



The AI supply chain

Leonardo Gambacorta (BIS) and **Vatsala Shreeti***

Economist, Innovation and the Digital Economy, Bank for International Settlements (BIS)

24 October 2025, The Digital Economy Amid Rising International Tensions, Banca d'Italia

* The views expressed here are those of the presenter and not necessarily those of the BIS.

How are AI applications provided?

The AI ecosystem



The AI ecosystem



Hardware

- Specialised AI chips such as field-programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs) and GPUs.
- GPUs are the most widely used, especially during the training phase of AI model development. Thousands of GPUs required for training AI models, as well as for inference.
- It takes **eight GPUs for Microsoft Bing to answer a single question** in less than one second (CNBC, 2023)

The AI ecosystem



Cloud computing

- Range of on-demand services, including data and model storage, processing, computation and analytics; can be accessed remotely.
- Three main service models for cloud computing: i) software as a service (SaaS), ii) platform as a service (PaaS) and iii) **infrastructure as a service (IaaS)**.
- AI models are typically trained and stored using the IaaS model.

The AI ecosystem



Training data

- Datasets including text, audio, video and images from both public and proprietary sources.
- Training data could also be synthetic.
- Automated or human-driven labelling to make it suitable for training AI models.

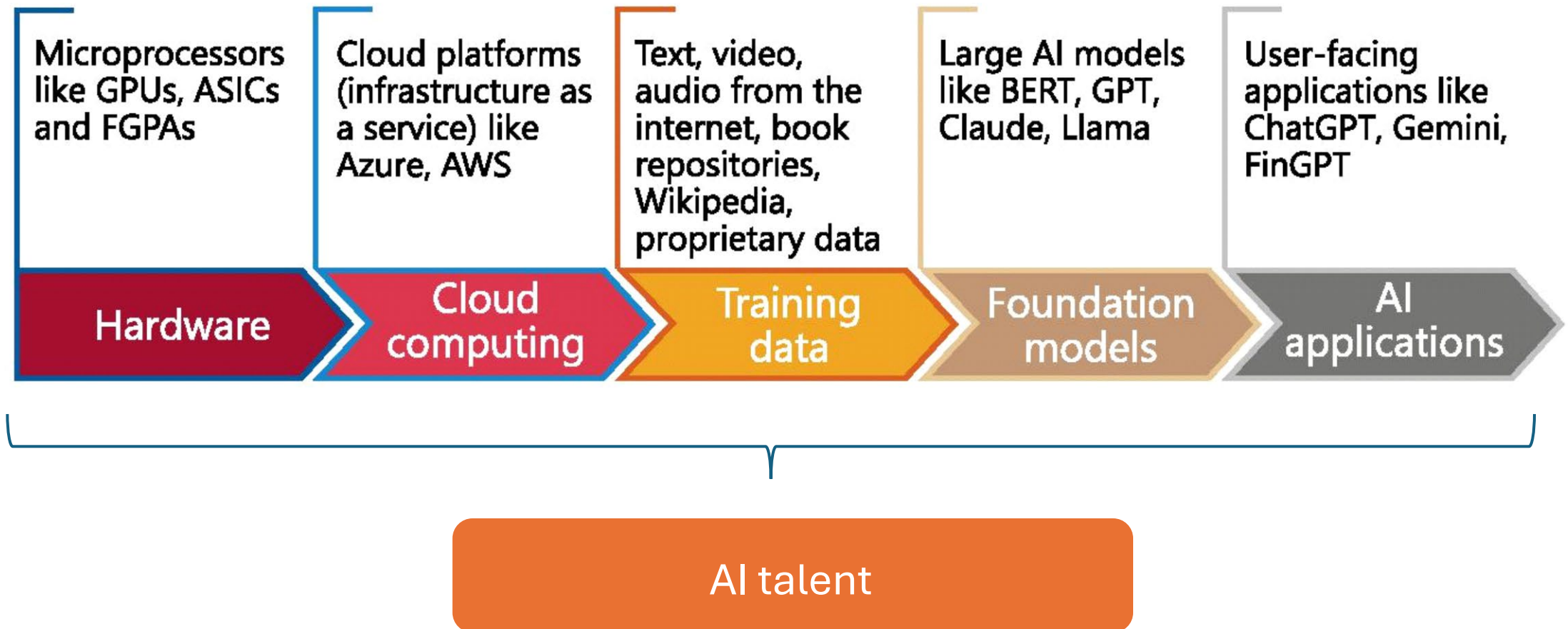
The AI ecosystem



Foundation models

- Large AI models that can be adapted for various functions and applications.
- Performance of a foundation model depends not only on its technical architecture but also on the volume and quality of the training datasets

The AI supply chain: human capital

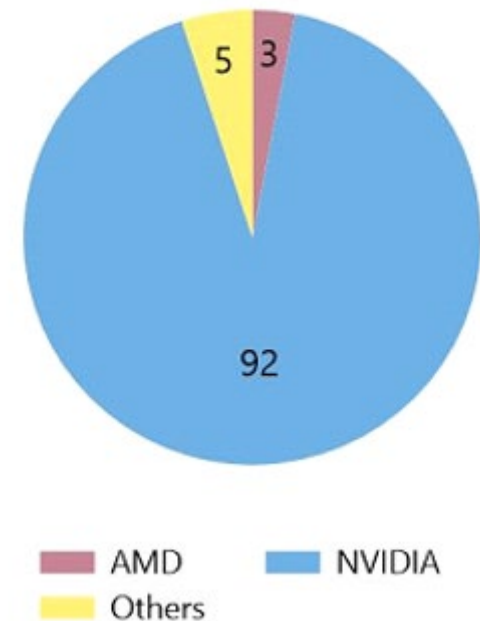


What does the market look like in each layer of the AI supply chain?

Hardware

- Nvidia serves **most of the global market** with gross margins of over 70%.
- Initially serving the video game market, first to leverage the **parallel computing capacity** of its GPUs for AI models.
- Nvidia's GPUs come in an **exclusive bundle with CUDA**, its parallel computing platform.
- CUDA has become the **industry standard for programmers** and can only be used with Nvidia's GPUs.
- **Strategic acquisition** of Mellanox in 2019, a technology company that provides the architecture to connect GPUs in a network.
- First-mover's advantage and benefits from **exclusively bundling** CUDA with its GPU offerings.

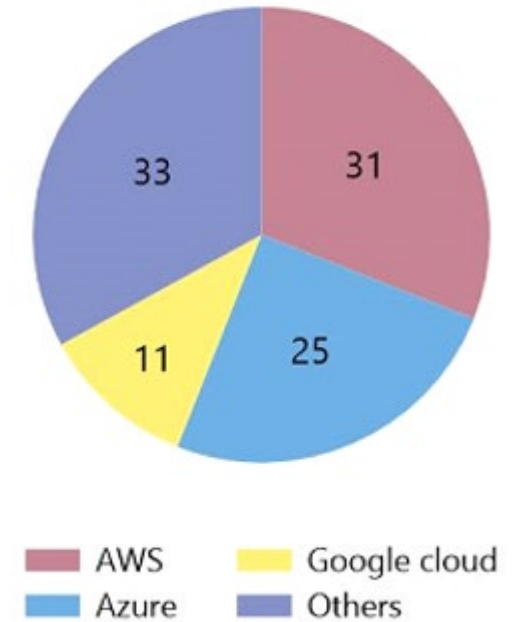
A. GPU revenues from data centres¹



Cloud

- Globally, the cloud computing market is dominated by **three big tech companies** : Amazon Web Services (31%), Microsoft Azure (24%) and Google Cloud Platform (11%).
- For IaaS, the **market is even more concentrated**, 74% market share globally of AWS, Microsoft and Google.
- High **switching costs** for end-users; switching difficult without extensive **retraining** of engineers.
- Egress fees: a **cost on users for transferring data** out of the cloud to a rival platform.
- Ecosystem of **vertically integrated services**.
- High fixed costs, lack of interoperability and significant **direct and indirect network effects** in usage.

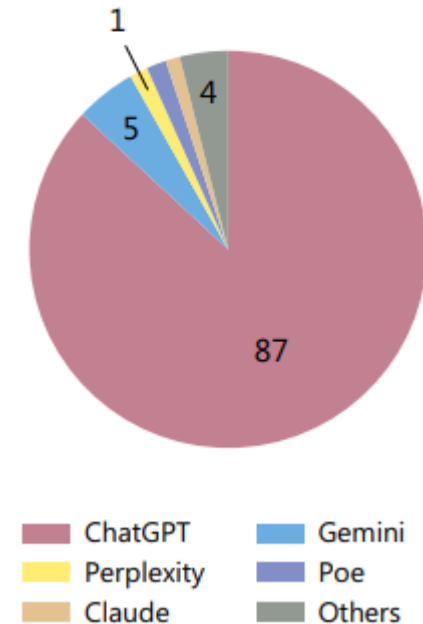
B. Cloud computing²



Foundation models and AI applications

- There are over **300 foundation models** in the market, provided by **14 different firms** (CMA, 2024; Korinek and Vipra, 2024).
- Competing business models: **open-source** (Llama, DeepSeek) vs **proprietary** (OpenAI).
- Nevertheless, market for foundation models currently is **dominated only by a handful of firms** like OpenAI, Google DeepMind, Anthropic and Meta.
- **Economic forces:** high fixed costs & low variable costs, economies of scale and scope, competition “for” the market, vertical integration.
- AI application layer: many downstream applications, but questions about the market structure within each market. **Winner takes all dynamics?**

C. AI applications³

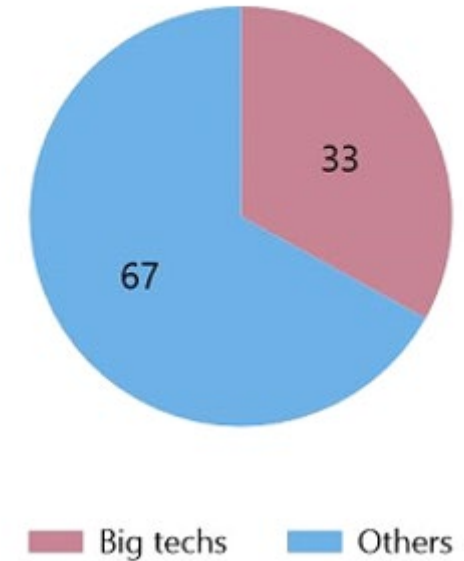


Big tech footprint in the AI supply chain

Big techs to big AI ?

- Big techs are **investing heavily in AI**: in 2023, they accounted for 33% of the total capital raised by AI firms, and nearly 67% of the capital raised by generative AI firms.
- They are also **vertically integrating** across all layers of the AI supply chain, especially the big cloud providers.
- Active in the **foundation model layer**, either through partnerships or as producers of foundation models themselves. Amazon, Google, Meta and Microsoft also **active in the hardware layer**.
- Unique position as **publicly available data run out**. Own pool of proprietary data, updating privacy policies and terms of use, acquisitions of data owners.
- A new **cloud-model-data** loop?
- Countervailing forces on data: **diminishing returns** from additional data, not all data will reinforce this loop.

D. Capital raised by AI firms



“Acquihires”

Billion-dollar ‘acqui-hires’ are bad for competition

From Google to Microsoft to Meta, some of the biggest tech deals of the past few years have been staff moves

INNOVATION > VENTURE CAPITAL

How ‘Acquihires’ Are Reshaping Silicon Valley’s AI Investments

Big Tech's acquihire deals face regulatory scrutiny, outgoing EU antitrust official says

What are the consequences for financial stability and other outcomes?

Potential impact on economic outcomes

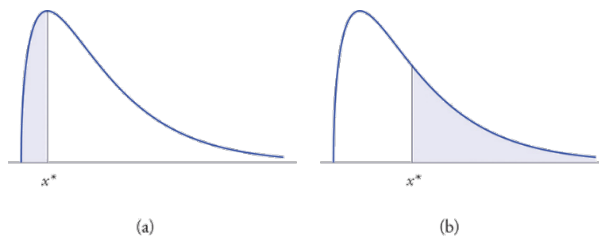
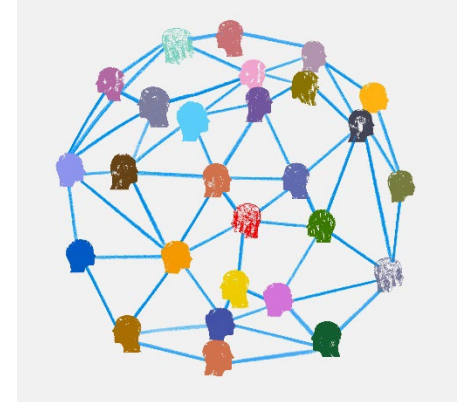
- **Consumer welfare:** limited consumer choice, lock-in, restricted access of smaller firms, for example, TSMC in 2021.
- Direction of **innovation** controlled by major cloud providers/big tech, **risk of misalignment** between socially desirable innovation and privately profitable innovation.
- Operational resilience of **critical infrastructure**. E.g. CrowdStrike incident, 2024.
- As AI model use is expanding in finance, a concentrated supply chain can create **systemic risk**.
- Use of **similar algorithms, datasets and models** can lead to flash crashes, market volatility and illiquidity during times of stress.

Financial stability and AI



Early rule-based systems were already important for financial stability: 1987 US stock market crash

Machine learning increases network interconnectedness, data uniformity, model herding.



[This Photo](#) by Unknown Author is licensed under [CC BY-SA-NC](#)

Gen AI leads to the fat tail problem, third party dependencies, model herding and uniformity.

Why is regulation hard?

Path ahead is not straightforward

- Achieving consensus on policy is difficult as the AI supply chain contains **many different markets that fall under the ambit of different regulatory authorities**, often with competing goals.
- International cooperation even more elusive, jurisdictions **differ in their legal frameworks, geopolitical goals and regulatory appetite**.
- Pace of technological progress in the field of AI is usually much faster than regulatory capacity.
- Need to have a constantly **evolving skillset** among policymakers and regulators.
- Antitrust measures are applied **ex post**, need to balance static effects on competition with dynamic effects on future innovation.



Thank you. Questions?
Vatsala.Shreeti@bis.org

Possible actions

Public data sets for model training?

Non-discrimination requirements for access to foundation models?

Antitrust: evidence collection

Sharing best practices?

Data sharing frameworks?

Multi-cloud strategies? Common APIs?

Harmonising regulatory frameworks?