AI frontrunners will benefit most, with Microsoft in the lead

Large technology firms are investing a lot to become AI services leaders. They are active across the AI value chain and will likely work to make their foundational model an industry standard. We expect that many sector-specific AI solutions will emerge. Given economies of scale, early leaders will likely benefit most

The hyperscale data centre market is currently dominated by four large technology companies in the West. These companies are well set to be leading the AI revolution. Moreover, we expect them to try to take control of the full value chain, which will enable them to extract maximum value. We expect the sector to grow by 10% per annum in the near future, in line with their electricity needs.

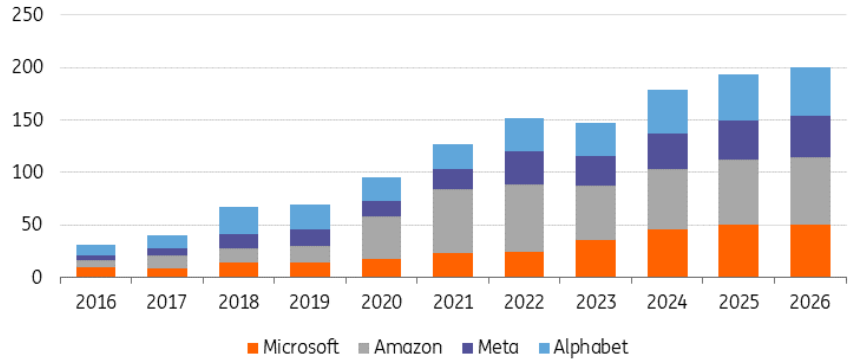
Although we expect to see many different business models, we see Microsoft as the market leader providing actual services. We expect the existing large cloud operators to consolidate substantial market share in the domain of Generative AI. Also, high capital investments provide a barrier to entry for followers. Outside the hyperscale operators, Nvidia also has a very strong AI offering, as explained in a related article.

The market leaders ramp up their asset base

The largest American technology companies should be able to benefit greatly from generative AI. To deploy Generative AI on a worldwide scale, we have recently seen a sharp increase in investment in digital infrastructure by Amazon, Alphabet, Meta and Microsoft. Digital infrastructure investments are a broader category than investments in Generative AI, but they do contain investments in Generative AI because this requires ultrafast microprocessors, and data centres. Counterpoint Research estimates that in 2023, roughly 13% of Microsoft's digital infrastructure spending was specifically for Generative AI. This percentage will likely increase in the future, as companies compete with their AI offerings.

Investments in data digital infrastructure are booming

Investments in digital infrastructure by Amazon, Microsoft, Meta and Google (US$, bn)



Source: Refinitive Eikon, ING (Microsoft capex held constant from 2025 to 2026 because of limited data availability)

Therefore, their investments are expected to keep growing for the foreseeable future. The capital investments by the four large operators of hyperscale data centres exhibit a compound annual growth rate of 11% between 2021 and 2025, where the 2025 number is based on consensus estimates. But note that the hyperscale data centres are expected to grow faster than smaller data centres, as we explained in our article on data centres.

Leading Generative AI developers and examples of AI models

OpenAI	GPT-4 Turbo GPT-4 GPT-3	Mistral	Mistral Large Mistral Medium Mistral Next
Alphabet	Gemini Ultra Gemini Pro Gemini Nano Palm -2	Anthropic	Claude 3 Opus Claude 2.1 Claude
Meta	LLaMA-3 LLaMA-2 LLaMA-1	Hugging Face	BLOOM

Source: ING

Foundational models are a key building block for AI applications

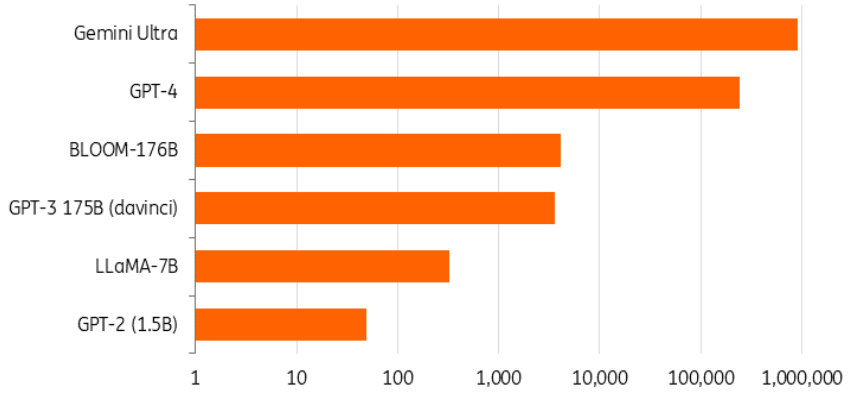
Breakthrough innovations in the field of AI have come in the form of some sophisticated foundational models which are general-purpose technologies. These general-purpose models can be used for a diverse range of use cases. As shown above, the development and training of these models require huge amounts of data and a vast infrastructure. There are quite a few prominent foundational models that can analyse and generate, for example, images, music, and languages. Prominent developers of foundational models are: OpenAI, Alphabet, and Meta, as can be seen in the table above. These companies are well-placed to lead the innovation race. This is important, as the technology sector typically benefits from economies of scale. In short, the company that comes first with the best technology has the best chance to take a leading, dominant, market share. This is because it makes the most money (which it can reinvest), but it also benefits from the interaction with customers and past experiences to further refine the service. This explains the race to have the leading foundational model. Below we show the computational requirements for a selection of models. Notably, the requirements went up from the old GPT-2 model to a more modern version, GPT-4. With it came a spectacular increase in user experience. And we have not seen all of it yet, as larger models will be around soon.

Many AI solutions are scalable

The above mainly discusses the large, breakthrough Generative AI models. Nevertheless, there also exist smaller (legacy) models, while companies are also working on smaller models that can be used for specific tasks. These smaller models can run on standalone servers and smartphones. The advantage of such solutions is that the operator has full control over the data that goes in and out of the system, which makes it more palatable for processing private data. Software providers, such as Alphabet, Meta and OpenAI offer families of models, offering different degrees of complexity. Notably, there are (trained) open-source models that can be used by developers to create their own, tailored, solutions.

Required time to train models increases with complexity

(Petaflop/s-days, exponential scale)



Source: Our World in Data, ING

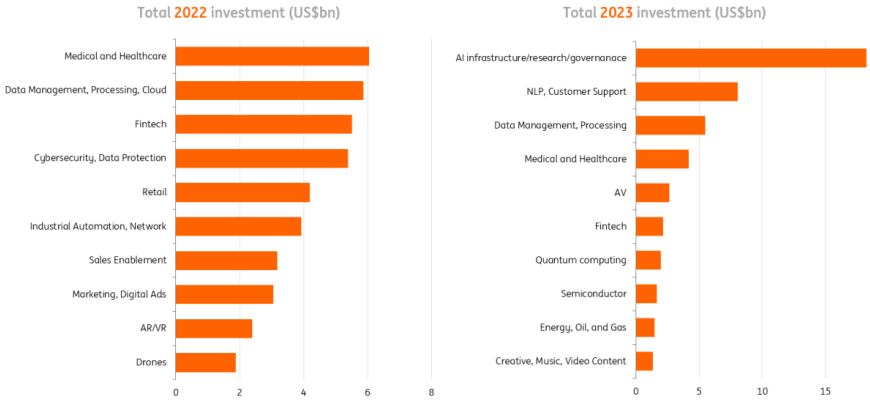
Server manufacturers will also find new opportunities as the AI market segment opens up. Server manufacturers are, for example, Hewlett Packard Enterprise, Dell Technologies and Super Micro Computer. As an example, Dell Technologies and Meta are selling a proposition (including servers and software) that can be used by corporations to run their own, private AI models for software development.

Asking a remote AI application for help is also possible. Today, the capabilities of low-tech hardware can be enhanced through communication with the intelligent cloud, running state-of-the-art AI models. Through a relatively simple device, a customer can send a request to a data centre that can perform more complex calculations or analysis. This operating model can therefore bring AI services to local, smaller, devices. What is needed are excellent, low-latency, communication networks, such as those based on the 5G standards. AI therefore provides another incentive to complete these networks, because they are an enabler of smart Internet of Things applications. Such developments are part of a recent transaction between Microsoft and Vodafone, and we expect many telecom companies and smartphone manufacturers to bring innovative services and applications in the future, including smartphones capable of running AI software.

It has proven difficult in history to alter business models

For companies, it is important not only to focus on the ongoing technical developments that are needed to pursue the full opportunities of AI but also on the business models. As we saw previously, it is sometimes difficult to embrace technological change in the right way. Two examples from a 2019 Forbes magazine article illustrate this. The industrial conglomerate GE experienced failure when it tried to build GE Digital. They wanted to develop an IoT service model around the traditional industrial equipment division, but the project was not managed properly, at a huge cost. Also, automaker Ford created a digital division, separate from its traditional automotive division. This also resulted in disappointment because the digital products were not well integrated. Before investing in new AI-related product developments, companies therefore should focus on the business case of a new, AI-inspired, product value proposition. A first example, with a good chance of success, is Copilot for Microsoft 365, a subscription-based revenue model.

Substantial investments are made to develop sector specific solutions



Source: Quid 2023; Netbase Quid 2022 taken from the Stanford University Artificial Intelligence Index Reports 2022 and 2023 (largest amounts only)

Sector specific applications are being developed

It is very likely that we are going to see the emergence of many AI-driven solutions that will be derived from the foundational models, such as user interfaces, chatbots or dedicated tasks for the business services industry. As can be seen in the figure above, this is becoming a very large industry. Investors see large opportunities in sectors like Medical and Healthcare, Customer Support, Data Management, Processing and Cloud, and Fintech. These sectors have attracted huge amounts of private investments in 2022 and 2023 alone.

Most of these solutions under development will likely not run in the public domain. Companies owning private information may want to develop services to unlock these in a more intelligent (AI-like) way, through proprietary models that could be based on (or combined with) foundational models. Companies, often, cannot share (customer) data with a public cloud model because they lose control over the information. The development and implementation of private (in company) generative AI tools could, therefore, be a solution for privacy and confidentiality challenges. A company like Together.AI specialises in the development of new applications. They can help train AI models for the purpose of clients but also help with the tailoring of datasets and training of the private AI models. A platform like Hugging Face provides tools for building machine learning applications. These tools will become available on the Amazon (AWS) platform.

Leading ecosystems are emerging

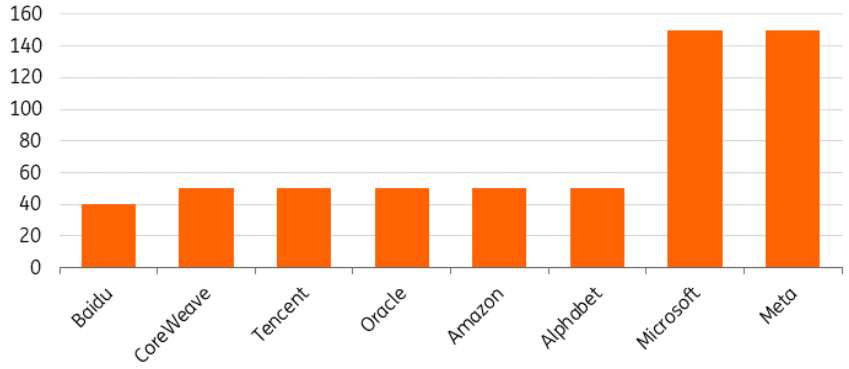
Alphabet, Microsoft and AWS are the leading global providers of data centre services and are among the largest owners of data centre and cloud infrastructure. They are developing the tools and models to provide a full suite of services for their clients that use their infrastructure. Likely, AI will become an important driver of growth in the cloud software market. Through a leading position in AI, a company such as Microsoft or Alphabet, will have the opportunity to solidify (or grow) its cloud service model. IDC expects the spending by businesses on Generative AI to reach $143 billion in 2027.

As described in the previous section, the development, maintenance and roll-out of AI tools require specific knowledge, tools and substantial amounts of capital. As described in the Stanford

University Artificial Intelligence Index Report 2024, the training of Generative AI models is expensive, likely costing millions. But, also, the inference, data, and maintenance of models is costly, especially since the most advanced semiconductors from Nvidia have been in scarce supply.

Nvidia's products were in scarce supply in 2023

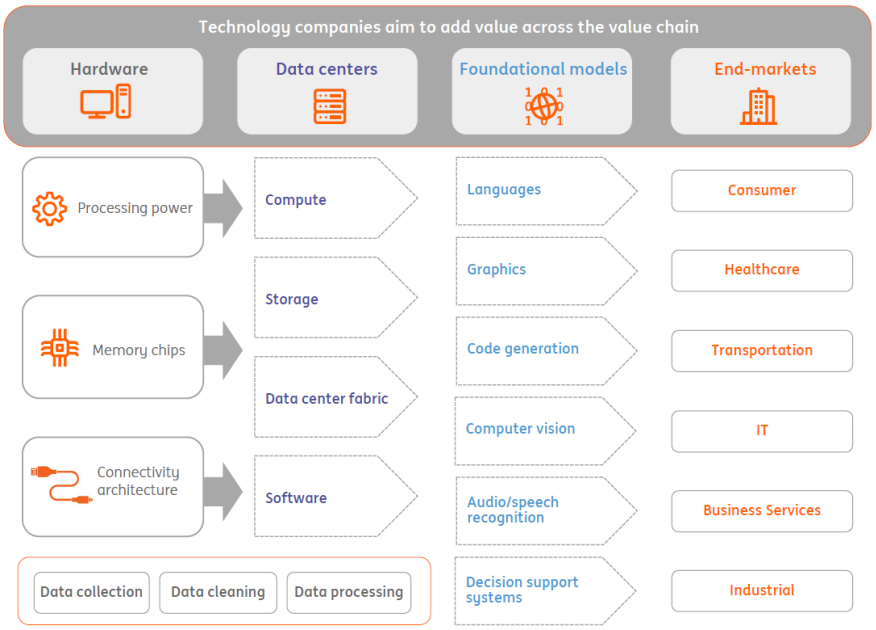
H100 GPU shipments by customer



Source: Omdia Research, through theverge.com

The figure above shows that Microsoft and Meta managed to build the largest portfolios of Nvidia H100 systems, while Amazon and Alphabet are building leading compute capacities. As we have seen, these companies are also the leading data centre operators, with extensive software (development) suites, and a global footprint, which makes them well-connected. Through bundling solutions, Microsoft, Alphabet and (over time) Amazon are well-positioned to provide a strong AI ecosystem, building on an existing strong cloud offering. The challenge to operate across the value chain can be seen in the diagram below.

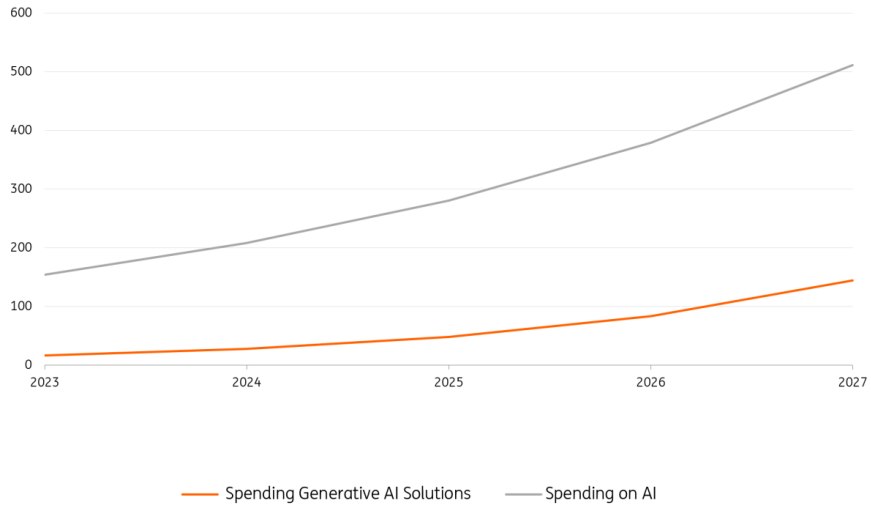
Technology companies aim to sell a full product suite



Source: ING

This will likely have implications for customers buying AI services. Notably, it is not simple for companies to migrate to another ecosystem. And it will be very difficult for new entrants to carve out market share at a later stage because the market leaders benefit from an early mover advantage.

The market for AI services will expand strongly



Source International Data Corporation, ING

Leading cloud operators and Nvidia will dominate the AI landscape for some time to come

We therefore expect that a few AI ecosystems will dominate the landscape for some time,

although there is possibly room for vendor-neutral hyperscale data centres, such as the Equinix xScale data centres to join the race now. Equinix, for example, will offer a private cloud service to enterprises which unlocks a private Nvidia DGX AI infrastructure which they can use to build and run custom Generative AI models.

Author

Jan Frederik Slijkerman
Senior Sector Strategist, TMT
jan.frederik.slijkerman@ing.com