China’s DeepSeek model is a major advance in AI technology

Bill Shaw
30 January 2025

Last week, DeepSeek, a startup company based in Hongzhou, China, released its newest artificial intelligence model, DeepSeek R1. Within days, the chatbot became the most-downloaded app in Apple’s App Store.

DeepSeek’s performance meets or exceeds that of state-of-the-art AI models from American companies such as Meta and Open AI, surpassing all open-source models previously available and many closed models on most standard benchmarks.

The achievement sent shockwaves through Wall Street, wiping out approximately $1 trillion in market value for corporations in one day. It also represents a major blow to US plans for sustaining AI dominance as part of its objective to prevent China from usurping the US as the top economic and military power in the world.

In addition, DeepSeek’s phone app connected to R1 quickly topped the charts on the Apple Store, surpassing the ChatGPT app. On the Google Play Store, it has already been downloaded 10 million times.

Model performance gains

Academia and industry measure how “good” an AI model is using standard benchmarks. These benchmarks are predefined tasks for which the answers are known. The model is applied to the tasks, and its outputs are compared to the known answers. Generally speaking, the greater the number of correct answers on the tasks, the better the model performs. A shared set of standard benchmarks enables comparing models against one another.

The DeepSeek team tested its R1 model on 21 benchmarks and compared the results to those achieved by industry-leading AI models from Meta, Open AI and others. The benchmarks included English-language, Chinese-language, software-programming and mathematics tasks.

They compared R1 to four industry-leading AI models as well as their previous version of DeepSeek. These models included Claude-3.5-Sonnet-1022 from Anthropic; three Open AI models—GPT-4o, o1-mini, and o1-1217; and R1’s predecessor DeepSeek-V3.

DeepSeek R1 outperformed the other models on 12 of the 21 benchmarks. For the remaining nine benchmarks, it placed second on eight and fourth on one.

It should be noted that o1-1217, given its purpose and design, was applicable to only 11 of the benchmarks. For those 11 benchmarks, R1 was the best model for four tasks, whereas o1-1217 was the best model for six tasks and Claude was the best model for one task. R1 bested o1-mini on 20 of 21 benchmarks.

Dramatic reduction in computation

What makes the DeepSeek achievement particularly dramatic is the massive reduction in the computational resources needed to build R1. DeepSeek used far fewer computational resources than required for the creation of its competitors.

Building R1 required approximately 2.8 million compute hours on a graphics card from NVIDIA called an H800. Such graphical processing unit or GPU cards are used to build AI models because they efficiently execute the complex mathematical computations required. DeepSeek used a computing infrastructure with 2,048 H800 cards.

By contrast, Meta required 30.8 million GPU hours to build its popular Llama-3.1 model, meaning the DeepSeek R1 model took only 9% as long. Because DeepSeek R1 is a larger model than Llama-3.1, the speedup is even greater than a 91% reduction.

Model size is typically given as the number of numerical parameters that comprise the model. DeepSeek R1 is 671 billion parameters compared to Llama-3.1’s 405 billion, or 66% larger.

The speedup in model construction is made even more impressive by the fact that the H800 GPU is a stripped-down version of NVIDIA’s H100 GPU to comply with United States export control restrictions to China. Meta’s estimate of 30.8 million GPU hours to construct Llama-3.1 405B is based on the faster H100 GPU card. Tests of the performance difference between the cards show that the H800 is approximately 11.5% slower than the H100.

Open source

The fact that DeepSeek R1 is open source means that the full set of 671 billion parameters and the software used to operate the model are freely available to download, inspect and modify. Open-source models are often preferred by software developers and AI engineers because they are easier to modify and adapt to various purposes.

Despite its name, Open AI's leading models are not open source. AI engineers cannot inspect or modify Open AI's leading o1 model, for example, or its immediate predecessor GPT-4o.

Additionally, R1 implements a "chain of thought" procedure, a technique originally developed by Open AI for its o1 model. Whereas o1 and other Open AI models hide the "reasoning" steps in the chain of thought, R1 lets the user see all the steps it takes to reach an answer.

Because open-source models can be used and modified by anyone, an industry of companies that host models has risen. For example, Meta's open-source Llama-3.1 model is hosted by several different companies that compete on the cost of using the model.

Observers quickly noted that queries to the DeepSeek-hosted version of R1 refused to answer queries such as "what happened at Tiananmen Square?". The open nature of the model does not imply that China is becoming less authoritarian. However, it does enable anyone outside China to host the model themselves without such restrictions and censorship.

Furthermore, the criticism also applies to Open AI models, which refuse to answer questions about the Gaza genocide when prompted. Censorship of closed models is much more difficult to overcome than with open-source models.

Low cost to use

DeepSeek also charges far less for the usage of R1 than its competitors. The largest models are too computationally expensive to run on personal computers or even most servers. The same large GPU infrastructure that is used to build the models is also typically used to run these models.

The result is that AI companies stage the models on their large GPU clusters and accept requests—known as prompts—over the Internet, input the prompts into the model and then return the model's output back to the user.

Running R1 via such application programming interface or API calls over the internet is far cheaper than for other leading AI models. DeepSeek is currently charging for R1 less than 4% of what Open AI charges to run its o1-1217 model. Specifically, o1 costs are $15 per million tokens (MT) input and $60 per MT output, whereas R1 costs $0.55 per MT input and $2.19 per MT output, a reduction of 27 times. A token is approximately equivalent to a word.

To achieve the lower costs to operate R1, DeepSeek uses an architecture called "Mixture of Experts." This means that for each token generated, only a fraction of the model (37B parameters of the 671B, i.e. an "expert") is activated. This reduces the computing power required for model output, resulting in the lower costs.

In addition, modifications to models through a process known as quantization can dramatically reduce the computational resources necessary to run a model. Although quantization does reduce model performance, various quantization schemes can dramatically lower computational requirements while only decreasing model performance somewhat.

Already two researchers, taking advantage of the open-source nature of R1, created multiple quantized versions of it. One version can run on a desktop or laptop computer with as little as 20GB of RAM, albeit slowly. These researchers published their modified versions of R1 as open source on an AI model repository known as Hugging Face.

Implications for US dominance in AI

The week prior to the DeepSeek news, President Trump announced a planned $500 billion initiative called StarGate to invest in technology to ensure United States dominance in AI. Stargate LLC, a company with investments from Open AI, Oracle, SoftBank and the investment firm MGX, seeks to build multiple AI data centers across the country, beginning with 10 centers in Texas. Trump also announced he would eliminate regulations on generating the massive quantities of electricity required to run the data centers.

In addition, Open AI announced on January 21 the pending release of its next AI model, o3-mini, in "a couple of weeks."

The DeepSeek achievement immediately eclipsed the StarGate initiative and Open AI's plans for o3-mini, turning the AI industry in general on its head. The perception that the US has a long lead in AI—whether previously justified or not—has vanished practically overnight, raising questions about the ability of the US to create or maintain dominance in AI. DeepSeek and its R1 model have become the central topic of conversation, shifting the work focus of vast swathes of the AI industry.

The Biden administration had not only put in place the export controls that resulted in the DeepSeek team using H800 instead of H100 GPUs, but it also expanded those restrictions in its final days in office. President Trump was already expected to ramp up economic and military confrontation with China further, but the DeepSeek achievement is likely to accelerate and intensify the planned escalation.