AI bill shock: could rising costs drive enterprises to run their own AI infrastructure?

July 2026

James Allen

The ICT industry has long debated whether computing resources are more efficiently delivered centrally (for example, a mainframe with dumb terminal) or through local, on-premises infrastructure (for example, workstations).

The mainframe era used a centralised system. The subsequent minicomputer era introduced a more hybrid approach, as powerful workstations such as Sun and Silicon Graphics became increasingly common for specific tasks. The PC/LAN era was a largely local era, while the internet and the rise of cloud computing moved back to a hybrid/more centralised model. Today, many users still have a laptop or desktop that can run office automation/graphics apps, while also using web services.

The boom in LLM-based agents, coupled with the recent shift by loss-making AI model providers such as Anthropic from per-seat subscriptions to API-based pricing, has driven a rapid increase in token costs for large enterprises, leading to AI bill shock. Many have run into this exact problem today, due to Microsoft changing the pricing for Cowork. In addition, these costs are difficult to forecast, especially as the relationship between tokens and correct output ('useful work') is, at best, indirect. At worst, some fraction of the tokens are simply wasted. In addition, none of these systems applies real-time hard constraints, unlike a mobile phone prepay billing system where access is suspended the instant a user's balance reaches zero. Agents that can spawn other agents and execute code using paid APIs can sometimes result in much higher spend. In combination these two factors create a significant risk of bill shock.

From token economics to hardware ownership

One way in which to manage your LLM costs is to use lower-cost, simpler models, since if less compute resource is needed, there is less cost. Some existing models automate this process where they can. Another related option is to use open models (which tend to be simpler and smaller) and move back to ownership of the compute hardware rather than buying 'tokens as a service'. This ownership model can be fully local. Relevant staff can use high-end workstations that can run the relevant model inference locally, either on a central processing unit (CPU) or a graphics processing unit (GPU), although performance may be relatively slow. Alternatively, the model could be centralised through investment in AI inference servers that are shared across a company or department, so this may more closely resemble a private cloud model. At the upper end of the scale, such owned server facilities would also necessarily be housed in data centres. If served at the most local level, for example on desks and devices, this token demand would not be

using data centre capacity at all. This could have implications for a fraction of the large number of data centre projects currently based on expectations of very large growth in token demand.

In the case of ownership, the uncertain opex is converted into more certain capex and associated opex (power bill, cooling, maintenance, and so forth). Costs are capped either way. Providing each end user with a specific workstation can give them more control over managing their spend and help organisations to extract value for money from the combination of employee and ICT resources.

Of course, there are downsides:

- the benefit (the effort saved by, or the additional useful output provided by the use of the agent or LLM) is also capped
- there is upfront capex on the required hardware (which will likely depreciate quickly)
- there is less flexibility compared with the current cloud API model.
- considerable effort (and a great deal of money) have gone into engineering the Frontier model inference servers to deliver low latency (time to first token) and high-throughput (tokens per second). If output is generated more slowly, end users interact with systems differently and get very different benefits. For example, interlacing two tasks so as to be able to use slow computer output for each is cognitively costly for the employee: they get less done. Those looking to take ownership should be wary of being 'penny wise, pound foolish': the devices you can afford may not give you the performance you need.
- the maintenance and eventual upgrade needs of the on-desk or centralised infrastructure becomes the organisation's own problem.

Current tokens-as-a-service providers are loss-making and benefit from large economies of scale (implying that, all things equal, it would be very challenging to replicate their outputs more cheaply yourself). However, the 'own hardware' trend is already present to some degree. End users have started to use Mac mini Pro desktops with more than 32GB RAM or, at the higher end, NVIDIA DGX Spark workstations (at a cost of approximately GBP4500 each, though two might be needed). Where the hobbyists go, the engineering department may follow, and at the right price point that may extend to other departments that are heavy users of LLMs.

What could accelerate or slow the ownership trend?

Factors that might deter a move towards 'own hardware' include the following:

- a rapid evolution in practically useful Frontier model capabilities, if these cause rapid changes in hardware requirements for competitive performance open LLMs (especially with regards to high-speed memory)

- end-user/corporate application lock-in to specific specialised LLM harnesses such as Claude Code that work only with specific LLMs (or, even if notionally technically agnostic, work better or more efficiently with specific LLMs, as we require 'useful work')
- high prices due to hardware shortages (for example, shortage of RAM or GPUs).

Tokens-as-a-service providers like Anthropic could limit the attractiveness of such options by providing:

- better ways to influence or limit agent token consumption; prompt-compression tools could reduce token requirements and lower costs
- better ways to predict how many tokens will be required for a specific task (enabling prioritisation and budgeting in the harnesses)
- alternative tariffs/commercial models with ownership-like cost structures (including dedicated prepaid capacity, rate limiting, with hard cut-offs).

Conversely, it may be that some of the commercial coding platforms become increasingly model agnostic, adopting a 'bring your own LLM' approach to appeal to organisations that own their own AI hardware. This might also decrease their reliance on a token-based model, given that tokens are currently loss-making for most players in the AI value chain, apart from NVIDIA and utilities. There are already several open source/model agnostic coding platforms, including Goose. We note, however, that SpaceX/xAI is moving in the opposite direction by buying Cursor. SpaceX has a significant valuation that is predicated on AI, and Cursor offers integration possibilities that could increase usage of Grok.

Organisations are increasingly rethinking how they deploy AI to balance performance, scalability and cost efficiency. To discuss strategies for reducing AI spend, avoiding bill shock and evaluating the evolving economics of AI infrastructure, contact James Allen, Partner, Oli Barnett, Managing Partner or Christian Fischer, Partner for expert insight and practical guidance.