Engineering 53 (2025) 3–5

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

News & Highlights

Chinese AI Model Shocks the World—What Comes Next?

Mitch Leslie

Senior Technology Writer

Monday, 27 January 2025, was a dismal day for technology investors. The share price of Nvidia (Santa Clara, CA, USA), the world's largest manufacturer of the graphics processing units (GPUs) that underlie artificial intelligence (AI), plummeted 17% [1]. The fall sliced 589 billion USD from the company's value, the biggest single day decrease in history [1]. The stocks of other AI heavyweights such as Meta (Menlo Park, CA, USA) and Alphabet (Mountain View, CA, USA), Google's parent company, also sank [2]. Power companies suffered as well. Shares of Constellation Energy (Baltimore, MD, USA), the country's largest owner of nuclear plants, fell by 20% [3].

The trigger for this sell-off was the announcement by DeepSeek, a start-up in Hangzhou, China, that it had released an AI reasoning model that matched or beat the performance of rival models but cost much less to build, train, and operate [4]. Big technology players such as OpenAI, the organization based in San Francisco, CA, USA, that produced ChatGPT, had already unveiled reasoning models [5,6]. "The idea was that they were impossible without significant amounts of resources," said Subbarao Kambhampati, professor of computing and augmented intelligence at Arizona State University (Tempe, AZ, USA). For a small, almost unknown company to release a model of this type was "a shot across the bow," he added. "They said, 'hey, we can catch up.'"

That prospect scared investors in big technology and power companies, who feared that DeepSeek's feat could start a trend toward smaller, cheaper AI models that gobbled less electricity [7]. Much of the initial reaction to DeepSeek's news was due to "obsessive hype," as one technology research company put it [8]. Still, experts say that DeepSeek deserves credit for using creative approaches to develop its reasoning model [9]. It produced the model, known as R1, even though it did not have access to Nvidia's most advanced GPUs—US technology export restrictions prohibit their sale to Chinese companies [10]. "They were very clever about how they used the compute power they had," said Anthony Cohn, professor of automated reasoning at the University of Leeds (UK) and a researcher at the Alan Turing Institute in London. In addition, industry observers say, the company's achievement is a testament to the growing AI capability of China, which is rapidly gaining on the companies from Western countries that have dominated the field [10,11].

The performance of standard large language models (LLMs) such as OpenAI's GPT-4 may be plateauing, so large reasoning models, or LRMs, could be the next big thing for the AI industry [12,13]. If these models can "think through" questions, they may be better at tasks such as solving complex mathematical problems and writing computer code. Specialized models are necessary because standard LLMs are poor at reasoning, said Kambhampati. OpenAI debuted the first LRM, dubbed o1, in September 2024 and set the standard for the field [5,14]. By the time of R1's release, companies such as Google had jumped in with their own products [6]. Whether these models reason in the same way as humans do is controversial [15], but they outperform ordinary LLMs on a variety of problems that require logic to solve.

Although R1 was not the first LRM, it made a splash because DeepSeek seemed to do more with less. The company did not specify how many GPUs it required to train R1, but it did reveal that only about 2000 Nvidia H800 GPUs were necessary to train V3, the LLM that R1 is based on [10]. In contrast, Meta enlisted 16 000 of the more powerful Nvidia H100 GPUs to train a comparable LLM [10].

DeepSeek also appeared to achieve high performance at a surprisingly low cost. The company said that it spent only about 6 million USD to train V3 (it did not provide an amount for R1) [16]. Big technology companies are tight-lipped about their costs, but estimates suggest that training some LLMs may now require 100 million to 1 billion USD [17]. DeepSeek's 6 million USD figure was likely the amount needed just for V3's final training, said Kambhampati. Still, it suggested that DeepSeek had managed to economize. R1's price tag remains unclear, however, because the company has not disclosed its total development cost [7].

Researchers know a lot about how R1 works because DeepSeek published a preprint that described the model, made the source code available for free, and provided the model weights that shape what R1 learns during training [18–20]. The model shows "a lot of very careful engineering," said Cohn. One example is how it learns [21]. After standard LLMs go through initial training on gigantic amounts of text, they undergo further refinement to improve their answers, including a stage known as reinforcement learning from human feedback (RLHF), during which they tailor their responses based on ratings by human evaluators [21]. DeepSeek's approach also included an initial training period, but the reinforcement learning stage did not rely on human feedback. Instead, R1 attempted to solve problems with known solutions, and algorithms graded its answers [21,22]. During the process, the model generates long sequences of so-called intermediate tokens, or raw output [22]. Some researchers refer to these sequences as reasoning

https://doi.org/10.1016/j.eng.2025.07.009

traces or chains of thought, although Kambhampati and colleagues object to that terminology because it implies that they are comparable to the steps of human reasoning [23]. The model can produce pages and pages of intermediate tokens for each problem it attempts to solve [23]. As Kambhampati and his coauthors put it, they resemble "better formatted and spelled human scratch work" [23]. However, they serve a useful purpose because they are the fodder for further training that allows the model to improve its odds of getting the right answers [22]. One way DeepSeek's approach to training may have permitted the company to cut development costs is by reducing the number of workers needed to refine the model [21].

DeepSeek's achievement could open the way for a range of powerful new models, said Fei Wu, professor of computer science and director of the Artificial Intelligence Research Institute at Zhejiang University in Hangzhou, China. "Building upon DeepSeek's work, we can develop numerous domain-specific small models to accomplish tasks within particular fields," he said. "By then integrating these specialized models, we can create a highly versatile general-purpose model capable of performing tasks across different domains." Moreover, Wu said, several large Chinese technology companies—including Baidu (Beijing), Tencent/WeChat (Shenzhen), and ByteDance (Beijing)—are integrating DeepSeek's models into their operations. "DeepSeek's open-source strategy is accelerating the universalization of AI technology."

DeepSeek jolted stock markets, outraged pundits and politicians who lamented that the United States was falling behind technologically, and sparked plenty of breathless news coverage. Its application (app) quickly became a bestseller in Apple's App Store (Fig. 1). However, whether DeepSeek's approach will revolutionize the AI industry remains uncertain. For instance, it is not a given that other companies will pursue leaner, cheaper models because of DeepSeek. Competitors may stick with the standard paradigm for AI advances, which emphasizes ever-larger and increasingly expensive models [7]. "It is not clear whether small models will dominate," said Kambhampati.

Moreover, DeepSeek's approach may not curb AI's growing appetite for energy, as some experts prophesied [24–27]. Although R1 appears to have been cheaper to train, the reasoning it performs is computationally more demanding and requires more power to answer questions than an LLM [28]. Even if companies can cut the power use of their models, they might respond by making yet larger ones, thus negating any energy savings [24].

And R1 does not stand apart from its competitors. DeepSeek's announcement claimed that the model topped o1 on three benchmarks and almost matched it on two others [29]. However, there are no standard benchmarks for evaluating the performance of AI models [30], and companies tend to cherry-pick the ones that show their models in the best light [31]. Now that researchers have had months to put R1 through its paces on many different challenges, they can say it excels in some areas but falls behind in others.

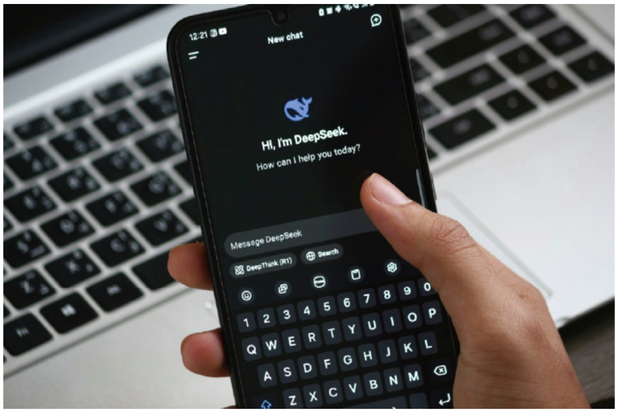

Fig. 1. DeepSeek was little-known outside of AI circles until late January 2025. With the release of its R1 LRM, however, its app quickly shot to number one on Apple's App Store, bypassing ChatGPT offerings. Credit: Abdelrahman Ahmed/Pexels (CC0).

A study published in April 2025 [32], led by Xueyan Mei, an instructor in biomedical engineering and imaging at the Icahn School of Medicine at Mount Sinai (New York City, NY, USA), and Zahi Fayad, a professor of radiology and medicine at the same school, illustrates this mixed record. Mei, Fayad, and colleagues compared three models—R1, o1, and a version of Meta's Llama—on four medically relevant tasks. The Llama version they used, which was an older, smaller model, got the lowest score each time. They found that o1 topped R1 on a multiple-choice exam that all United States doctors must pass. The two models were about equal at classifying tumors and making diagnoses from case descriptions. However, R1 provided clearer reasoning for its diagnoses. And o1 outperformed R1 at writing summaries of imaging studies. The researchers asked radiologists to grade the imaging summaries, and R1 scored lower because it was verbose and, in some instances, hallucinated, or made-up answers, said Mei. Overall, she said, "it has very good common sense, but I would be very careful about the outputs."

Cohn and his team have also been evaluating the model's capabilities on various tasks, including spatial reasoning. They found that o1 was superior. R1 also tends to be slow, Cohn said. His take on the model is that "it was a wake-up call, but in the long run I do not think it is going to make a huge difference." Investors now seem to agree. Within a month, Nvidia's stock had bounced back from its DeepSeek-prompted tumble [33]. And power companies say they still anticipate large increases in electricity demand from AI [34].

R1 suffers from a further drawback. Many organizations—including several government agencies in the United States, a number of universities, and companies such as Microsoft—have banned the company's models from their systems because of concerns over data security [35]. Mei and Fayad were only able to analyze R1's capabilities in their study by running it on an isolated platform.

R1 may be most important for what it suggests about the assumed leadership of Western companies in AI development. Despite its limitations, the model shows that Chinese companies are making rapid advances in the field. And given China's growing competitiveness in AI research, Cohn said, "it will be a challenge for the West to keep up."

References

[1] Saul D. Biggest market loss in history: Nvidia stock sheds nearly $600 billion as DeepSeek shakes AI darling [Internet]. New York City: Forbes; 2025 Jan 27 [cited 2025 Jun 1]. Available from: https://www.forbes.com/sites/dereksaul/2025/01/27/biggest-market-loss-in-history-nvidia-stock-sheds-nearly-600-billion-as-deepseek-shakes-ai-darling/.

[2] Goldman D, Egan M. A shocking Chinese AI advancement called DeepSeek is sending US stocks plunging [Internet]. Atlanta: Cable News Network (CNN); 2025 Jan 27 [cited 2025 Jun 1]. Available from: https://www.cnn.com/2025/01/27/tech/deepseek-stocks-ai-china.

[3] Kearney L, Hampton L. US power stocks plummet as DeepSeek raises data center demand doubts [Internet]. London: Reuters; 2025 Jan 27 [cited 2025 Jun 1]. Available from: https://www.reuters.com/business/energy/us-power-stocks-plummet-deepseek-raises-data-center-demand-doubts-2025-01-27/.

[4] Rajkumar R. DeepSeek's new open-source AI model can outperform o1 for a fraction of the cost [Internet]. New York City: ZDNET; 2025 Jan 21 [cited 2025 Jun 1]. Available from: https://www.zdnet.com/article/deepseeks-new-open-source-ai-model-can-outperform-o1-for-a-fraction-of-the-cost/.

[5] Jones N. 'In Awe': scientists impressed by latest ChatGPT model o1. Nature 2024;634:275–326.

[6] Wiggers K. Google releases its own 'reasoning' AI model [Internet]. San Francisco: TechCrunch; 2024 Dec 19 [cited 2025 Jun 1]. Available from: https://techcrunch.com/2024/12/19/google-releases-its-own-reasoning-ai-model/.

[7] Roose K. Why DeepSeek could change what Silicon Valley believes about A.I. [Internet]. New York City: The New York Times; 2025 Jan 28 [cited 2025 Jun 1].

Available from: https://www.nytimes.com/2025/01/28/technology/china-deepseek-ai-silicon-valley.html.

[8] Patel D, Kourabi AJ, O'Laughlin D, Knuhtsen R. DeepSeek debates: Chinese leadership on cost, true training cost, closed model margin impacts [Internet]. San Francisco: SemiAnalysis; 2025 Jan 31 [cited 2025 Jun 1]. Available from: https://semianalysis.com/2025/01/31/deepseek-debates/.

[9] Yang Z. How Chinese AI startup DeepSeek made a model that rivals OpenAI [Internet]. San Francisco: WIRED; 2025 Jan 25 [cited 2025 Jun 1]. Available from: https://www.wired.com/story/deepseek-china-model-ai/.

[10] Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. Nature 2025;638:300–1.

[11] Gordon N. How DeepSeek erased Silicon Valley's AI lead and wiped $1 trillion from U.S. markets [Internet]. New York City: Fortune; 2025 Mar 30 [cited 2025 Jun 1]. Available from: https://fortune.com/asia/2025/03/30/deepseek-ai-china-us-silicon-valley/.

[12] Wiggers K. 'Reasoning' AI models have become a trend, for better or worse [Internet]. San Francisco: TechCrunch; 2024 Dec 14 [cited 2025 Jun 1]. Available from: https://techcrunch.com/2024/12/14/reasoning-ai-models-have-become-a-trend-for-better-or-worse/.

[13] Gent E. AI models embrace humanlike reasoning [Internet]. New York City: IEEE Spectrum; 2025 May 8 [cited 2025 Jun 1]. Available from: https://spectrum.ieee.org/chain-of-thought-prompting.

[14] Kahn J. 9 things you need to know about OpenAI's powerful new AI model o1 [Internet]. New York City: Fortune; 2024 Sep 13 [cited 2025 Jun 8]. Available from: https://fortune.com/2024/09/13/openai-o1-strawberry-model-9-things-you-need-know/.

[15] Samuel S. Is AI really thinking and reasoning—or just pretending to? [Internet]. New York City: Vox; 2025 Feb 21 [cited 2025 Jun 1]. Available from: https://www.vox.com/future-perfect/400531/ai-reasoning-models-openai-deepseek.

[16] DeepSeek-AI Team. DeepSeek-V3 technical report. 2025. arXiv: 2412.19437v2.

[17] Robison K, Lopatto E. Why everyone is freaking out about DeepSeek [Internet]. New York City: The Verge; 2025 Jan 28 [cited 2025 Jun 1]. Available from: https://www.theverge.com/ai-artificial-intelligence/598846/deepseek-big-tech-ai-industry-nvidia-impac.

[18] DeepSeek-AI Team. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. 2025. arXiv: 2501.12948v1.

[19] Smith MS. What DeepSeek means for open-source AI [Internet]. New York City: IEEE Spectrum; 2025 Jan 31 [cited 2025 Jun 1]. Available from: https://spectrum.ieee.org/deepseek.

[20] Leslie M. Open-source artificial intelligence—how open? How safe? Engineering 2025;47:8–10.

[21] Heaven W. How DeepSeek ripped up the AI playbook—and why everyone's going to follow its lead [Internet]. Cambridge: MIT Technology Review; 2025 Jan 31 [cited 2025 Jun 1]. Available from: https://www.technologyreview.com/2025/01/31/1110740/how-deepseek-ripped-up-the-ai-playbook-and-why-everyones-going-to-follow-it/.

[22] Kambhampati S, Stechly K, Valmeekam K. (How) do reasoning models reason? 2025. arXiv: 2504.09762v1.

[23] Kambhampati S, Stechly K, Valmeekam K, Saldyt L, et al. Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! 2025. arXiv: 2504.09762v2.

[24] Calma J. AI is 'an energy hog,' but DeepSeek could change that [Internet]. New York City: The Verge; 2025 Feb 1 [cited 2025 Jun 1]. Available from: https://www.theverge.com/climate-change/603622/deepseek-ai-environment-energy-climate.

[25] Marshall C. 'Game changer'? What 'DeepSeek' AI means for electricity [Internet]. Washington, DC: E&E News by Politico; 2025 Jan 29 [cited 2025 Jun 8]. Available from: https://www.eenews.net/articles/game-changer-what-deepseek-ai-means-for-electricity/.

[26] Leslie M. Could artificial intelligence's soaring demand for electricity spark a nuclear power revival? Engineering 2025;48:9–11.

[27] Leslie M. Artificial intelligence, like cryptocurrency, eats energy—lots of it. Engineering 2024;32:7–9.

[28] O'Donnell J. DeepSeek might not be such good news for energy after all [Internet]. Cambridge: MIT Technology Review; 2025 Jan 31 [cited 2025 Jun 1]. Available from: https://www.technologyreview.com/2025/01/31/1110776/deepseek-might-not-be-such-good-news-for-energy-after-all/.

[29] DeepSeek Inc. DeepSeek-R1 release [Internet]. Hangzhou: DeepSeek; 2025 Jan 20 [cited 2025 Jun 1]. Available from: https://api-docs.deepseek.com/news/news250120.

[30] Eriksson M, Purificato E, Noroozian A, Vinagre J, Chaslot G, Gómez E, et al. Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation. 2025. arXiv: 2502.06559v2.

[31] Mulligan SJ. The Way We Measure Progress in AI is Terrible [Internet]. Cambridge: MIT Technology Review; 2024 Nov 26 [cited 2025 Jun 1]. Available from: https://www.technologyreview.com/2024/11/26/1107346/the-way-we-measure-progress-in-ai-is-terrible/.

[32] Tordjman M, Liu Z, Yuce M, Fauveau V, Mei Y, Hadjadj J, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med 2025 [forthcoming].

[33] Fellman J. Nvidia stock has clawed almost all the way back from the DeepSeek selloff [Internet]. New York City: Quartz; 2025 Feb 18 [cited 2025 Jun 1]. Available from: https://qz.com/nvidia-stock-ai-chips-deepseek-china-1851765288.

[34] Kearney L. US power companies increase data center demand spending as DeepSeek fears wane [Internet]. London: Reuters; 2025 Feb 13 [cited 2025 Jun 1]. Available from: https://www.reuters.com/business/energy/aep-fourth-quarter-profit-rises-data-centers-boost-demand-power-2025-02-13/.

[35] Rollet C. Microsoft employees are banned from using DeepSeek app, President says [Internet]. San Francisco: TechCrunch; 2025 May 8 [cited 2025 Jun 1]. Available from: https://techcrunch.com/2025/05/08/microsoft-employees-are-banned-from-using-deepseek-app-president-says/.