SUERF
The European Money and Finance Forum

SUERF Policy Brief
No 1219, July 2025

Unpacking the AI supply chain: market structure and policy implications


Leonardo Gambacorta | Bank for International Settlements (BIS) and CEPR
Vatsala Shreeti | Bank for International Settlements (BIS)

Keywords: Artificial intelligence (AI), generative AI, market structure, big techs, competition, financial stability, cyber risk

JEL codes: E31, J24, O33, O40

Abstract

Behind the rapid advancement of artificial intelligence (AI) lies a complex supply chain containing five layers: hardware, cloud infrastructure, training data, foundation models and AI applications. The economic forces shaping the market of these layers include high fixed costs, economies of scale, network effects, rapid technological progress and strategic actions by big technology companies. The way that AI is provided today influences not only consumer welfare but also operational resilience, cyber security and financial stability. Building an inclusive and resilient AI ecosystem is paramount for aligning technological progress with broader social goals.

Disclaimer: This policy brief is based on BIS paper no 154. The views expressed in this policy brief are those of the authors and do not necessarily reflect those of the Bank for International Settlements.

Elements of the AI supply chain

Artificial intelligence (AI) applications rely on a supply chain made up of multiple layers, each with distinct market dynamics. This AI supply chain has, at its core, five interconnected layers (Graph 1). The first is hardware, which includes specialised chips such as graphics processing units (GPUs), field-programmable gate arrays (FPGAs) or application-specific integrated circuits (ASICs). These specialised chips are especially well suited to AI as they excel at performing thousands of complex computations simultaneously. GPUs are most commonly used for training and inferring from AI models.

The next layer is cloud computing. Cloud computing platforms like Amazon Web Services (AWS), Microsoft Azure and Google Cloud provide the computational backbone for training, storing and deploying AI models. The third layer is the training data, which are the lifeblood of large AI models. Training data can include text, audio, video and images from both public and proprietary sources. These data are then fed into the fourth layer, large foundation models, which can be adapted for many different functions and applications.

The last layer contains user-facing AI applications, like ChatGPT, Claude, Gemini, FinGPT, DALL-E or Github copilot. These applications leverage foundation models to deliver diverse functionalities, ranging from natural language processing, financial forecasting or coding assistance to image generation and protein fold prediction.

The AI supply chain — Graph 1

Source: Gambacorta and Shreeti (2025).

Economic forces shaping the input layers

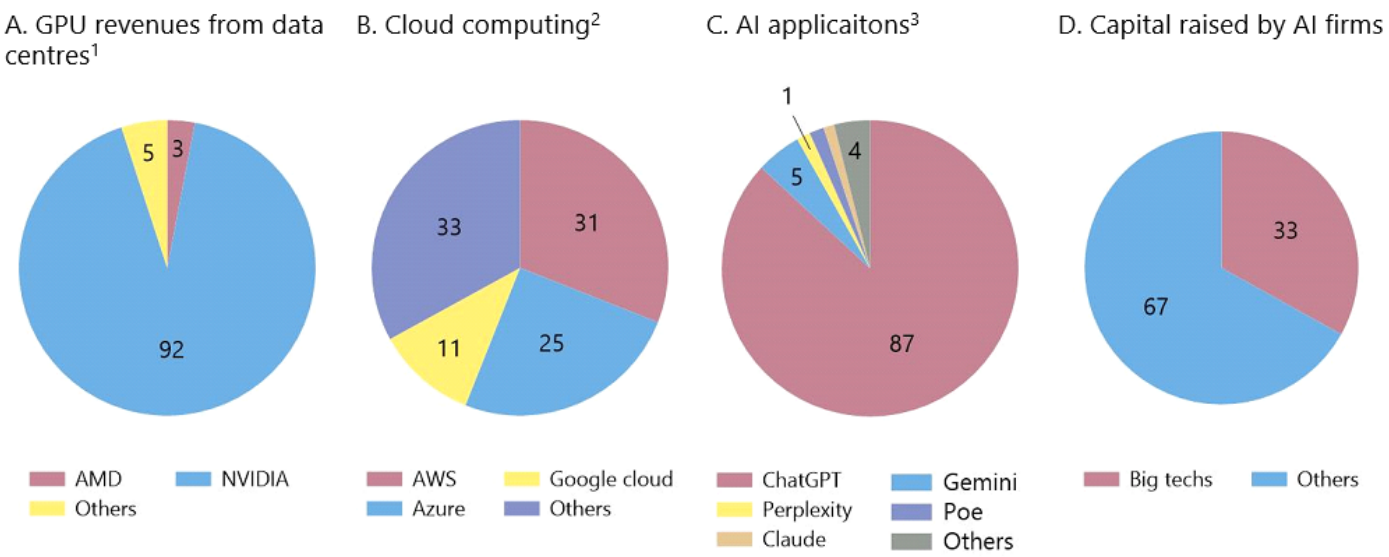
The market structure of each layer looks different, with varying degrees of competition and concentration. The hardware layer, which refers to the production of specialised chips like GPUs, is largely dominated by Nvidia, which controls over 90% of the market (Graph 2A). Nvidia has gross margins of over 70% and its revenues increased by 405% between 2023 and 2024 (Nvidia press release, 2024).

Two key factors explain this success. First was Nvidia's early entry into the AI space. Nvidia's GPUs were originally made for the video gaming market, but their parallel computing capabilities proved to be crucial for AI applications. The second factor behind Nvidia's dominance in the hardware layer is the edge that its parallel computing platform, CUDA, provides. Nvidia's GPUs come in an exclusive bundle with CUDA, which has become the industry standard, enabling programmers and software developers to simplify the process of using GPUs and to enhance their performance. While competitors such as AMD, Intel and some Chinese firms have entered the market, Nvidia's entrenched advantages, coupled with high switching costs for users, make it difficult for rivals to challenge its dominance.

The cloud computing layer is similarly concentrated, with three major providers – AWS, Microsoft Azure and Google Cloud – accounting for nearly three-quarters of the global market (Graph 2B). This dominance is underpinned by high fixed costs, strong network effects, and significant barriers to switching, such as costly egress fees and proprietary software restrictions. Additionally, these providers often bundle their cloud services with other offerings, further reinforcing customer lock-in and limiting competition.

Market structure of the AI supply chain

In per cent — Graph 2



[1] Based on global revenues of GPU producers for GPUs used in data centres in 2023. [2] Based on global cloud computing revenues for Q1 2024. [3] Based on monthly visits data. For further details see Liu and Wang (2024). [4] Based on total capital invested in 2023 in firms active in artificial intelligence & machine learning collected by the authors from Pitchbook. Big techs correspond to Alibaba cloud computing, Alibaba group, Alphabet, Amazon industrial innovation fund, Amazon web services, Amazon, Apple, Google cloud platform, Google for startups, Microsoft, Tencent cloud, Tencent cloud native accelerator and Tencent holdings.

Sources: Liu, Y and H Wang (2024): "Who on Earth Is Using Generative AI?", *World Bank Policy Research Working Paper*, no 10870; IoT Analytics Research (2023), *Generative AI Market Report 2023–2030*; PitchBook Data Inc; Statista; authors' calculations..

The training data layer, while less concentrated than hardware and cloud computing, is increasingly tilting in favour of larger firms. Historically, AI models have relied on publicly available data, but as these datasets become scarcer, companies with access to proprietary user data will likely have an advantage. While data feedback loops can amplify these advantages, the extent of their effectiveness varies by application, and there may be diminishing returns to additional data in certain contexts.

The foundation model layer seems more contestable, with over 300 models in the market (CMA (2024), Korinek and Vipra (2025)). The development of foundation models requires significant upfront investment and access to vast computational resources, potentially creating high barriers to entry.[1] Economies of scale and scope further reinforce the market structure, while user inertia can make it difficult for new players to gain traction. As a result of this, while there are several hundred foundation models and the dynamics of the market are rapidly evolving, only a handful of firms produce them. Currently, OpenAI, Google DeepMind, Anthropic and Meta still dominate this space.

[1] While training foundation models typically involves substantial costs, there are some signs that these expenses may be decreasing. One example of potentially declining costs is the launch of DeepSeek, a Chinese foundation model reportedly trained at significantly lower cost than GPT-4 (Financial Times, 2025 a,b).

Finally, the user-facing AI applications layer is thriving, with a wide range of tools and services built on foundation models. However, within each narrow market, there is a chance of "winner takes all" dynamics emerging. For instance, despite the rise of competing chatbots, ChatGPT captured 60% of the market in 2024, underscoring the importance of being an early entrant (Graph 2C).

Big techs are everywhere

A key feature of the AI supply chain is the expanding influence of big tech companies in every input layer. These companies are leveraging their existing market power in digital services to expand their influence in emerging AI markets. In 2023, big techs accounted for a significant share of global AI investments, contributing 33% of the total capital raised by AI firms and 67% of the funding for generative AI startups (Graph 2D). Their deep financial resources, control over data and established cloud infrastructure give them a competitive edge, enabling them to integrate vertically across the AI supply chain.

Three big techs dominate the cloud computing market and often use exclusivity agreements to strengthen their position. For instance, Microsoft's partnership with OpenAI requires the latter to rely exclusively on Azure cloud services, while Amazon's investment in Anthropic comes with a similar condition to use its cloud infrastructure. Such arrangements limit competition by locking AI startups into specific cloud providers, further consolidating the dominance of big techs in this layer.

In the training data layer, big techs have a significant advantage due to their access to proprietary datasets generated by their platforms. Companies like Google, Meta and Microsoft are using data from services such as Gmail, Facebook and LinkedIn to train their AI models. As the availability of high-quality public data declines, proprietary data become increasingly valuable. Big tech firms are also acquiring companies with unique data assets and modifying their privacy policies to expand their access to user data. For example, Google recently updated its terms of service to allow data from Google Docs and Google Maps to be used for AI training.

Big techs are also expanding their activities in other layers of the supply chain. They are building their own hardware and foundation models, integrating user facing AI applications with their existing services and even securing their own sources of nuclear power to run energy-hungry data centres. This vertical integration creates a "cloud-model-data loop", where big tech firms can use their dominance in one layer to strengthen their position in others. For instance, their control over computational resources, proprietary data and cloud infrastructure can allow them to produce better AI models, which in turn generate more data for training future iterations.

A distinct aspect is the potential impact on financial stability: as AI adoption expands in the financial sector, reliance on the same data and models can amplify systemic risks during periods of financial stress, creating procyclical behaviour or herding effects. Most fundamentally, a few firms' dominance over AI gives them disproportionate influence over the direction of future innovation, increasing the risk of misalignment between private incentives and social welfare (Acemoglu, 2021).

Towards an inclusive AI ecosystem

Addressing these risks is a challenging task, but several policy measures are on the table. Promoting data sharing will be key. Open data initiatives and the creation of public datasets for model training can help level the playing field for smaller firms and reduce reliance on proprietary datasets held by big tech companies. Investments in open-source foundation models can encourage competition and allow smaller players to participate in the AI ecosystem. Cloud computing is one of the main bottlenecks in the AI supply chain, and measures targeted at reducing switching costs for end users can go a long way. Given the global nature of AI markets, international cooperation remains essential, even if it is challenging. Harmonising regulatory frameworks and sharing best practices can help address cross-border challenges and ensure a fair and competitive market. In the meantime, regular monitoring of market dynamics is crucial.

References

Acemoglu D. (2021), "Harms of AI", NBER Working Paper 29247.

Biglaiser G., J. Cremer and A. Mantovani (2024), "The economics of the cloud", TSE Working Paper 1520.

CNBC (2024), "Why Big Tech is turning to nuclear to power its energy-intensive AI ambitions", 16 October.

Competition and Markets Authority (CMA) (2024), "AI foundation models: update paper", April.

Financial Times (2025a), "OpenAI says it has evidence China's DeepSeek used its model to train competitor", 29 January.

Financial Times (2025b), "DeepSeek's 'aha moment' creates new way to build powerful AI with less money", 29 January.

Gambacorta, L. and V. Shreeti (2025), "The AI supply chain", BIS Working Paper 154.

Hagiu, A. and J, Wright (2025), "Artificial intelligence and competition policy", *International Journal of Industrial Organization*, January, 103134.

Korinek, A. and J Vipra (2025), "Concentrating intelligence: scaling and market structure in artificial intelligence", *Economic Policy*, *40*(121), pp.225-256.

About the author(s)

Leonardo Gambacorta is the Head of the Emerging Markets unit at the Bank for International Settlements. Prior to his current role, he served as Head of Innovation and Digital Economy, Research Adviser and Head of Monetary Policy in the Monetary and Economic Department. His primary research interests include monetary transmission mechanisms, the effectiveness of macroprudential policies in curbing systemic risk, and the effects of technological innovation on financial intermediation. He is a research fellow of the Centre for Economic Policy Research.

Vatsala Shreeti joined the BIS in September 2022. She is broadly interested in empirical industrial organisation and digital economics, with a focus on emerging market economies. She is a research affiliate of CESifo. Her recent work has focused on the adoption of digital technologies in emerging markets, design and pricing in fast payment systems, AI in finance, and incentives for innovation in high-technology industries. She holds a PhD in Economics from the Toulouse School of Economics and a Masters in Economics from the London School of Economics and Political Science.

SUERF Policy Notes and Briefs disseminate SUERF Members' economic research, policy-oriented analyses, and views. They analyze relevant developments, address challenges and propose solutions to current monetary, financial and macroeconomic themes. The style is analytical yet non-technical, facilitating interaction and the exchange of ideas between researchers, policy makers and financial practitioners.

SUERF Policy Notes and Briefs are accessible to the public free of charge at https://www.suerf.org/publications/suerf-policy-notes-and-briefs/.

Editorial Board: Ernest Gnan, David T. Llewellyn, Donato Masciandaro, Natacha Valla

Designed by the Information Management and Services Division of the Oesterreichische Nationalbank (OeNB)

SUERF Secretariat
c/o OeNB, Otto-Wagner-Platz 3A-1090 Vienna, Austria
Phone: +43 1 40 420 7206
E-Mail: suerf@oenb.at
Website: https://www.suerf.org/