POLAR CAPITAL

The next wave of AI innovation

Fred Holt & Nick Dumas-Williams, Investment Analysts
Polar Capital Global Technology Team

Fred Holt, Investment Analyst: Until recently, I would guess that not many would recognise the DeepSeek logo. Even today perhaps it is not that well known, but on 27 January the Chinese company released their R-1 reasoning model into the world, sending shockwaves through the market

In one day NVIDIA, the poster child for AI, lost c$600bn of market cap, roughly equivalent to the market cap of Boeing, Ford, PepsiCo, PayPal and Starbucks combined.

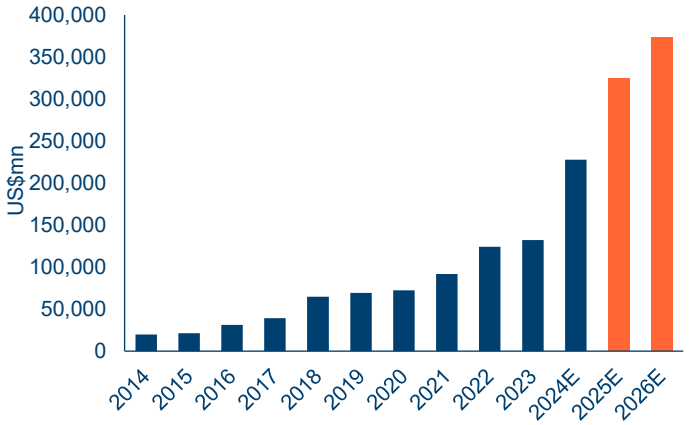
So, what did DeepSeek do? DeepSeek's model boasted performance similar to OpenAI GPT-4o at reportedly significantly lower costs, sparking concerns around AI capex requirements. They claimed to have done this for about $6m of spend, although this has been widely debunked with spend likely closer to $1-2bn.

So, why has NVIDIA recovered almost all of that $600bn market-cap loss? This is because this was a case of some novel algorithmic improvements, but only really resulting in a model that was equivalent to OpenAI models of some 6-9 months ago and, crucially for us, on a relatively standard technology improvement curve. In technology waves we get rapid discontinuous change, and rapid innovation tends to follow rapid innovation. It is disconcerting because it was from China but don't forget that this model was also enabled by large frontier model development from the largest hyperscalers.

So what are the largest hyperscalers saying and doing?

Over the past few weeks, and since the announcement of DeepSeek, the world's largest spenders on compute, data centres and AI have been raising their capex forecasts for 2025 (Fig.1). Consensus expectations are now for around 41% y/y capex growth in 2025 to around $325bn, and we expect another growth year in 2026. In addition, NVIDIA is ramping its new Blackwell chip as we speak, and we have had numerous companies come out to lend their support around the value they see in AI.

Fig.1: Top 10 cloud providers: Cloud capex y/y (excl. Amazon)[1]



Source: Polar Capital, November 2024. 1. Morgan Stanley, 7 February 2025.

OpenAI and other technology partners including Softbank, Oracle and ARM have announced the Stargate project, with a proposed $500bn investment in AI data centres over the next four years, with building already started in Aberlene, Texas. Amazon made clear that AI is a once in a lifetime opportunity and the biggest technology shift since the internet.

This rate of change can be unsettling, but it is not unusual to us or to technology waves in general.

Humans by their nature struggle to think about exponential levels of improvement and typically tend to think linearly. This means that we often encounter innovations like DeepSeek where there is a tendency to believe that the pace of innovation has peaked and that we have perhaps made all the breakthroughs we will ever make. When, in reality, the step change just witnessed is merely an innovation on a typical path of technology improvement.

Polar Capital LLP 16 Palace Street, London, SW1E 5JD
T +44 (0)20 7227 2700 E investor-relations@polarcapitalfunds.com polarcapitalfunds.co.uk
For further information please contact your Polar Capital representative

For example, in 1971 Intel released the 4004 microchip with c2,300 transistors. It was widely believed at the time that this would likely be more compute than we would ever need for almost all applications. As we speak, NVIDIA is ramping its Blackwell GPU with 208 billion transistors and the pace of innovation continues apace.

Furthermore, in the 1950s breakthroughs in nuclear power technology led Lewi Strauss, the Chair of the US Atomic Energy Commission, to say; 'It is not too much to expect that children will enjoy electricity too cheap to meter" – i.e. we had solved the challenge of getting ubiquitous and cheap energy.

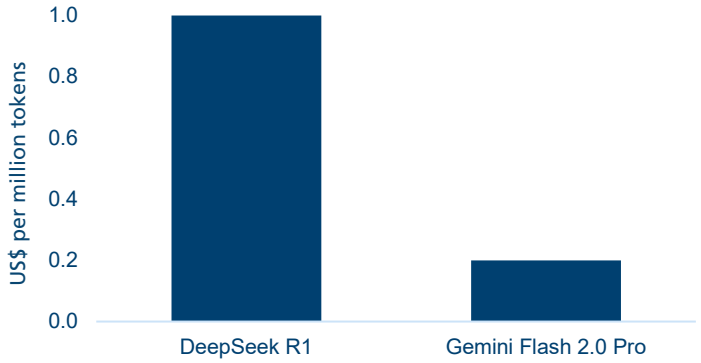
Finally in 1997, perhaps the most relevant example today, is when IBM's Deep Blue beat Chess Grandmaster Garry Kasparov at chess. This was hailed, rightly, as a major tech breakthrough and the moment at which technology had become smarter than humanity. Indeed, the cover of *Newsweek* at the time hailed this as 'The Brain's Last Stand'. Of course, this was just the beginning of AI and the path to AGI i.e. AI that is actually smarter than a human.

In short, we are seeing technology innovation happen rapidly, entirely consistent with other technology waves, and we expect to see more over the coming months and weeks.

Nick Dumas-Williams, Investment Analyst: Was the DeepSeek announcement, which at a headline level introduced very impressive reduction in costs in running AI (inferencing), one of these moments? Have we solved low-cost inferencing?

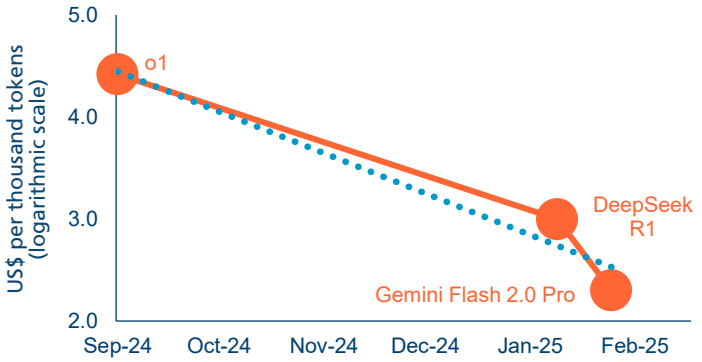
What if I told you that less than one month after the DeepSeek announcement, Google reduced the equivalent cost a further 5x (Fig.2&3). It can be debated whether the market would have had such a strong reaction had the Google announcement come first, but regardless, that adds up to a significant 130x cost reduction in a very short space of time.

Fig.2: Cost of inference



Source: artificialanalysis.ai, 17 February 2024.

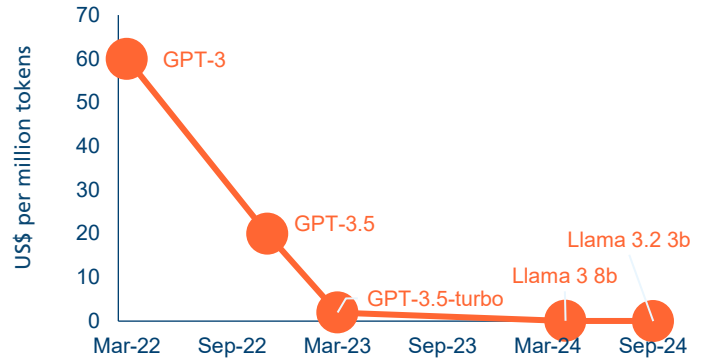
Fig.3: Cost of inference



Source: Polar Capital, 19 February 2025.

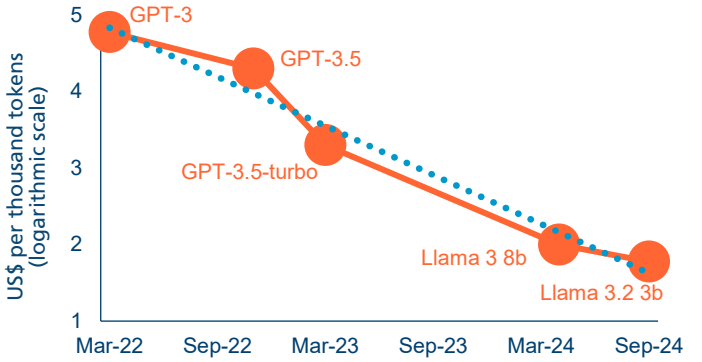
But that trend is exactly what is supposed to happen. We have seen this before. Fig.4&5 show the cost of GPT-3 intelligence, which has become 1,000x cheaper within three years. This is a key characteristic, and indeed a requirement, of tech progression. It can be hard to fathom but the technology is doing exactly what it should.

Fig.4: Cost of inference



Source: artificialanalysis.ai, 17 February 2024.

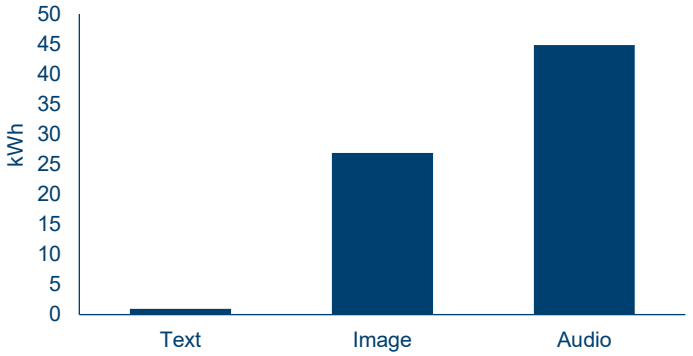
Fig.5: Cost of GPT-3 level intelligence



Source: Polar Capital, 19 February 2025.

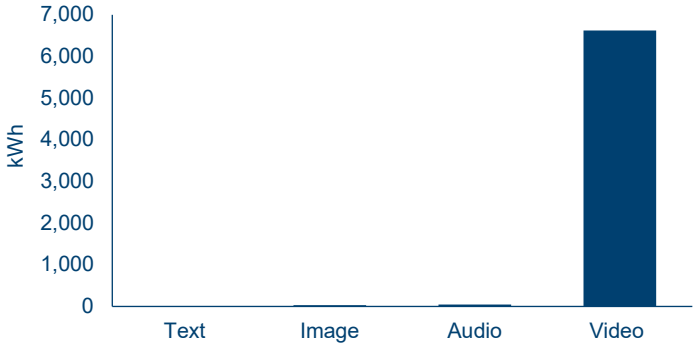
Why does this matter? Because it allows us to do more complex work at greater scale. Fig.6 is a proxy for the intensity inferencing in different types of media and unsurprisingly more complex media requires more power and more cost. That explodes when you get to video (Fig.7). The collapsing cost of inferencing is absolutely required to unlock this higher value work.

Fig.6: Normalised estimated power consumption per inference task



Source: Wells Fargo, 17 February 2025.

Fig.7: Normalised estimated power consumption per inference task



Sorce: Wells Fargo, 17 February 2025.

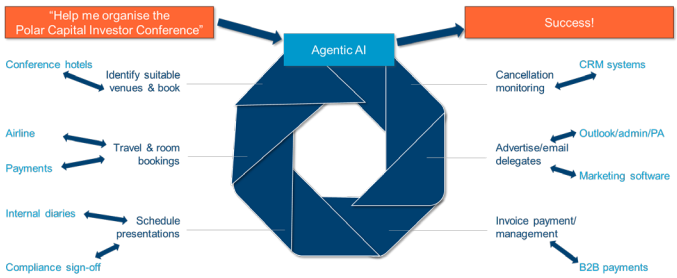
The statistics also prove that we are seeing the volume response so, as the costs come down, the volume increase more than offsets that to deliver growth – another key function of tech progression.

Finally, this is perhaps a forward-looking view on where that volume response could really take off.

You may have heard term 'agentic AI', this is AI that can act on your behalf (i.e. has agency). Fig.8 is a thought experiment showing how Agentic AI could support both the excellent team we have here at Polar Capital and also the wider ecosystem, who put on our recent annual Investor Conference.

In each of these steps the agent can interact with other systems and other people, it can make payments, chase invoices, check against databases and systems. The next ring is where hotels are going out to suppliers, and from there the web keeps expanding.

Fig.8: 2025 - the year of agentic AI?



Source: Polar Capital, February 2025.

Each of these interactions is many, many inferencing steps and so the continued cost decline is the only way to facilitate this kind of work, which we are extremely excited about. It increasingly looks like we will see some form of agentic AI this year and so we would expect continued strong volume response.

Fred Holt and Nick Dumas-Williams

Polar Capital Global Technology Team

25 February 2025