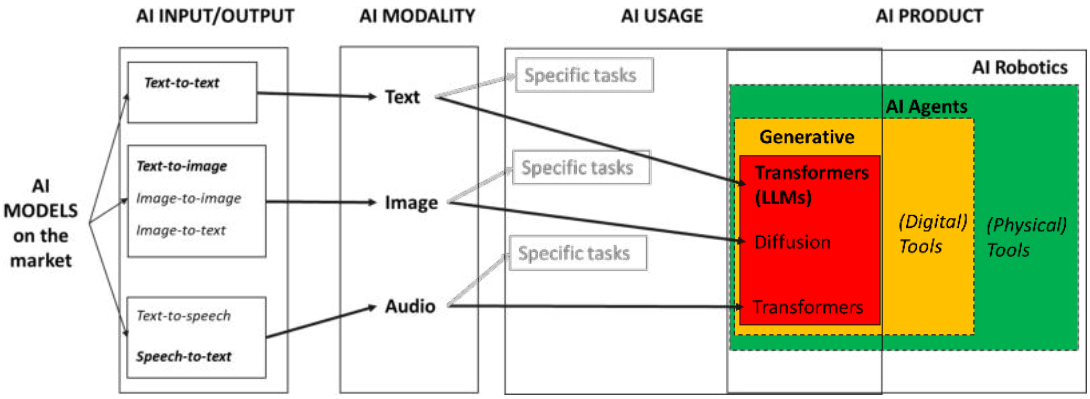
Table B.1. Definition, concepts and measurement in AI models

AI characteristics	Description
Modality	Modality of the input (prompt) and output (answer) of the AI. Modalities can be text-to-text, text and image-to-text (multimodal), text-to-image or image-to-image (image), text-to-sound or sound-to-text (audio), text-to-video (video) (Bommasani et al., 2021; Stanford University, 2024).
Performance	Capability of the model on industry benchmarks. Benchmarks are standardized tests covering different capabilities (reasoning, coding, summarising, etc.) and aimed at evaluating and comparing the quality of the models (Artificial Analysis, 2024; Chiang et al., 2024; Hendrycks et al., 2020; Srivastava et al., 2022).
Speed of inference	Velocity of the interaction between the user and the AI. Usually measured in tokens per second, it corresponds to an indicator of how fast it takes for the model to read the input (prompt) and provide an (answer) (Artificial Analysis, 2024).
Price of inference	Cost for the usage (interaction) with the AI through the API (Application Programming Interface) of a cloud provider. The prices per input (prompt) and output (answer) are usually different. We follow a convention summarizing the total price as a blended price that combines 3 inputs for 1 output (Artificial Analysis, 2024). Prices are expressed in USD dollars per million tokens in the case of text models For image, the price depends on the number of images generated and for audio on the number of seconds of audio input (Quentin, Biderman and Schoelkopf, 2023; Artificial Analysis, 2024). [49]
Openness	Availability of the weights of the model, but also the richness of information available concerning the training characteristics (architecture of the model, size and type of dataset for training; characteristics of the compute etc.) (Bommasani et al., 2023; White et al., 2024; Liesenfeld and Dingemanse, 2024]; Solaiman, 2023; Open-Weights-definition, 2024).

Figure B.1 illustrates how specialized AI models are integrated to form AI agents capable of performing complex tasks. It begins with various AI input/output modalities, such as text-to-text, image-to-text, and speech-to-text, which are processed by models specialized in handling text, image, or audio data. These models are used for specific tasks, like language translation or image captioning. The core of the figure highlights AI agents, which combine these specialized models to execute multi-step tasks that require handling multiple data types simultaneously. For example, an AI agent might use speech recognition, language understanding, and text generation together to function as a virtual assistant. The figure also differentiates between digital AI agents, which operate in software environments, and AI robotics, which apply these capabilities in physical systems. This figure is an illustration of the integration of diverse AI models to create versatile AI agents that can tackle economic relevant problems.

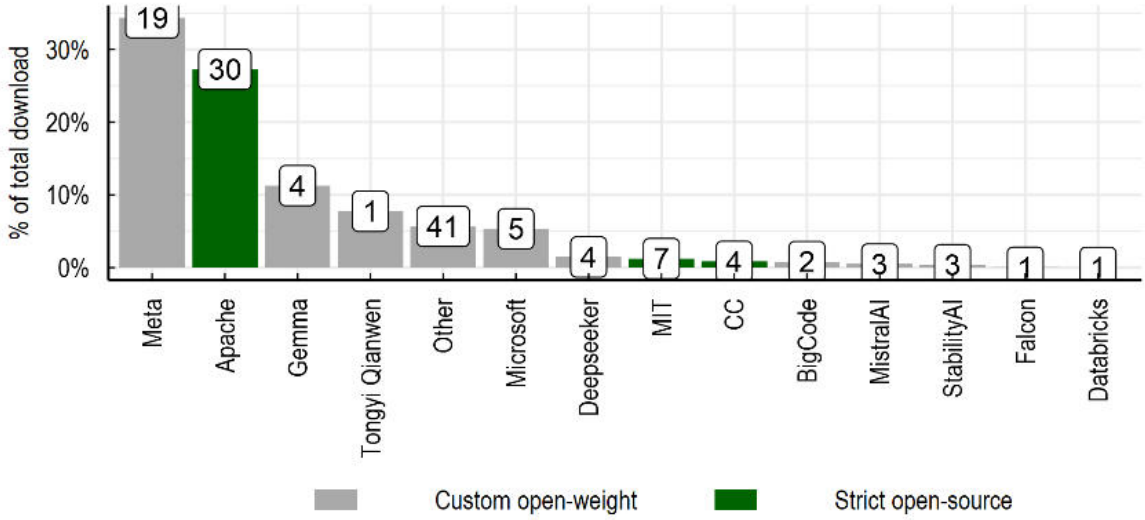
[49] In most cases, prices are the same across regions. However, several providers charge different prices for the same models in different locations, reflecting differences in inference costs. It is also important to note that prices are pre-VAT prices. Some companies provide different discount policies; for example, OpenAI proposes a 50% discount for queries with a 24-hour delay, and Anthropic and Google propose different prices when storing intermediary output (catching).

Figure B.1. From specialized AI models to AI agents



Source: authors' elaboration.

Figure B.2. Popularity of open-source AI models per license family



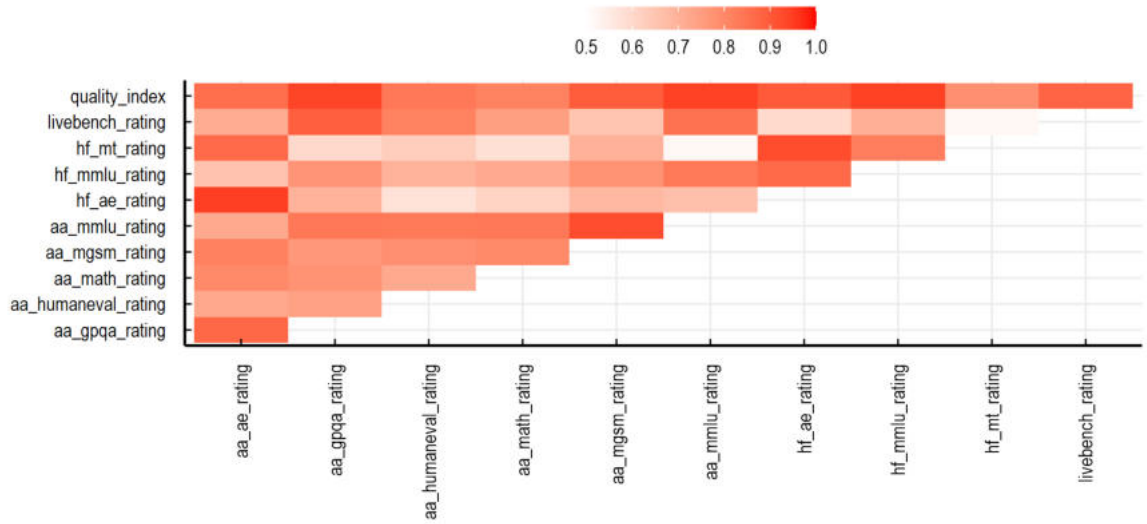
Note: Numbers on top of the bar refer to the number of models. A custom license refers to a license of open-weight models under the terms and conditions of the developing company. Open-source license refers to a license that complies with the licensing terms of the Open Source Initiative (Open Source Initiative, 2024[9]) and identified as being Apache 2.0, MIT or Creative Commons (CC) licences. Data collected in October 2024.
Source: Author's calculation based on data from Ollama.

Table B.2. Contribution of Open source to the AI value chain

Research and Development	- **Accessible Frameworks:** Open-source frameworks like TensorFlow, PyTorch, and Keras lower the barrier to entry for AI development. - **Reproducibility:** Open-source code and datasets enable researchers to reproduce experiments, validate findings, and build upon work. - **Cumulative Innovation**: Fosters cumulative innovation in software, hardware, and usage.
Data Collection and Preparation	- **Open Data Initiatives:** Projects like Hugging Face's Datasets library and The Pile provide open access to diverse and large-scale datasets. - **Data Annotation Tools:** Open-source tools for data annotation, such as LabelStudio and VGG Image Annotator (VIA), facilitate labeled data creation.
Model Training	- **Distributed Training:** Open-source libraries like Horovod enable distributed training of AI models across multiple GPUs and machines. - **Pre-trained Models:** Open-weight pre-trained models on platforms like Hugging Face allow for transfer learning and fine-tuning.
Deployment and Integration	- **Containerization:** Open-source containerization tools like Docker and orchestration platforms like Kubernetes facilitate model deployment. - **Model Serving:** Open-source serving frameworks, such as TensorFlow Serving and TorchServe, enable efficient deployment and scaling.
Evaluation and Monitoring	- **Benchmarking:** Open-source benchmarks and leaderboards provide standardized metrics for evaluating model performance. - **Monitoring Tools:** Open-source monitoring tools like Prometheus and Grafana help track the performance and health of AI models. Community and Ecosystem -**Knowledge Sharing:** Open-source projects encourage knowledge sharing through documentation, tutorials, and forums.
Effects on the main AI bottlenecks	**Compute** **-Compute Efficiency**: Eases compute bottlenecks, allowing demand for smaller models that run on traditional computers. **-Price Pressure:** Puts price pressure on cloud-based model provisions and limits cloud gatekeeping. **-Cost Transparency:** Enforces transparency in costs of inference. **-Limiting Rent Capture:** Reduces rent capture of cloud providers with exclusive access to closed AI licenses by allowing new entrants to compete with equivalent models. **-Cost-Free Solutions:** Provides new entrants with cost-free solutions and avoids dependence on incumbent competitors. **-Limiting Gatekeeping**: Limits gatekeeping from vertically integrated companies. **Skills** **-Skill Development:** Eases the skill bottleneck by raising the skills of AI developers, data scientists, and IT engineers. **-Accessibility:** Fosters the accessibility and democratization of access to AI. **-Product Quality:** Increases the quality of products by allowing companies to fine-tune and tailor foundation models for specific niche markets. **Model Customization:** Increases flexibility and customization of models, ensuring distributed power over information creation. **Data** **- Data access:** allow smaller players to access data for training, finetuning and improving models **-Independence and Control:** Provides gains in independence, privacy, and control over companies data

Source: author's elaboration.

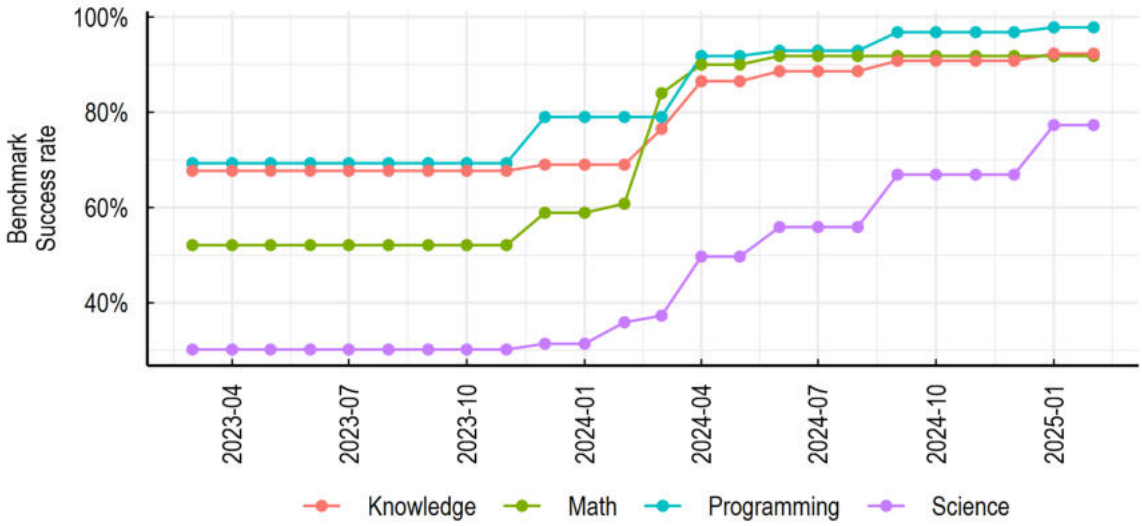
Figure B.3. Correlation across AI quality indicators



Note: The figure displays the pairwise correlation of performances on common benchmarks.
Source: author's calculations.

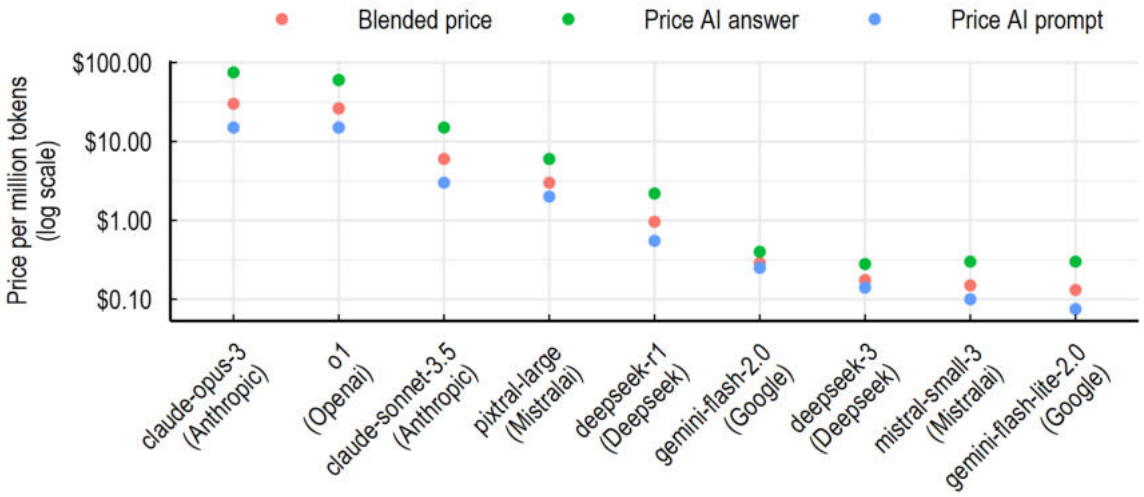
Figure B.4. Evolution of AI model Performance on industry benchmarks

Best performance achieved by AI models available on the market



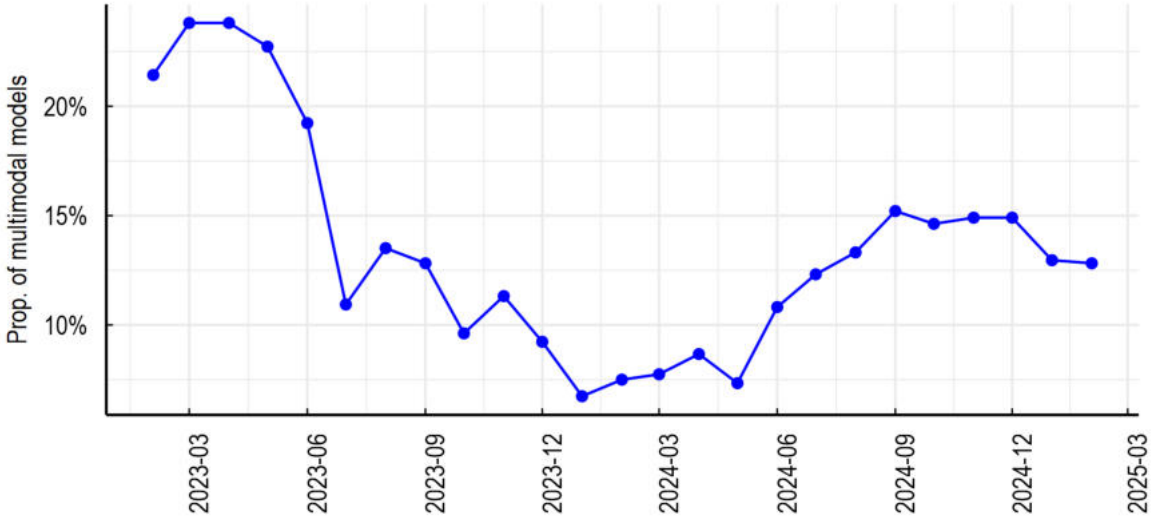
Note: The success rate of AI corresponds to the highest success rate of active models on industry benchmarks. Industry benchmarks are standardized test aimed at measuring the performance of AI models along several dimensions (see Annex A for more details). Knowledge is measured based on the MMLU benchmark, Math based on the MGSM benchmark, Programming on the HumanEval benchmark and Science on the GPQA benchmark.
Source: author's calculations based on data from Artificial Analysist.

Figure B.5. Price of selected AI models



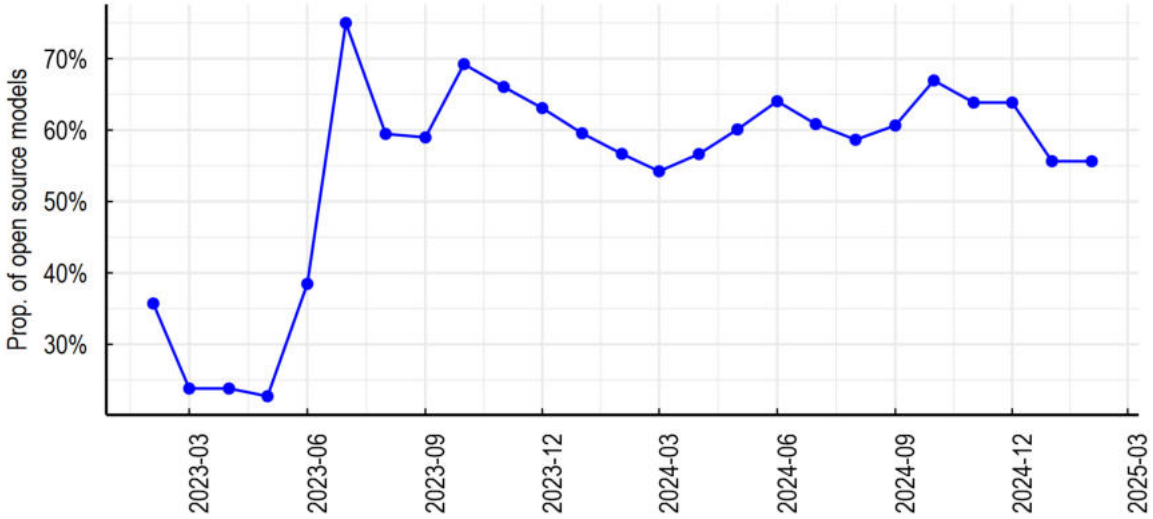
Note: The "Price of AI answer" corresponds to the output price while the "Price of AI prompt" corresponds to the input price of accessing the model via the API of the respective companies in February 2024. Blended price is the representative price of the use of AI per million tokens following the methodology in Annex C and calibrating the representative usage to 1 input token for 3 output token.
Source: author's calculations.

Figure B.6. The share of multimodal models among all Text-to-Text models



Source: Author's calculations.

Figure B.7. The share of open-weight models among all models



Source: author's calculations.

Annex C. Indicators to measure competition

Environment:

AI-adopting firms use AI as an input of production, and at each period, they arbitrage and choose the model that maximizes the quality at each given price and for their given usage.

The model requires three foundational assumptions on the market for AI (Table B.3):

- **Perfect model substitutability** = all models have the same architecture and technical integration for the AI-adopting firm, models only differ by their price and quality.
- **Heterogeneous taste for quality** = AI-adopting firms have heterogeneous preferences for quality
- **Perfect information on model quality and price** = information on the quality (performance) and the price of models are common knowledge for AI-adopting firms.

Table C.1. Assumptions on the determinants of market structure in AI

Type of market friction	Description	Assessment on their size
Model substitutability	Whether models are comparable, perform the same task and require the same complementarity factor of production	High
Heterogeneous taste for quality	AI-adopting firms perform different tasks and face different budget constraints which translates in different sensitivity to the cost of marginal improvements in the quality of the AI models they use.	High
Perfect information	AI-adopting firms can perfectly monitor the difference in quality and prices across the different suppliers and models	High
Switching costs	Capacity for AI-adopting firms to switch to the best model, company or provider at low cost and lag.	Low
Focality	Coordination problem for entrants here incremental increases in quality do not generate adoption	High
Cultural preference	Despite equivalent quality AI adopting firms may choose one developer rather than another for no direct economic reasons like national preference, security consumer	Low

Definition of AI models

Each AI model is uniquely characterised by four dimensions (Table B.1): the modality of the model defines the type of interaction with the user (text, image or sound), the performance (quality) of the model is a measure of the capability to successful perform a cognitive task, the speed of inference is the number of output generated by the model per unit of time assumed to be constant for all models, the price of inference is the cost of usage of the model and reflect the price of input and the price of output.

Definition of the AI Economic Frontier

At a given point in time, the AI Economic Frontier identifies the set of models available on the market that minimize price for a given level of quality. The Economic Frontier is upward-sloping since better performance is available at a higher price but at the expense of slower performance.

Formally, the optimal quality on the market is defined as:

$$g(p) = \max\{q_i | \, p_i \leq p\} \, \forall \, i \, \epsilon \, \{1, m\} \tag{4}$$

With,

p_i= the price of AI inference per million tokens for *Text-to-Text*, per 100 images of output for *Text-to-Image,* and per hour of audio transcription such that:

$$p_i = \alpha * p_i^{Input} + \sigma_i * (1 - \alpha) * p_i^{Output} \tag{5}$$

with

p_i^{Input}= the price of input (prompt from the user)

p_i^{Input}= the price of output (response from the AI)

$\alpha = \frac{size \; of \; the \; user \; prompt}{size \; AI \; answer}$ = Input to output ratio.

$\sigma_i = \frac{reasoning \; tokens}{output \; tokens}$ the reasoning coefficient of model i.

m = all the models available on the market

Quality (q) is model as an index representative of a cognitive tasks that includes 4 different sub tasks of various difficulty ($q^{HS} = high \; school \; tasks$, $q^{grad} = graduate \; level \; tasks$, $q^{pref} = qualitative \; preference$ and $q^{novel} = novel \; tasks$)

$$q = \mathrm{w}_{\mathrm{ungrad}} * q^{ungrad} + \mathrm{w}_{\mathrm{pref}} * q^{pref} + \mathrm{w}_{\mathrm{grad}} * q^{grad} + \mathrm{w}_{\mathrm{novel}} * q^{novel} \tag{6}$$

The AI economic Frontier is then defined as the subset of M_t models in F_t such that

$$F_t = \{i \mid q_i = Q(p_i) \, \forall \, i \; \in \{1 \, , M_t \}\} \tag{7}$$

Definition of AI market segments

Tier 1 models are the most capable (largest) and most expensive models at the frontier. These are multimodal variants and display the most advanced reasoning capabilities, typically necessary to perform general, multi-step tasks.

$$Q_1(p) = \max\big(Q(p_i)\big) \, \forall \, i \in F \tag{8}$$

Tier 2 models are highly capable (medium) models, near the median price at the frontier. These models are likely to be the most cost-effective general-purpose models for most tasks.

$$Q_2(p) = \text{median}\big(Q(p_i)\big)\ \forall\ i \in F \tag{9}$$

Tier 3 models are the cheapest and smallest models at the frontier, typically by an order of magnitude cheaper (and faster) than Tier 2 variants. Those models are optimised for specific, more routine tasks such as categorisation, classification or summarising text at a large scale.

$$Q_3(p) = \min\big(Q(p_i)\big)\ \forall\ i \in F \tag{10}$$

A simple model of AI revenues in a competition environment

The revenue of each model can be decomposed into 3 components: the price component, the adoption component (intensive and extensive margin) and the competition component (switching costs and focality).

R_{ijt}= Revenue of model i of company j at time t

P_{ijt} = Price of model i of company j at time t

A_t = Number of users of model i of company j at time t (extensive margin of adoption)

U_{ijt} = Intensity of utilisation of model i of company j at time t (intensive margin of adoption)

C_{ijt} = Competition pressure on model i of company j at time t (switching costs)

F_{jt} = user retention of model i of company j at time t (focality)

$$R_{ijt} = f(P_{ij}, A_t, U_{ijt}, C_{it}, F_{jt}) \tag{11}$$

Number of AI users

The number of users at time t is modelled applying a growth rate of g per cent per month of new AI users since the beginning of the period. We assume that the growth rate is constant across periods and companies.

$$A_t = (1+g)^t * N_0 \tag{12}$$

where,

g= the monthly growth rate in the number of AI users.

N_0= Number of AI users at the beginning of the period.

Competition components

Competition pressure and switching costs

The competition component includes two dimensions. First, the competition pressure (C_{ijt}) the component that captures whether a model is at the frontier at each period (the numerator in equation (13)) and among how many equivalent frontier models are the revenue shared (the denominator in equation (13)).The interpretation of this parameter is that the higher the value of C_{ijt} the lower the competition pressure.

Formally,

$$C_{ijt} = \frac{1[Q_{ij} \geq g(p_{kT} * (1-\gamma))]}{M_{kT}} \quad (13)$$

where

Q_{ij} = Quality of model i of company j

$g(p * (1-\gamma)) =$ Quality at the frontier of a model of price p given a frontier tolerance (switching) cost of γ percent.

$\gamma \in [0,1]$ measuring the size of the switching costs. For example, y=0.5 means that users continue using the model despite a price 50% more expensive than the reference model at the frontier.

M_{kT} = Number of models at the frontier equivalent to model k at time t.

Reputation

The second competition dimension is the reputation of a company (focality parameters) that allow companies to continue making revenues on models that have been displaced at the frontier by new models.

Reputation is modelled by averaging the revenues made on a model across the F precedent periods and weighting each period by an exponential decay θ.

Formally,

$$R_{jt} = \sum_{T=t-f}^{F} \theta^{T} * R_{jT} \quad (14)$$

with $\theta \in [0,1]$ the reputation parameter of model i of company j. For simplicity, we assume that the reputation parameter is the same across companies and periods. Intuitively, a company that make it to the frontier retain θ^{F} of the users is acquired in the past F months even the company is displaced from the frontier.

AI Market segments

Before aggregating at the company levels, we first define the AI market segments and assign each model to its relative market segment based on its price. We define the number of market segments (S) as the number of subsections of the AI economic frontier, The price and quality of each segment (tier 1, tiers 2 and tiers 3) is represented by the price and quality of the respective model at the frontier of each segment.

Utilisation of AI

Intensity of utilisation

The utilisation of AI can be decomposed in two dimensions. The intensity of utilisation (U) that reflects the number and size of interaction of each user with the AI. It corresponds to the intensive margin in the demand for AI. To simplify the model, we assume that it is constant across segments and normalized to one.

Usage of AI

AI foundation models can be used for a variety of tasks where specialized small models can sometimes perform as good generalist models for a fraction of the costs. We model different usage scenarios by adjusting the proportion of total demand addressed to each segment of the market (w_s). For example, $w_s = 1$ suppose that all the demand is addressed to the more capable and expansive model at the frontier.

Revenues from AI for each market segment

Combining all the previous equations we derived the estimated revenues of company j at the segment S at time t using the information on the AI economic frontier (best price and quality), the number of players at the frontier (M_{st}) and the price and quality of models at each model, making several assumptions on global demand (number of users (A), intensity of utilisation (U) and type of usage (w) and calibrating three competition parameters (θ, F and γ).

$$R_{jst} = \sum_{T=t-f}^{F} \sum_{i=1}^{M} w_s * A_t * U * p_{st} * \theta^T * \frac{1[Q_{sj} \geq g(p_{sT} * (1-\gamma))]}{M_{sT}} \quad (15)$$

Revenues from AI for each company

We derive simply the revenue of company j at time t as the sum of the revenues in each market segment

$$R_{jt} = \sum_{k=1}^{S} R_{kjt} \quad (16)$$

Equation (15) provide the fully fledge equation for the total AI revenues of each company j at time t.

$$R_{jt} = \sum_{k=1}^{S} \sum_{T=t-f}^{F} \sum_{i=1}^{M} w_k * (1+g)^t * N_0 * U * (\alpha * p_{kt}^{input} + \sigma_{kt} * (1-\alpha) * p_{kt}^{output}) * \theta^T * \frac{1[Q_{ij} \geq g(p_{kT} * (1-\gamma))]}{M_{kT}} \quad (17)$$

Revenue for each company

To provide a full picture of the revenues of AI, we aggregate the revenue across the different AI modalities by applying the following weighted sum

$$R_{jt} = \sum_{g=1}^{G} \delta_g * R_{gjt} \quad (18)$$

With δ_g the proportion of total AI demand that is addressed to the AI modality g.

Market shares for each company per modality

Finally, the market share of company j at time t is defined as the share of the revenue of the company divided by the sum of the revenue generated by all companies in the market.

$$MS_{jt} = \frac{R_{jt}}{\sum_{n=1}^{D} R_{nt}} \quad (19)$$

Calibrations of demand scenarios

- *The Baseline demand scenario* corresponds to our central scenario where demand is more balanced across different use cases, following reported usage (demand) from qualitative sources (e.g. OpenRouter)[50] , and roughly corresponds to suggested use by model developers: demand for AI is set to 5% for the Tier 1 segment (Larger), 70% for Tier 2 (Medium), and 25% for Tier 3 (Small).
- *"AGI (Artificial General Intelligence) demand scenario"* assumes demand focuses on the strongest performance; in particular, the best model (Tier 1) captures 70% of the demand for AI, tiers 2 25% and tiers 3 5%. This scenario captures a particularly strong willingness to pay for quality from AI-adopting firms and assumes no supply restrictions due to compute bottlenecks (for inference).
- *"Edge demand scenario"* assumes that the smallest, most cost-efficient models (Tier 3) – that are often deployed on local, lower-performance hardware such as mobile devices, on the "*edge*" of networks – attract 70% of the demand while Tier 1 only attracts 5% and Tier 2 attracts 25%. This scenario reflects the use, at large scale, of *"good enough"* specialized AI models in a scenario where taste for quantity rather than quality prevails. It also implicitly assumes stronger compute bottlenecks where cutting-edge hardware to run the most capable models are in short supply.

Calibrations of competition scenarios

- *"High switching" costs* refer to a scenario where consumers switch to the model at the frontier only if it costs significantly less (γ as expensive) than the reference model at the frontier, *and* reputation driven concerns are strong, that is, a proportion θ^F of consumers remain with the model they have choose F period in the past although this model does not belong to the frontier anymore.

[50] According to the data of OpenRouter, demand for the most capable o1-preview model represents less than 1% of the total demanded tokens of the platforms between the 10th of November 2024 and the 10th of December 2024, with 70% going for models in the tiers 2 segments like Claude sonnet 3.5, GPT4o, Gemini 1.5 and 30% for small models in the Tier 3 segment.

- *"Low switching" costs* reflect a scenario where each month, all consumers switch to the models at the frontier that corresponds to their usage ($\gamma = 0$), regardless of the company of origin (absence of *focality*).

Table C.2. Calibration of scenarios

Parameter	Calibration	
$\boldsymbol{w_s}$	AGI	$(w_1 = 0.7\,, w_s = 0.25\,, w_s = 0.5)$
	Baseline	$(w_1 = 0.05\,, w_s = 0.7\,, w_s = 0.25)$
	Edge	$(w_1 = 0.05\,, w_s = 0.25\,, w_s = 0.7)$
$\mathbf{g}$	0.03	
$\boldsymbol{N_0}$	100	
$\boldsymbol{U}$	1	
$\boldsymbol{\theta}$	1	
$\boldsymbol{F}$	High reputation	22
	Low reputation	0
$\boldsymbol{\gamma}$	High switching costs	1
	Low switching costs	0.05

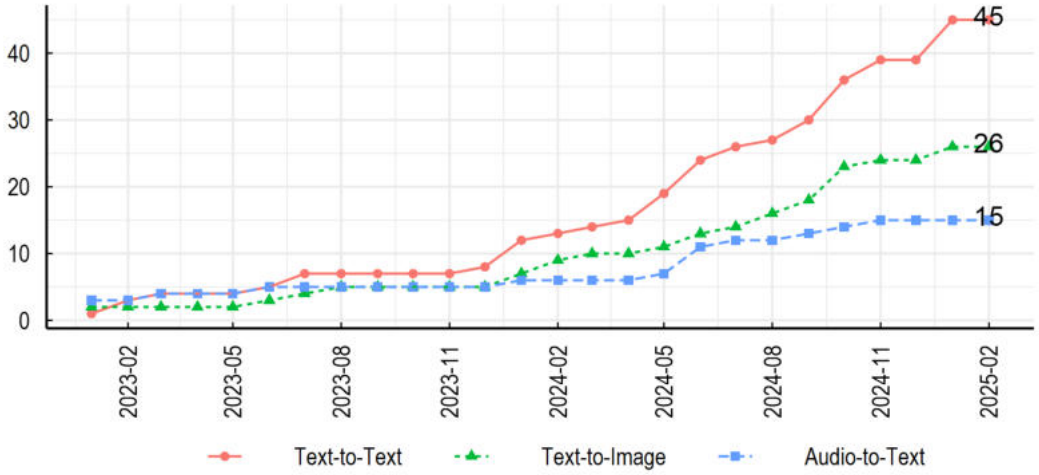
Table C.3. Key indicators for monitoring competition in AI markets

Dimension	Indicator	Description
AI model Market structure	**Number of companies at the frontier**	Number of companies that develop (AI developers) or provide (AI providers) the models at the AI Economic Frontier (among developers or providers). The indicator measures how many players compete with each other when offering a given quality of model at the lowest price.
	Price gap with the frontier	Difference between the price of the model and the price of the comparable model at the AI Economic Frontier. For each level of performance, the indicator measures the reduction of price necessary for the model to become an optimal choice for AI-adopting firms. $$G_t = \frac{p\_t}{p(Q^t(p_i, s_i))} - 1$$
	Distance to the frontier	Number of months (k) for an active model in period t to be at the AI Economic Frontier. $$DF_t = \{k \mid g^t(p_i, s_i) > q_i^t \geq g^{t-k}(p_i, s_i)\}$$
	Persistence at the frontier	Number of months at the AI Economic Frontier $$PF_{jt} = \sum_{i=1}^{M} \sum_{k=1}^{t} 1[Q_{ij} \geq g^k(p_{iT} * (1-\gamma))$$
	Price index of AI	Quality-adjusted price index of AI for each segment. The indicator provides the change over time in the price of frontier AI models for each three segments of the market. $$I_t = \omega_1 * \frac{(\frac{P_t^{T1}}{Q_t^{T1}})}{(\frac{P_0^{T1}}{Q_0^{T1}})} + \omega_2 * \frac{(\frac{P_t^{T2}}{Q_t^{T2}})}{(\frac{P_0^{T2}}{Q_0^{T2}})} + \omega_3 * \frac{(\frac{P_t^{T3}}{Q_t^{T3}})}{(\frac{P_0^{T3}}{Q_0^{T3}})}$$
	Churning rate of AI models	The proportion of models at the frontier at time t-1 but no longer at the frontier at time t. $$\mathrm{d_t} = \frac{N_t}{N_{t-1}}$$

Source: Authors' elaboration.

Annex D. Data sources

Figure D.1. Evolution of the number of cloud providers



Source: author's elaboration.

Table D.1. Data sources on AI prices

id	country	region	url_price
MistralAI	FRA	Europe	https://mistral.ai/technology/#pricing
OpenAI	USA	North Am.	https://openai.com/pricing
Google	USA	North Am.	https://cloud.google.com/vertex-ai/generative-ai/pricing?
Anthropic	USA	North Am.	https://www.anthropic.com/pricing#anthropic-api
Cohere	CAN	North Am.	https://cohere.com/pricing
AI21	ISR	Mena	https://www.ai21.com/studio/pricing
Amazon	USA	North Am.	https://aws.amazon.com/fr/bedrock/pricing/
StabilityAI	GBR	Europe	https://platform.stability.ai/pricing
IBMWatsonxAI	USA	North Am.	https://www.ibm.com/products/watsonx-ai/foundation-models#generative
Alep-Alpha	DEU	Europe	https://docs.aleph-alpha.com/docs/pricing/
Infomaniak	CHE	Europe	https://www.infomaniak.com/fr/hebergement/ai-tools/tarifs
MixedbreadAI	DEU	Europe	https://www.mixedbread.ai/pricing
Deepseek	CHN	Asia	https://www.deepseek.com/
Deepinfra	USA	North Am.	https://deepinfra.com/pricing
ZhipuAI	CHN	Asia	https://open.bigmodel.cn/pricing
BaichuanAI	CHN	Asia	https://platform.baichuan-ai.com/price
OpenRouter	USA	North Am.	https://openrouter.ai/docs/models
AIMLAPI	USA	North Am.	https://aimlapi.com/ai-ml-api-pricing
Minimaxi	CHN	Asia	https://www.minimaxi.com/en/price
Databricks	USA	North Am.	https://www.databricks.com/product/pricing/foundation-model-serving
ClarifayAI	USA	North Am.	https://www.clarifai.com/hubfs/Pricing_Page.pdf
TogetherAI	USA	North Am.	https://www.together.ai/pricing

ReplicateAI	USA	North Am.	https://replicate.com/pricing
RekaAI	USA	North Am.	https://www.reka.ai/reka-api
Deepgram	USA	North Am.	https://deepgram.com/pricing
PerplexityAI	USA	North Am.	https://docs.perplexity.ai/guides/pricing
Groq	USA	North Am.	https://groq.com/enterprise-access/
Lepton AI	USA	North Am.	https://www.lepton.ai/pricing
Octo AI	USA	North Am.	https://octo.ai/docs/getting-started/pricing-and-billing
NovitaAI	SGP	Asia	https://novita.ai/model-api/pricing
UpstageAI	KOR	Asia	https://www.upstage.ai/pricing?utm_term=Pricing&utm_content=%2Fgnb
Hyperbolic	USA	North Am.	https://docs.hyperbolic.xyz/docs/hyperbolic-ai-inference-pricing
Sensenova	CHN	Asia	https://platform.sensenova.cn/pricing
Simplismart	IND	Asia	https://www.simplismart.ai/pricing#Tab%201
Sambanova	USA	North Am.	https://cloud.sambanova.ai/pricingb
Fal	USA	North Am.	https://fal.ai/pricing
Lamini	USA	North Am.	https://www.lamini.ai/pricing
Segmind	IND	Asia	https://www.segmind.com/pricing
Fireworks	USA	North Am.	https://fireworks.ai/models
Midjourney	USA	North Am.	https://docs.midjourney.com/docs/model-versions
Ideogram	USA	North Am.	https://ideogram.ai/pricing
RevAI	USA	North Am.	https://www.rev.ai/pricing
Cloudfare	USA	North Am.	https://developers.cloudflare.com/workers-ai/platform/pricing/
Gladia	FRA	Europe	https://www.gladia.io/pricing
Speechmatics	GBR	Europe	https://www.speechmatics.com/pricing
Assemblyai	USA	North Am.	https://www.assemblyai.com/pricing
Cartesia	USA	North Am.	https://www.cartesia.ai/pricing
RecraftAI	GBR	Europe	https://www.recraft.ai/blog/pricing-update#appendix
FriendliAI	USA	North Am.	https://friendli.ai/pricing/dedicated-endpoints
Nebius	NLD	Europe	https://nebius.com/prices-ai-studio
blackforestlabs	DEU	Europe	https://docs.bfl.ml/pricing/
grok	USA	North Am.	https://docs.x.ai/docs/models?cluster=us-east-1
Bytedance	CHN	Asia	https://www.volcengine.com/pricing?product=ark_bd&tab=1
Stepfun	CHN	Asia	https://platform.stepfun.com/docs/pricing/details
Avian	USA	North Am.	https://avian.io/pricing/
Luma	USA	North Am.	https://lumalabs.ai/dream-machine/api/pricing

Table D.2. AI developers and providers by country

country	Date developers	AI developers	Date providers	AI cloud providers
ARE	2023-07-01-2025-02-01	Tii(Text-to-Text)		
CAN	2023-07-01-2025-02-01	Cohere(Text-to-Text), Ideogram(Text-to-Image)	2023-07-01-2025-02-01	Cohere(Text-to-Text), Ideogram(Text-to-Image)
CHN	2023-06-01-2025-02-01	Zhipuai(Text-to-Text/Text-to-Image), 01ai(Text-to-Text), Baai(Text-to-Text), Alibaba(Text-to-Text), Deepseek(Text-to-Text), Baichuanai(Text-to-Text), Minimaxi(Text-to-Text), Stepfun(Text-to-Text/Audio-to-Text), Sensenova(Text-to-Text), Baidu(Text-to-Text), Bytedance(Text-to-Text/Text-to-Image), Tencent(Text-to-Text), Tencentarc(Text-to-Image)	2023-06-01-2025-02-01	Zhipuai(Text-to-Text/Text-to-Image), Deepseek(Text-to-Text), Baichuanai(Text-to-Text), Minimaxi(Text-to-Text), Stepfun(Text-to-Text/Audio-to-Text), O1ai(Text-to-Text), Sensenova(Text-to-Text), Baidu(Text-to-Text), Bytedance(Text-to-Text), Tencent(Text-to-Text)
DEU	2023-02-01-2025-02-01	Alep-Alpha(Text-to-Text), Blackforestlabs(Text-to-Image), Lykon(Text-to-Image)	2023-02-01-2025-02-01	Alep-Alpha(Text-to-Text), Blackforestlabs(Text-to-Image)
FRA	2023-02-01-2025-02-01	Mistralai(Text-to-Text), Huggingface(Text-to-Text), Openassistant(Text-to-Text), Gladia(Audio-to-Text), Cartesia(Text-to-Audio)	2023-06-01-2025-02-01	Mistralai(Text-to-Text), Gladia(Audio-to-Text), Cartesia(Text-to-Audio)

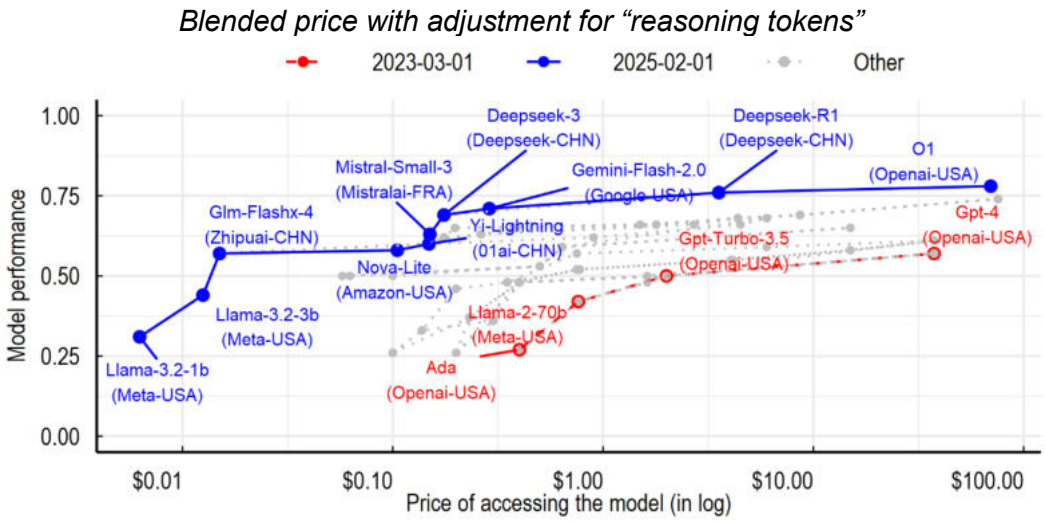
GBR	2023-01-01-2025-02-01	Stabilityai(Text-to-Text/Text-to-Image), Recraftai(Text-to-Image), Speechmatics(Audio-to-Text)	2023-01-01-2025-02-01	Stabilityai(Text-to-Image), Recraftai(Text-to-Image), Speechmatics(Audio-to-Text)
IND	2024-10-01-2024-12-01	Segmind(Text-to-Image)	2024-05-01-2025-02-01	Simplismart(Text-to-Text/Text-to-Image/Audio-to-Text), Segmind(Text-to-Image)
INT	2023-07-01-2025-02-01	Eleutherai(Text-to-Text), Bigcode(Text-to-Text), Bigscience(Text-to-Text)		
ISR	2023-03-01-2025-02-01	Ai21(Text-to-Text)	2023-03-01-2025-02-01	Ai21(Text-to-Text)
JPN	2024-05-01-2025-02-01	Elyza(Text-to-Text)		
KOR	2024-06-01-2025-02-01	Upstageai(Text-to-Text)	2024-09-01-2025-02-01	Upstageai(Text-to-Text)
POL	2024-07-01-2025-02-01	Elevenlabs(Text-to-Audio)	2024-07-01-2025-02-01	Elevenlabs(Text-to-Audio)
SAU	2024-07-01-2025-02-01	Sdaia(Text-to-Text)		
USA	2023-01-01-2025-02-01	Openai(Text-to-Text/Text-to-Image/Audio-to-Text/Text-to-Audio), Meta(Text-to-Text), Stanford University(Text-to-Text), University Of Washington Nlp(Text-to-Text), Bair(Text-to-Text), Togetherai(Text-to-Text/Text-to-Image), Lmsys(Text-to-Text), Databricks(Text-to-Text), Google(Text-to-Text/Text-to-Image/Text-to-Audio), Salesforce(Text-to-Text), Mosaic(Text-to-Text), Nous(Text-to-Text), Replit(Text-to-Text), Gryphe(Text-to-Text), Openchat(Text-to-Text), Anthropic(Text-to-Text), Lepton Ai(Text-to-Text), Perplexityai(Text-to-Text), Microsoft(Text-to-Text/Text-to-Audio/Audio-to-Text), Replicateai(Text-to-Text/Text-to-Image/Text-to-Audio), Cognitivecomputations(Text-to-Text), Amazon(Text-to-Text/Text-to-Image/Text-to-Audio), Ibmwatsonxai(Text-to-Text), Core42(Text-to-Text), Mindsandcompany(Text-to-Text), Nexusai(Text-to-Text), Olmoai(Text-to-Text), Teknium(Text-to-Text), Remmai(Text-to-Text), Defog(Text-to-Text), Snorkel(Text-to-Text), Fireworks(Text-to-Text), Neversleep(Text-to-Text), Lynn(Text-to-Text), Snowflake(Text-to-Text), Rekaai(Text-to-Text), Nvidia(Text-to-Text), Disco Research(Text-to-Text), Llava-Hf(Text-to-Text), Gradientai(Text-to-Text), Allen Institute For Ai(Text-to-Text), Nexusflow(Text-to-Text), Tinyllama(Text-to-Text), Alpindale(Text-to-Text), Inflection(Text-to-Text), Liquid(Text-to-Text), Pygmalionai(Text-to-Text), Grok(Text-to-Text), Austism(Text-to-Text), Test(Text-to-Text), Carson(Text-to-Text), Midjourney(Text-to-Image), Wavymulder(Text-to-Image), Prompthero(Text-to-Image), Playgroundai(Text-to-Image), Luma(Text-to-Image), Assemblyai(Audio-to-Text), Deepgram(Audio-to-Text), Lmnt(Text-to-Audio)	2023-01-01-2025-02-01	Openai(Text-to-Text/Text-to-Image/Audio-to-Text/Text-to-Audio), Deepinfra(Text-to-Text/Audio-to-Text), Togetherai(Text-to-Text/Text-to-Image), Amazon(Text-to-Text/Text-to-Image/Text-to-Audio), Lepton Ai(Text-to-Text/Text-to-Image/Audio-to-Text), Perplexityai(Text-to-Text), Replicateai(Text-to-Text/Text-to-Image/Text-to-Audio), Clarifayai(Text-to-Text/Text-to-Image), Octo Ai(Text-to-Text/Text-to-Image), Google(Text-to-Text/Text-to-Image/Text-to-Audio), Databricks(Text-to-Text), Groq(Text-to-Text/Audio-to-Text), Ibmwatsonxai(Text-to-Text), Aimlapi(Text-to-Text/Text-to-Image/Audio-to-Text), Anthropic(Text-to-Text), Openrouter(Text-to-Text), Rekaai(Text-to-Text), Microsoft(Text-to-Text/Text-to-Image/Text-to-Audio/Audio-to-Text), Cloudfare(Text-to-Text/Text-to-Image), Fireworks(Text-to-Text/Text-to-Image/Audio-to-Text), Hyperbolic(Text-to-Text/Text-to-Image), Lamini(Text-to-Text), Sambanova(Text-to-Text), Assemblyai(Text-to-Text/Audio-to-Text), Grok(Text-to-Text), Midjourney(Text-to-Image), Fal(Text-to-Image/Audio-to-Text), Luma(Text-to-Image), Deepgram(Audio-to-Text), Lmnt(Text-to-Audio)
CHE			2024-06-01-2025-02-01	Infomaniak(Text-to-Text/Text-to-Image/Audio-to-Text)
NLD			2024-11-01-2025-02-01	Nebius(Text-to-Text/Text-to-Image)
SGP			2024-09-01-2025-02-01	Novitaai(Text-to-Text/Text-to-Image)

Note: Country classification is based on the location of the main headquarters. INT stands for International and reflects labs with no specific headquarter location or regrouping researchers from different countries.
Source: Author's elaboration.

Annex E. Sensitivity analysis of key results

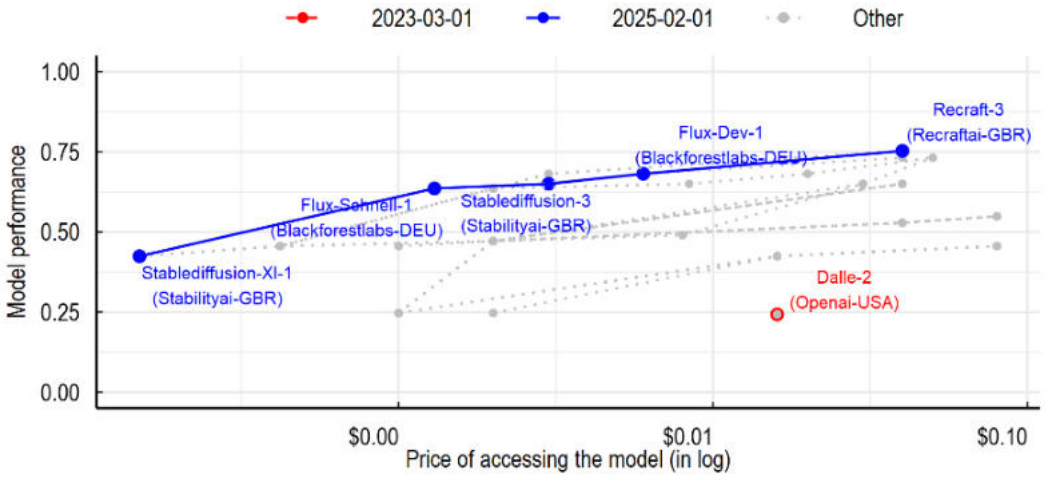
Details of the AI Economic Frontier

Figure E.1. Alternative specification of the AI Economic Frontier, Text-to-Text



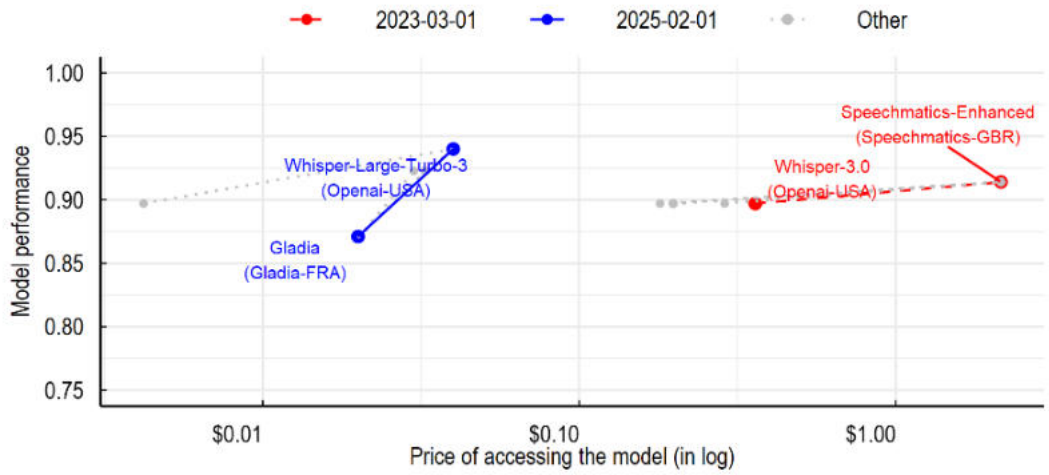
Source: Author's calculations.

Figure E.2. AI Economic Frontier, Text-to-Image



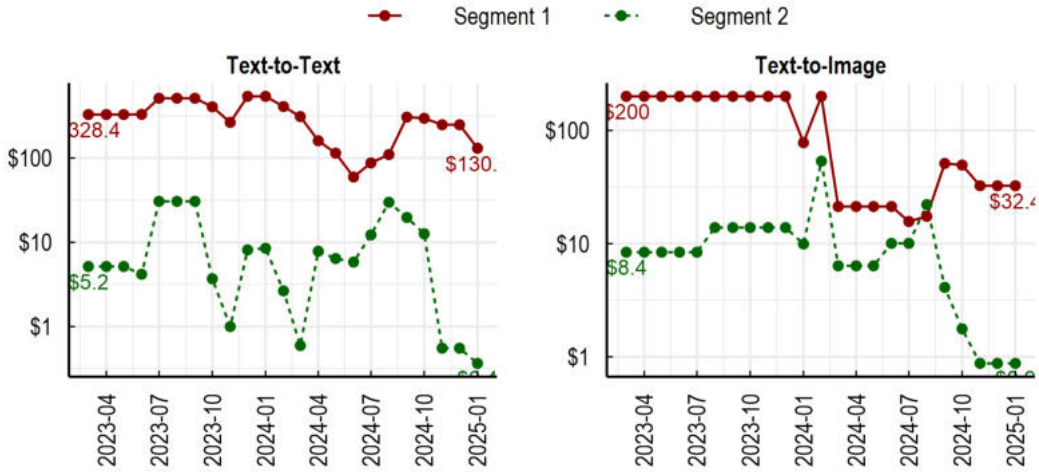
Note: Each dot represents the model with the best price-performance trade-off over the full sample of Text-to-Image active models available each month. The solid (dashed) line represents the AI economic frontier in January 2025 (March 2023). Performance is defined by a normalised weighted index of performance based on common benchmarks of the industry. For more details about the methodology, including performance and price measurement, see Annex A and B.
Source: author's calculations.

Figure E.3. AI Economic Frontier, Audio-to-Text



Note: Each dot represents the model with the best price-performance trade-off over the full sample of Text-to-Image active models available each month. The solid (dashed) line represents the AI economic frontier in January 2025 (March 2023). Performance is defined by a normalised weighted index of performance based on common benchmarks of the industry. For more details about the methodology, including performance and price measurement, see Annex A and B.
Source: author's calculations.

Figure E.4. Evolution of the slope of the AI Economic frontier

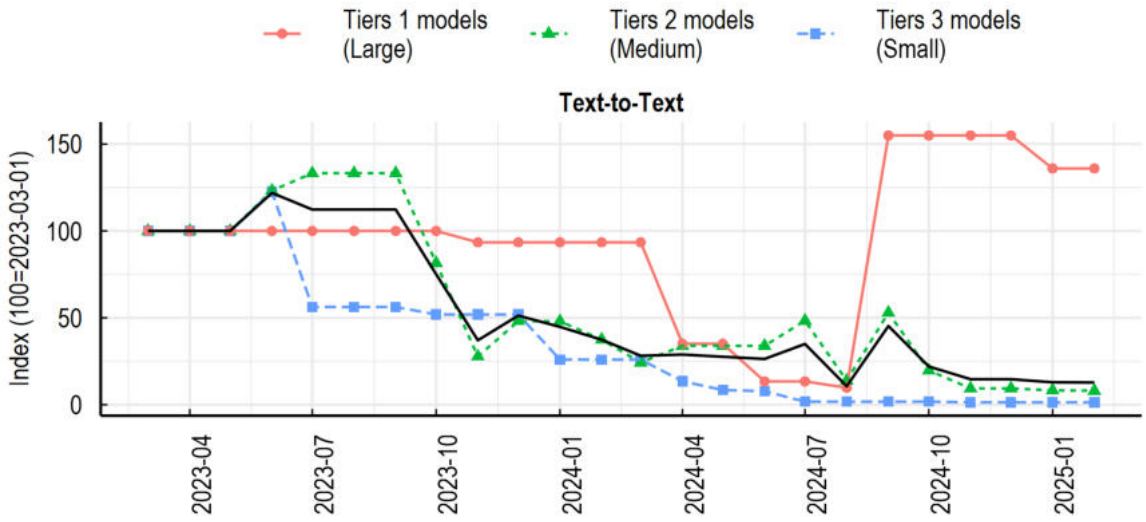


Note: The figures show the estimated slope of the linear estimation of the AI Economic Frontier, between the representative model in tiers 3 and tiers 2 market segment (segment 2) and between the model in Tier 2 and Tier 1 (segment 1). The y-axis denotes the additional cost of accessing a 10 p.p higher performance on benchmarks.
Source: author's calculations.

Details on the AI price index

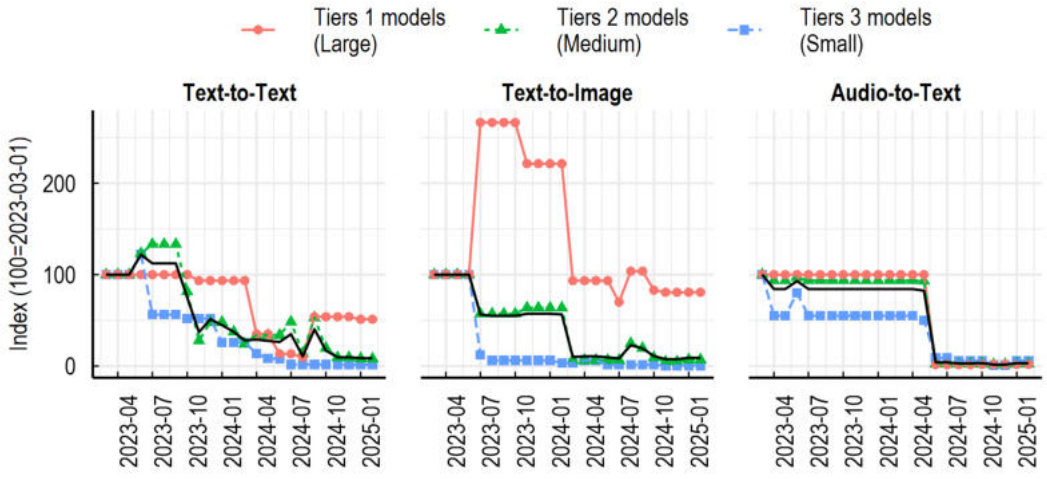
Figure E.5. Quality-adjusted price index Text-to-Text, alternative price variable

Blended price with adjustment for "reasoning tokens"



Note: The figure shows the Quality adjusted price index by market segment as defined in Annex C and includes in the price of AI output the "hidden" tokens charged in reasoning models. The black line denotes the aggregate index under the baseline demand scenario where 70% of demand is addressed to tiers 2 models, 25% to tier 3 and 5% to tier 1 models.
Source: Author's calculations.

Figure E.6. Quality adjusted price index by AI model segment and modality



Note: Black line denotes the aggregate index for each modality under the baseline demand scenario as defined in Annex C.
Source: Author's calculations.

Alternative method to estimate quality adjusted prices: Hedonic regression method

An alternative approach to compute the price index is to adopt a hedonic method by computing and index based on the estimated time fixed effects of equation (20) and showed in Figure E.7. The results shows that the declining trends is similar with a cumulative decline around 80% during the period

$$P_{ticp} = \alpha_1 + \beta_1 Q_i + \beta_2 Q_i^2 + \beta_3 S_i + \beta_4 O_i + \beta_4 H_i + T_t + \varepsilon_{itcp} \quad (20)$$

Q_i= A quality index measuring the performances of models on industry benchmarks

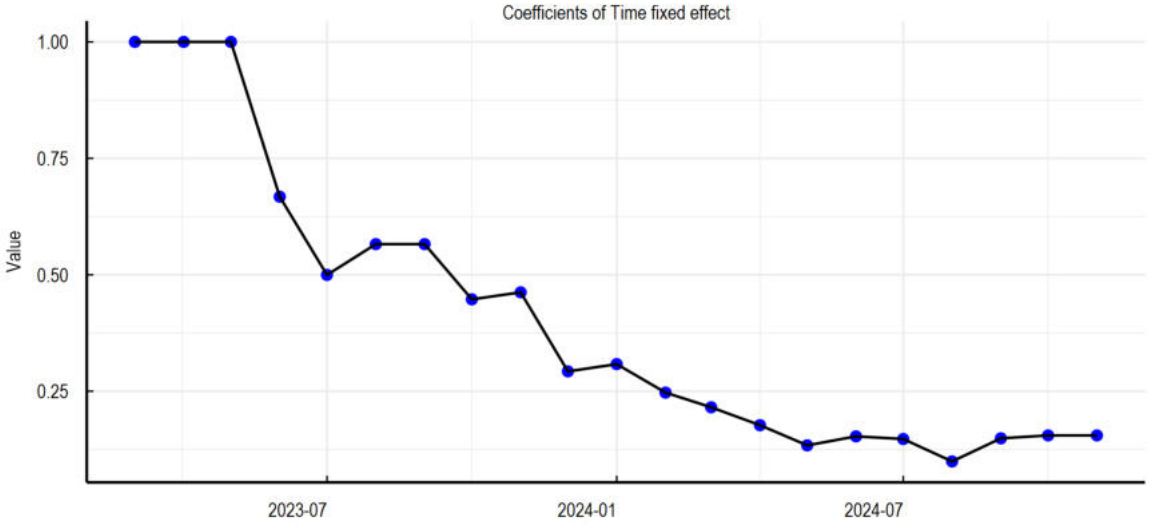
S_i= A speed index measuring the number of token per second

O_i= a dummy variable denoting whether the model is open-source

H_i= a dummy variable for models served by hyperscalers (AWS, Amazon or Google)

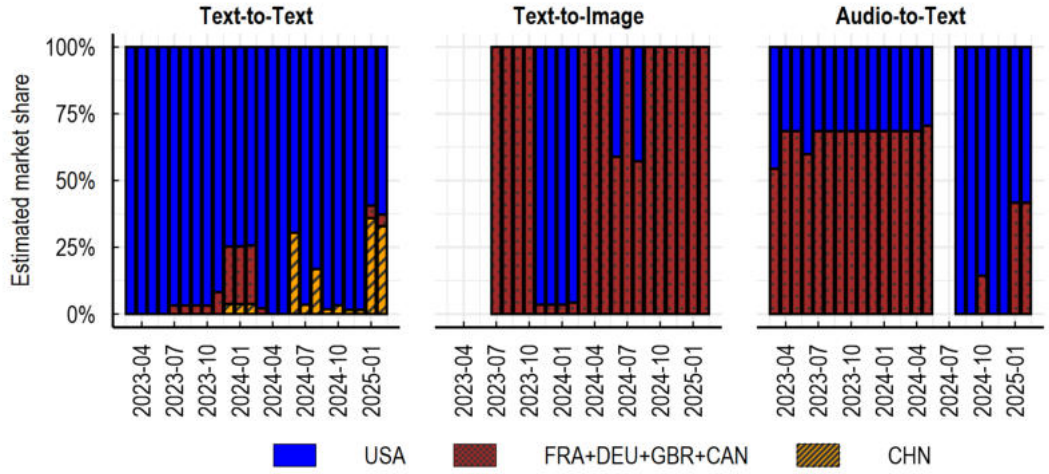
T_t= a year-month fixed effect

Figure E.7. AI price index - hedonic methodology



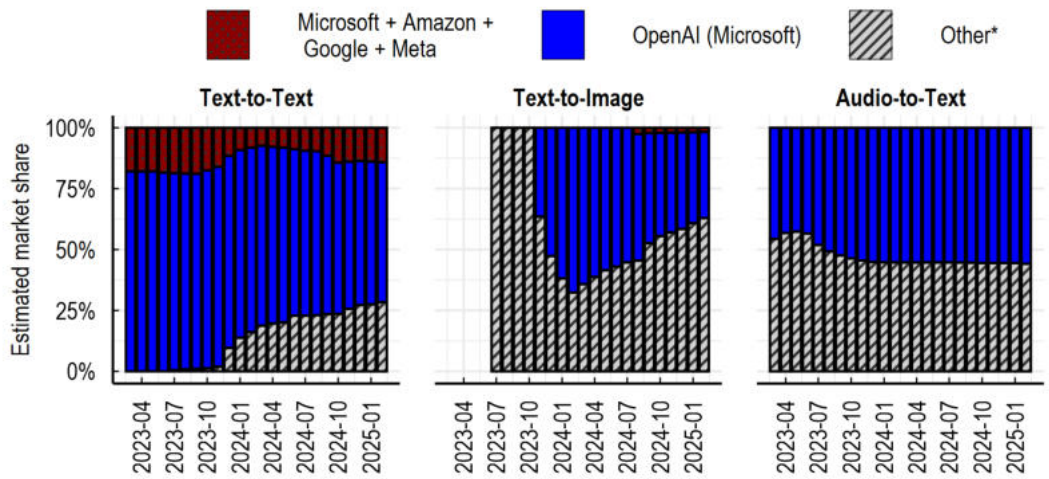
Market share under alternative scenarios and calibrations

Figure E.8. Simulated market shares by region, baseline scenario under low switching costs



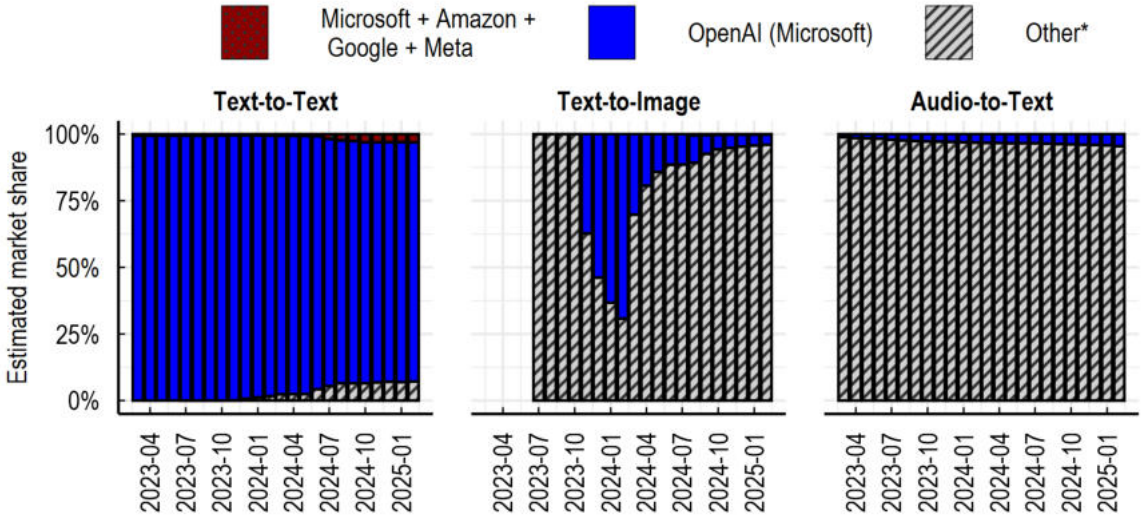
Note: Simulated market share of AI model revenues per region of the AI developing company for different calibrations reflecting several assumptions on the determinants of market structure as detailed in Annex C.
Source: authors' calculations.

Figure E.9. Simulated market shares by company status, baseline demand scenario under high switching costs



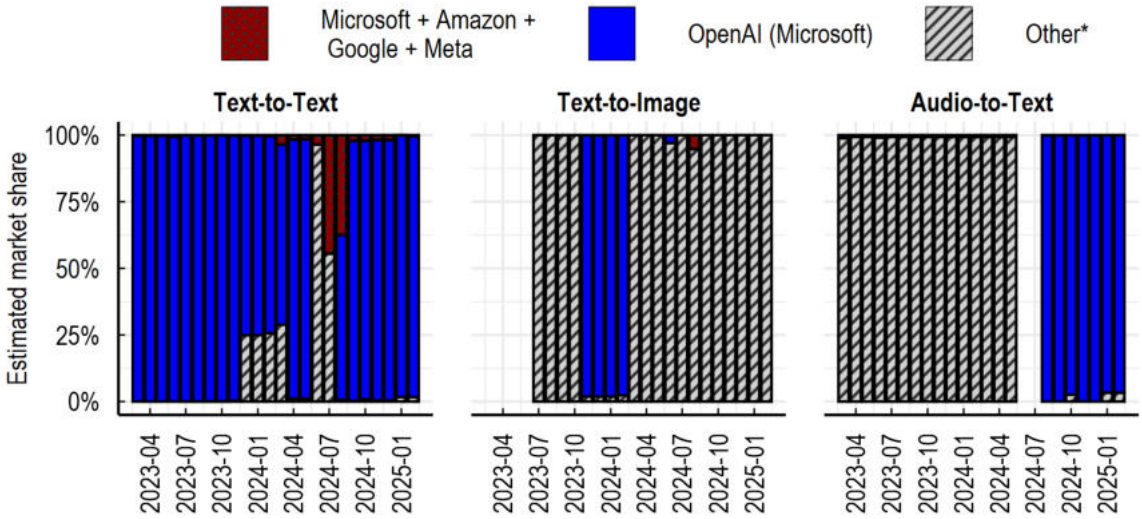
Note: Simulated market share of AI model revenues per region of the AI developing company for different calibrations reflecting several assumptions on the determinants of market structure as detailed in Annex C.
Source: authors' calculations.

Figure E.10. Simulated market shares by company status, AGI demand scenario under high switching costs



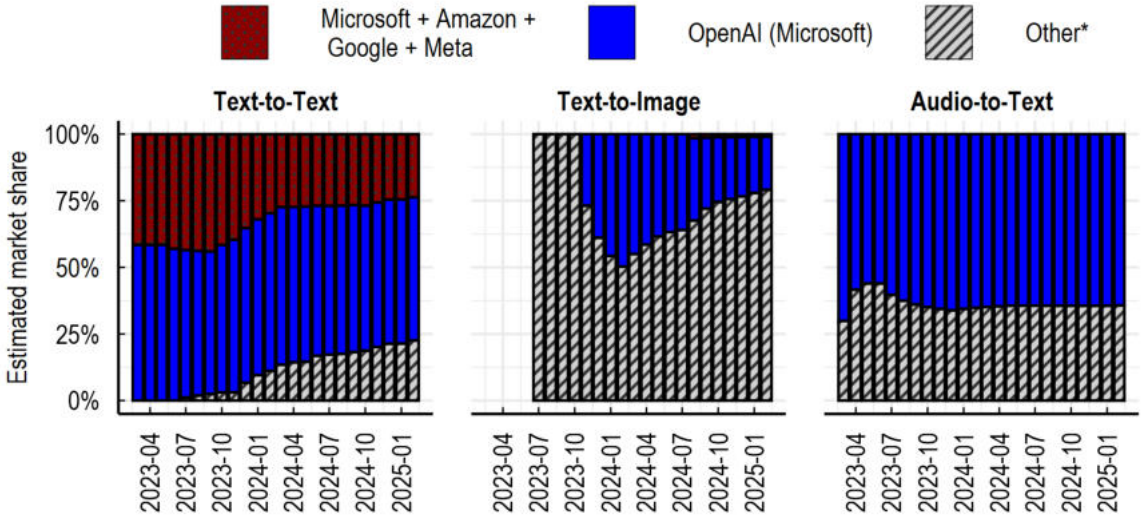
Note: Simulated market share of AI model revenues per region of the AI developing company for different calibrations reflecting several assumptions on the determinants of market structure as detailed in Annex C.
Source: authors' calculations.

Figure E.11. Simulated market shares by company status, AGI demand scenario under low switching costs



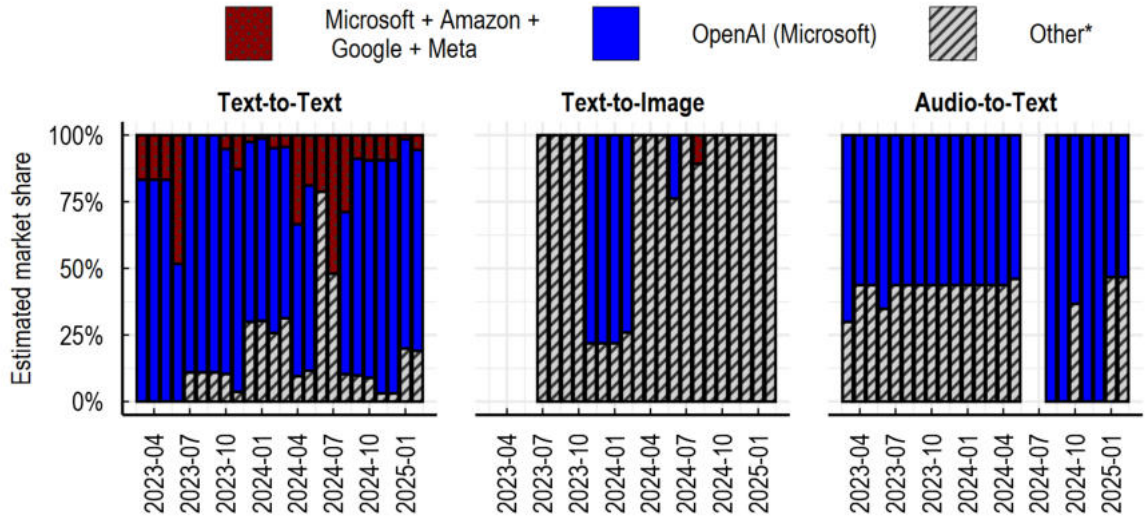
Note: Simulated market share of AI model revenues per region of the AI developing company for different calibrations reflecting several assumptions on the determinants of market structure as detailed in Annex C.
Source: authors' calculations.

Figure E.12. Simulated market shares by company status, Edge demand scenario under high switching costs



Note: Simulated market share of AI model revenues per region of the AI developing company for different calibrations reflecting several assumptions on the determinants of market structure as detailed in Annex C.
Source: authors' calculations.

Figure E.13. Simulated market shares by company status, Edge demand scenario under low switching costs



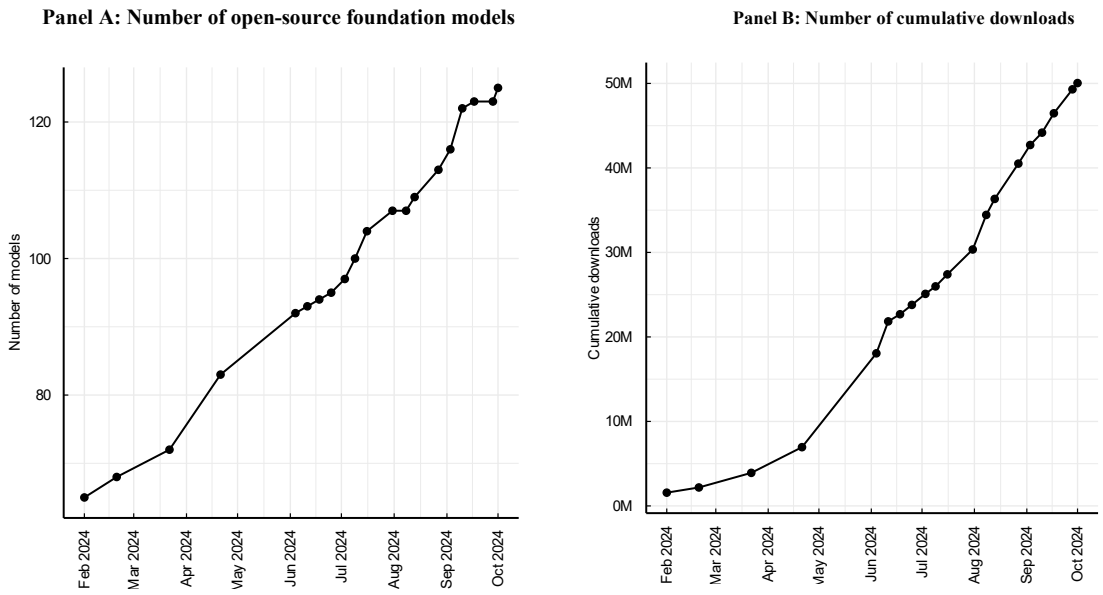
Note: Simulated market share of AI model revenues per region of the AI developing company for different calibrations reflecting several assumptions on the determinants of market structure as detailed in Annex C.
Source: authors' calculations.

Alternative method to simulate market shares using downloads of Open source models

Open-weight models can be downloaded from several AI (open source) platforms, such as *Kaggle* and *Huggingface,* or from the developer's website. On those platforms, many more models are available than

those available on demand from cloud providers. For example, in March 2024, close to 20,000 AI models were available from the platform *Hugginface*[51] (OECD.AI, 2024[10]). The local deployment includes open-source models for which data is publicly available and can be analysed, but it may also include local deployment of closed-licenced models. Figure E.14 shows the evolution of the number of open-source foundational models (Panel A) and the number of cumulative downloads on the Ollama platform.

Figure E.14. The rapid rise of open-source enables AI local deployment



Note: The figures display the cumulative number of downloads ("pulls") of open-source LLMs from *Ollama*. *Ollama* is an open-source platform that simplifies the installation and execution of LLMs locally on the user's hardware.
Source: Authors's calculation based on data from *Ollama*.

Popularity varies widely across open-source models. 75% of the downloads ("Pulls") are concentrated in the most popular models llama3.1 (Meta), llama3 (Meta), llama2 (Meta), Gemini (Google), Mistral (MistralAI). Those Opensource models available from cloud providers for direct inference and belong to the development frontier as defined in the previous sections.

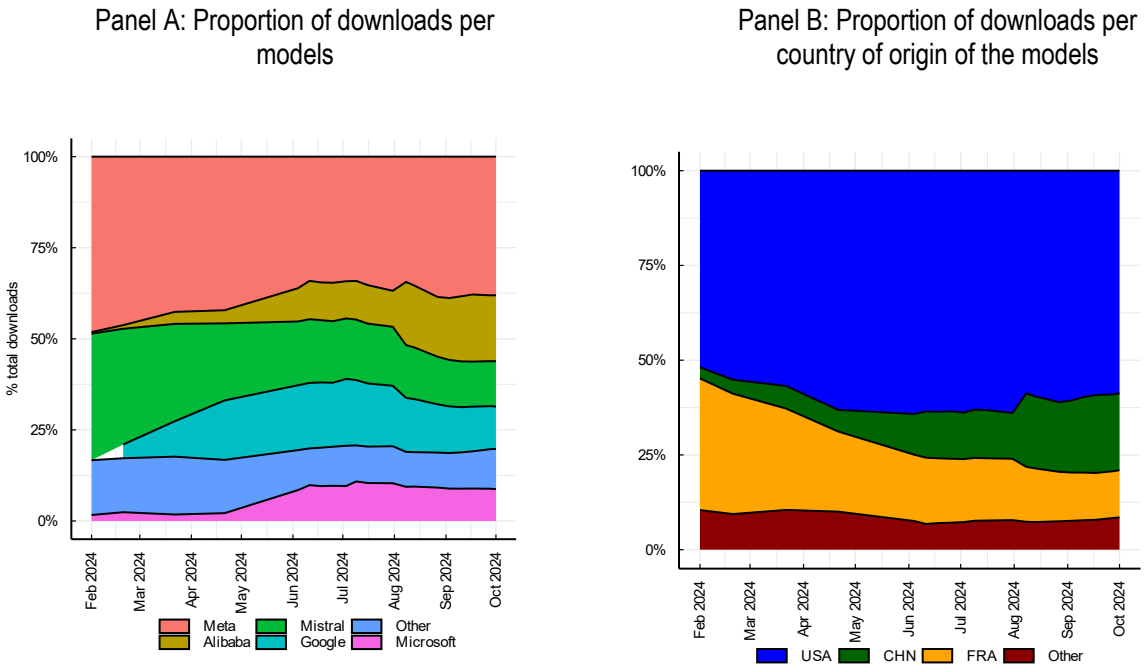
The market for the provision of open-source models is highly concentrated with 90% of downloads concentrated on models from five companies (Meta, Google, Mistral, Alibaba and Microsoft). Among the leading companies, only Mistral and Microsoft provide models strictly open-source licenced (Figure B2). It is interesting to note that only two companies (Meta and Mistral) provide foundation models that serve as the foundation for specific fine-tuned models by other companies, which are then available in open source on the Ollama platform. In September 2024, 13% of the downloads of models based on the llama 2, 3, and 3.1 foundation models were provided by a company that is not Meta (the parent company of llama models). For the Mistral model, this number is even larger (25%), suggesting that Llama and Mistral models are the leading models of the open-source innovation ecosystem as of September 2024.

In terms of country of origin (Figure E10 Panel B), the most popular open-source models from AI Labs come from the United States with more than 60% market share (mostly driven by models from Meta), with

[51] The number of AI models on Huggingface reflects mostly the adoption of AI by data scientists and individual developers that finetune open-source foundation models for specific usages, on specific and custom datasets.

the consolidation of the US since May 2024 coming from the popularity of Google (*Gemma*) and Microsoft (*Phi*) open-source models. While France (*Mistral*) is the second most popular country of origin it has lost significant market share (divided by two, from 37% to 15%). China has gained significant market share-between May and October 2024 (x5) and is challenging France as the second main player in open-source[52]. Those results are consistent with results reported using the market share methology based on the AI Economic Frontier defined in Annex C.

Figure E.15. Evolution of the market shares of AI models for local deployment



Note: The figures display the proportion of "Pulls" (downloads) per foundation model family (Panel A). Meta models include all models of the Llama family (2, 3, 3.1 and codellama) including the finetuned version by other providers than Meta. Similarly, Mistral includes both Mistral and Mixtral models as well as all the finetuned version of those foundation models. Google models include the Gemma family (1 and 2). Alibaba models include Qwen family (1 and 2). Microsoft models refer to the Phi family. Other include models based on llava, Deepseek, mxbai, nomic, yi,, Orca and others. Ollama is an open-source platform that simplifies the installation and execution of LLMs locally on user's hardware. As of October 2024, the platforms provide access to 120 LLMs and multimodal models. USA includes models from Meta (Llama), Microsoft (Phi), Google (Gemma), Snowflake (Artic), Databricks (Dbrx), Llava. FRA corresponds to models from Mistral. CHN includes models from Alibaba (Qwen), Deepskeers and 01.AI (yi).

[52] It is important to note that the data collection may be biased toward Western-based AI labs because downloads may not correctly reflect local deployment from Chinese platforms. The contribution to open source in software development and in particular in the field of AI is not a surprise as open-source development strategies are included in the PCR's Five-year plan in 2021 for the first time.

Annex F. A simple model of the price of AI inference

This section provides a simple structural model of the price of inference of LLM models following Quentin, Biderman and Schoelkopf (2023) and isolate three main components of the price per usage of AI: a software component, a hardware component and a demand component. The equation models the breakeven price of AI per usage p_m^*, i.e the price of inference charged for the use of AI that equates costs C_I and revenues R_I for the cloud provider, while assuming no fixed costs. Formally:

$$R_{I(p_m^*)} = C_{I(p_m^*)},$$

where $R_I =$ revenue generated from the inference of the model and $C_I =$ variable costs generated by the inference, and

$$R_I = p_m . Q_I . \gamma . w \quad \text{and} \quad C_I = X_H . N_{max} . p_I$$

with p_m capturing the price of the model per million tokens, Q_I stands for the number of prompts (model requests) per hour, γ capturing the size (in million tokens) of the prompt, *w* capturing the number of users. X_H equals the number of GPUs necessary to run AI model inference (ie using the model), N_{max} is the total number of users that can be served, p_I is the cost per hour of running the server that hosts the model.

The breakeven price p_m^* can be derived as follows and is determined by cost related characteristics (compute power need X_h and compute power price P_I) and demand (captured by capacity utilisation *CU* and the intensity of use *IU*):

$$p_m^* = \frac{X_H . p_I}{CU . IU} = \frac{Cost\ parameters}{Demand\ parameters},$$

In more detail, these variables are given as follows:

- $X_H = \frac{S_m}{\beta . G_H} =$ Computing capacity relative to model parameter size
  - S_m = size of the AI model in billion parameters (i.e llama3 70B)
  - $\beta = \frac{32}{1,2.4.Q\_b} =$ Hardware configuration following the approximation in Quentin, Biderman and Schoelkopf (2023)
  - $G_H =$ The capacity of the GPU in giga-bytes (GB) that is a function of the price (i.e H100 ~ 80GB)
- $p_I =$ compute cost of inference
- $\frac{1}{CU.IU} =$ Demand adjustment parameter
  - $CU = \frac{w}{N_{max}} =$ Capacity of utilisation in terms of user numbers (*extensive margin)*: Number of users relative *w* to the number of possible simultaneous users N_{max}. The number of users is a function of the performance of the model which is in turn a function of compute.

- $IU = \gamma . Q_I$ = intensity of utilisation: size (in million tokens) γ times the number of questions (per hour) Q_I, that can be assumed to be a decreasing function of the quality of the answers (fewer prompts are needed to get the same quality answer).[53]

In turn, the breakeven price can be expressed as follows:

To formally introduce our econometric specification, we first derive the theoretical breakeven price of AI inference following (Quentin, Biderman and Schoelkopf, 2023) with the following equation (see Annex for details):

$$P_m^* = \frac{S_m}{\beta} \cdot \frac{P_I}{G_H} \cdot \frac{1}{CU.IU} \quad (21)$$

Equation 21 shows that the prices charged for AI depend positively on the cost of compute per hour (variable costs of compute) and negatively on demand. The quality of the hardware has an ambiguous relationship as on the one hand it impacts the compute costs per hour, but also the number of GPUs necessary to host the models, as well as the maximum number of clients that can be served simultaneously[54]. The costs of compute include electricity costs, cooling costs, maintenance of the hardware, and broader costs of data clusters like the price of land and the depreciation cost of the hardware. The multiplicity of global and local factors explains the diversity of the price offer of AI on top of the difference in model capability. It also highlights that while current prices are largely unique per model and provider, some evidence of price discrimination across regions is observed and could be the norm in the future.

Transforming Equation (21) into logarithmic form provides a linear relationship between the price of AI inference and the cost parameters of the model that can be estimated by OLS using the data presented in section 2.3.

$$\log (P_{m)}^*) = \log\left(\frac{S_m}{\beta}\right) + \log\left(\frac{p_I}{G_H}\right) - \log(CU) - \log(IU) \quad (22)$$

Equation (22) shows that the breakeven log price of inference can be decomposed into three linear components. First, the *software price component* captured by the term $\log\left(\frac{S_m}{\beta}\right)$ that includes the sizes of the model in billion parameters and the configuration of the models. The *scaling law* (Hoffmann et al., 2022) relates the size of the model and its performance, with an optimal relationship also between the size of the training datasets. Consequently, we can approximate the model size by its performance $\frac{S_m}{\beta} \sim Q$, with the final performance of the model dependent on the quality of the training data, the modelling choices, and the specific know-how of the developing company[55].

Second, the *hardware price component* captured by the term $\log\left(\frac{p_I}{G_H}\right)$ that includes the variable compute cost of running a GPU instance to serve one inference point and the size of the memory of the GPU that is used. As models get larger (involving more parameters), higher quality hardware is required to serve the model, which increases the cost of running the hardware and serving the requests to the model. The costs

[53] The intensity of utilisation is also determined by the speed of the model.

[54] It is not modelled in this equation but is also essential for the long-run price, as it includes the models' depreciation costs and the provider's margin.

[55] While the model size is known for open-source models, it is usually not disclosed for closed (proprietary) models.

mainly depend on the provider of the model and the quality and price of the hardware chosen to serve the model.[56]

The third set of factors are related to demand. Since revenues depend on the use of the models (number of tokens generated) and costs are largely independent of the number of users, the breakeven price is such that the per-client price should depend on capacity utilisation *CU,* and the intensity of utilisation *IU*. In other words, when demand is high, the price that equates revenues and costs declines as the costs of the infrastructure are shared among more clients.

In the regressions linking model price to model and provider characteristics (Section 3.3) several of these variables are unobserved. However, they are partly controlled for by provider- and time fixed effects. that respectively control for the characteristics of the hardware and compute infrastructure used by the provider and the general market trend in the cost of compute and the increasing adoption rate and infrastructure investments.

[56] Since the quality of the hardware for AI (GPU) has improved over time and is dominantly supplied by NVIDIA chips, on could potentially approximate the quality of the hardware (G_H) and the price of compute (p_I) by the number of GB of memory and the price of renting the best-in-class hardware available at the moment of the release of the model.