C WORLDWIDE
ASSET MANAGEMENT

DeepSeek – A Sputnik Moment

Equity markets globally reacted abruptly on Monday, 27 January, due to the introduction of the DeepSeek language model.

The model's reception has been impressive. Over the weekend, it became the number one most-downloaded app in Apple's US App Store, sending shock waves through Silicon Valley and causing the share price of major tech stocks to plummet. DeepSeek R1's benchmark surpassed OpenAI's o1 model on multiple fronts. According to DeepSeek, they only spent $5.6mn on training the model, which is less than 10% of OpenAI's GPT4o.

The true disruption of this news event is the cost reduction, on the back of market concerns that the hardware Total Addressable Market for model training might shrink to a fraction of prior expectations, and investment in AI might become biased towards software, thereby opening the playing field.

Never before have we seen a technology evolve as quickly as GenAI. The latest iteration of advancement from DeepSeek comes on top of another major innovation from September 2024, namely Chat GPT o1, where compute moved from training (the model) to inference (use of the model). This was a major event that, over the longer term,

"Open-source models can potentially democratise AI, threatening Big Tech's ability to dominate it. We view the open-source methodology as positive for a broader tech ecosystem, which ultimately may be positive for all AI users and AI end demand."

Morten Springborg
Global Thematic Specialist
C WorldWide Asset Management

in our view, will likely impact and shift market shares amongst chip manufacturers (in a relative sense away from GPUs).

DeepSeek now presents smaller, more nimble models than o1 built on open source and a mixture of expert models (MoE) that prove to be much cheaper to build and use for inference, which will most likely further influence the chips market.

Open-source models can potentially democratise AI, threatening Big Tech's ability to dominate it. We view the open-source methodology as positive for a broader tech ecosystem, which ultimately may be positive for all AI users and AI end demand.

DeepSeek adopted an MoE architecture for their model. LLM and MoE are two different approaches to AI development. LLM is designed for general-purpose training. MoE is a sparse architecture where the model is divided into different experts (i.e. submodels). The MoE architecture brings a significant cost advantage to model inference and requires much less hardware.

While most cloud service providers are scaling up their computing clusters, DeepSeek is training their

model under a constrained budget and using hardware resources to optimise performance. We believe this trend, because of the downsizing of the models, will benefit edge AI devices.

As demand for inference is expected to increase massively, costs become critical. In other words, without serious cost decreases, inference will not ramp up as expected. According to news reports, DeepSeek is 10-20 times less expensive for inference than existing models from, for example, OpenAI. For the maturation of AI, this is a great development.

Necessity is the mother of invention

This proves that LLMs are commoditising, and cost and scale will be vital for commercial success. Several other Chinese model developers are about to introduce simpler and more optimised models. As they say, necessity is the mother of invention. The US ban on exports of advanced semiconductor technology to China has forced China to innovate in model design and train on older versions of Nvidia GPU. It will be interesting to see whether the US model developers will change tactics and shift away from brute force models and begin to focus on model optimisation. We think they will; in that sense, this is a small Sputnik moment.

The market price movements are telling us that it's negative for the demand for leading-edge GPU/ASICs – at least in the short term. In the longer term, it's more uncertain because of what is known as the Jevons Paradox. Jevons paradox occurs when technological progress increases the efficiency with which a resource is used (reducing the amount necessary for any one use). Still, the falling cost of use induces increases in demand enough that resource use is increased rather than reduced.

This has always been the case in compute and software. DeepSeek is just another downward shift in costs. And this will amplify demand for compute longer-term.

Also, these smaller models open the market for edge AI, where the models will reside on devices. This would not be possible with the large brute-force models.

While still early days, preliminary investment conclusions are:

1. Aggressive growth expectations for leading-edge compute for training models will be questioned

2. Longer-term demand for compute will increase because lower costs drive demand (Jevons Paradox)

3. Lower costs accelerate the roadmap for edge AI. Advanced scaled-down models will run locally on devices.

4. Demand for compute and silicon will increase in totality.

China is an innovative and deflationary force

Finally, considering DeepSeek's sudden and unexpected geopolitical and economic impact, there are important parallels to other advances in Chinese industrial might,

"Have the US sanctions in the name of national security backfired? Have the restrictions unintentionally made Chinese engineers more creative and resourceful?"

like EVs, renewables, robotics, etc. China is an innovative and deflationary force, which is a significant challenge for advanced technology and manufacturing in the West.

The US chose AI as the arena for competition and confrontation. Have the US sanctions in the name of national security backfired? Have the restrictions unintentionally made Chinese engineers more creative and resourceful? Could DeepSeek have come this far if the Chinese race for AI supremacy had happened on US terms of brute force? What could be the next step for the US to assert its leadership in AI supremacy? More sanctions? Discussions about banning or restricting DeepSeek in Western markets have not yet been explicitly reported. Still, its rapid rise could potentially lead to regulatory actions or calls for restrictions in the future. However, no matter what happens explicitly to Chinese models, the genie is out of the bottle in the sense that model development and resource use will change dramatically also for Western-developed models.

Therefore, this is a positive supply shock to the economy and, as such, positive for productivity and growth. It is good for every consumer and company that utilises the technology. The value will migrate to the technology users.

Also, the fact that a relatively small outfit in China has managed to force itself into the top of the league table of GenAI should question the "exceptionalism" of certain US tech companies. The premise that massive value will be extracted from excessive investments in model training and compute clusters should also be questioned. In that way, DeepSeek can also be seen as a term of Trade shock, which could have macroeconomic implications in the form of capital flows into the US equity market and then into dollar assets, as described in our recent White Paper From Diversification to Distortion: The Impact of Passive Investment Flows - C WorldWide Asset Management. The US technology sector is not the only game in town.

"The fact that a relatively small outfit in China has managed to force itself into the top of the league table of GenAI should question the "exceptionalism" of certain US tech companies"

Disclaimers