

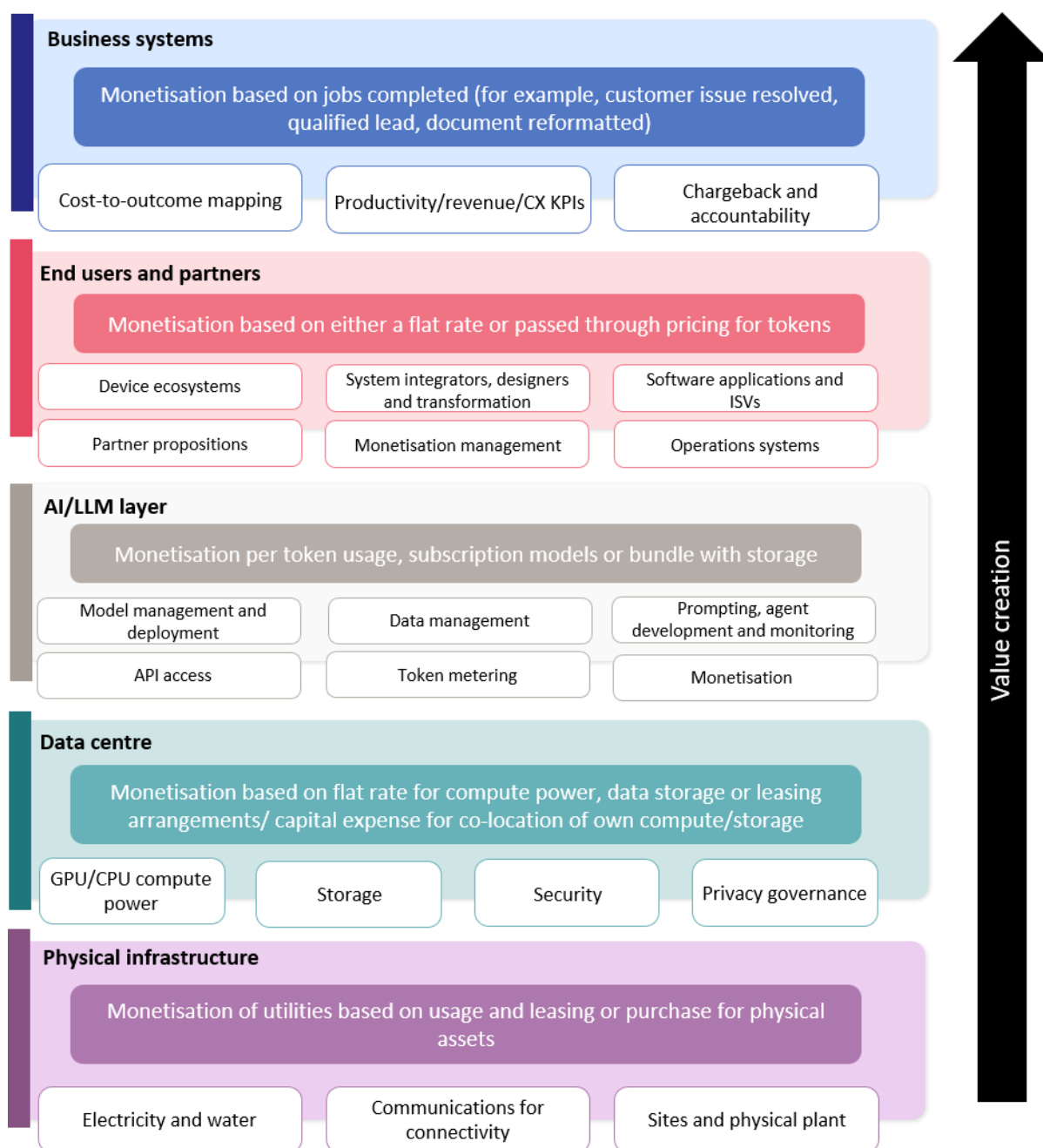
Monetisation of the AI value chain must reflect both the benefits of AI and the underlying cost structures

September 2026

Justin van der Lande

The AI value chain, from chips and data centre infrastructure to models, applications and business processes, creates different monetisation opportunities. The strongest near-term advantage sits with providers of scarce infrastructure, particularly GPU suppliers. However, long-term success will depend on whether value can be captured across the ecosystem and whether enterprises can link AI spend to measurable outcomes.

In the gold rushes of the early 1800s, the largest financial rewards often went to those selling the ‘picks and shovels’ rather than to the miners themselves. In AI, the equivalent beneficiary is currently the chip manufacturer NVIDIA, which reported revenue of more than USD120 billion last year. Other ecosystem vendors that buy or use NVIDIA GPUs – including developers of large language models (LLMs) and AI tools, such as OpenAI and Anthropic – have yet to demonstrate sustained annual profitability.

Figure 1: AI ecosystem¹

Source: Analysys Mason

All layers in the AI value chain must add value and be able to monetise their part of the stack (as expanded on in the following paragraphs), but ultimately it all needs to work at the top of the ecosystem. Here, the enterprise business systems and processes or consumers must ultimately generate value from AI. These benefits are needed to justify the costs incurred across the underlying layers of the AI value chain. Each layer has different business and monetisation models associated with it.

¹ CX' in Figure 1 stands for customer experience.

Data centres are monetised based on time and capacity

Data centres must monetise GPU-based services to cover the costs of site acquisition, utilities, physical security and connectivity. These facilities include those operated by hyperscalers such as AWS, Microsoft Azure, Google Cloud Platform and Oracle Cloud Infrastructure, and specialist neocloud providers such as CoreWeave and Lambda Labs. Telecoms operators are also active in the market and include Deutsche Telekom, SK Telecom, Singtel, TELUS, Ooredoo, Claro and Iliad.

Data centre pricing models are typically structured around time-based consumption, reserved capacity or token-based models, often referred to as GPU-as-a-Service (GPUaaS). Co-location leasing is also being used by some enterprises that operate their own GPUs but require co-location facilities.

Foundational and LLM providers have a consumption-based model or subscription

Beyond flat-rate consumer and enterprise subscriptions, revenue for LLM and foundational model providers are largely derived from pay-as-you-go access for developers using APIs and from cloud partners that provide access to models. Usage is generally measured in tokens and billed per million tokens. Pricing varies by provider and model, with input-token prices typically lower than output-token prices.

Pricing also varies according to the model being accessed: higher-value flagship and reasoning models command premium rates. Volume discounts are applied to larger users, and major ecosystem vendors often negotiate specific commercial agreements. This business model should support the path to profitability for vendors as token usage rises into the billions and trillions. Azure has reported processing 100 trillion tokens in a single quarter, with 30% growth quarter on

quarter. In addition, Google Cloud has reportedly processed 16 billion tokens per minute through APIs.

Both enterprises and consumers purchase AI directly and through a wider ecosystem of suppliers

LLMs and AI ecosystems are consumed either directly by users, including consumers and enterprises, or indirectly through a broad ecosystem of partners. These partners include:

- ecosystem vendors, such as Oracle and Microsoft, that enable developers and consumers to access multiple LLMs through the platform services that they offer
- independent software vendors (ISVs) that embed AI agents, models and capabilities within their applications
- systems integrators and consultancies that create AI-enabled solutions for enterprise customers
- device ecosystem vendors that enable their customers to access AI-based software, frameworks and other capabilities.

Pricing for AI usage is bundled as part of the service, or is a pre-requisite for the use of the software and the purchasers have their own commercial agreements with the providers.

Most major AI providers operate tiered partner programmes that offer earlier access to new technologies and recognise advanced users of their platforms.

Monetisation must ultimately be based on value to the purchaser

For every enterprise that uses AI, a process or system must ultimately change to create value and justify the investment. The novelty of new technology quickly fades, and organisations increasingly require clear commercial justification for continued AI spend. For enterprises experiencing rising AI costs, allocating those costs to specific business outcomes is becoming critical. Attributing costs to a particular process, use case or business issue helps to support investment decisions and future funding. Where consumption-based monetisation is deployed, usage may need to be monitored at the level of individual business processes or systems.

Outcome-based pricing, built around time/costs saved or measurable customer revenue, is often delivered in the form of “revenue share” models or KPIs as part of a contract. Historically, the effort involved in measuring outcomes and the definition of results has made these models harder to implement. Each customer engagement defines its KPIs differently, making engagements more complex. Accessing systems and outcomes is often possible through business applications, but other engagements require each party in the agreement to have access to suitable data, which is not always desirable.

Partners that implement AI or embed it within their own solutions also need to understand how their costs vary with usage. They may need to manage the commercial risk of supplying flat-rate solutions when the underlying AI services are priced on a consumption basis.

Consumption-based pricing is the way forward

Few suppliers can provide comprehensive products across the full AI value chain; most offer discrete solutions within individual layers. Vertically integrated AI solutions are only possible for the largest organisations, such as Google Cloud, while most vendors will continue to compete within specific horizontal layers.

Vertically integrated AI stacks offer some advantages, but most opportunities will remain within horizontally integrated solutions. Consolidation within each horizontal layer is likely to continue as organisations seek greater simplicity across the AI ecosystem. Major vendors will dominate through acquisitions and larger internal development resources. Consumption-based monetisation will increase as pricing becomes more closely aligned with underlying costs.

Explore Analysys Mason's [AI and Data Platforms](#) and [Monetisation Platforms](#) research programmes for expert analysis of how AI, automation and next-generation monetisation strategies are reshaping the telecoms industry. To discuss your organisation's priorities and learn how our research can support your strategy, contact [Justin van der Lande](#).