DOI:10.1145/3732796

Michael A. Cusumano

Technology Strategy and Management

DeepSeek Inside: Origins, Technology, and Impact

How a previously unknown startup company sought to balance cost and performance in large language model development.

The release of DeepSeek V3 and R1 in January 2025 caused declines in the stock prices of companies providing generative artificial intelligence (GenAI) infrastructure technology and datacenter services.[28] These two large language models (LLMs) came from a 200-employee Chinese startup compared to at least 3,500 employees for industry-leader OpenAI.[2,14] DeepSeek seemed to have developed this powerful technology much more cheaply than previously thought possible, potentially disrupting the economics of the entire GenAI ecosystem and the dominance of U.S. companies ranging from OpenAI to Nvidia.[5,6]

DeepSeek-R1 defines itself as "an artificial intelligence language model developed by OpenAI, specifically based on the generative pre-trained transformer (GPT) architecture."[a] Here, DeepSeek acknowledges that the transformer researchers (who published their landmark paper while at Google in 2017) and OpenAI developed its basic technology. Nonetheless, V3 and R1 display impressive skills in neural-network system design, engineering, and optimization, and DeepSeek's publications provide rare insights into the technology. This column reviews, for the non-expert reader, what we know about DeepSeek's origins, technology, and impact so far.

Company Origins

The parent of DeepSeek is High-Flyer, a quant hedge fund founded in 2015 in the city of Hangzhou, China.[b] It relies on AI and machine learning (ML) for stock-trading decisions. Cofounder Liang Wenfeng (born 1985), is a graduate of Zhejiang University. He studied machine vision and started the company with two classmates after researching the topic since 2008. High-Flyer is reported to have 113 employees and $59 million in revenues while managing $8 billion in assets.[11,c]

During 2019–2022, High-Flyer spent approximately $139 million and stockpiled 10,000 Nvidia A100 GPUs to build a supercomputer. The A100 is lower in performance than the datacenter-standard Nvidia H100 (the U.S. government banned both for export to China in October 2023).[9,24,29] In 2021, the Chinese government began a regulatory crackdown on high-frequency speculative stock trading. In response, Liang created a second team of researchers to focus on fundamental AI and ML research. He hired graduates from top Chinese universities and worked closely with Chinese academics. The team evaded government scrutiny because it was separate from the investment group and not

a In response to the question: How do you define what you are? See https://bit.ly/3FbIEOk

b See https://bit.ly/3FawS6O

c Datanyze, High-Flyer Profile and History; https://bit.ly/4dh92CV

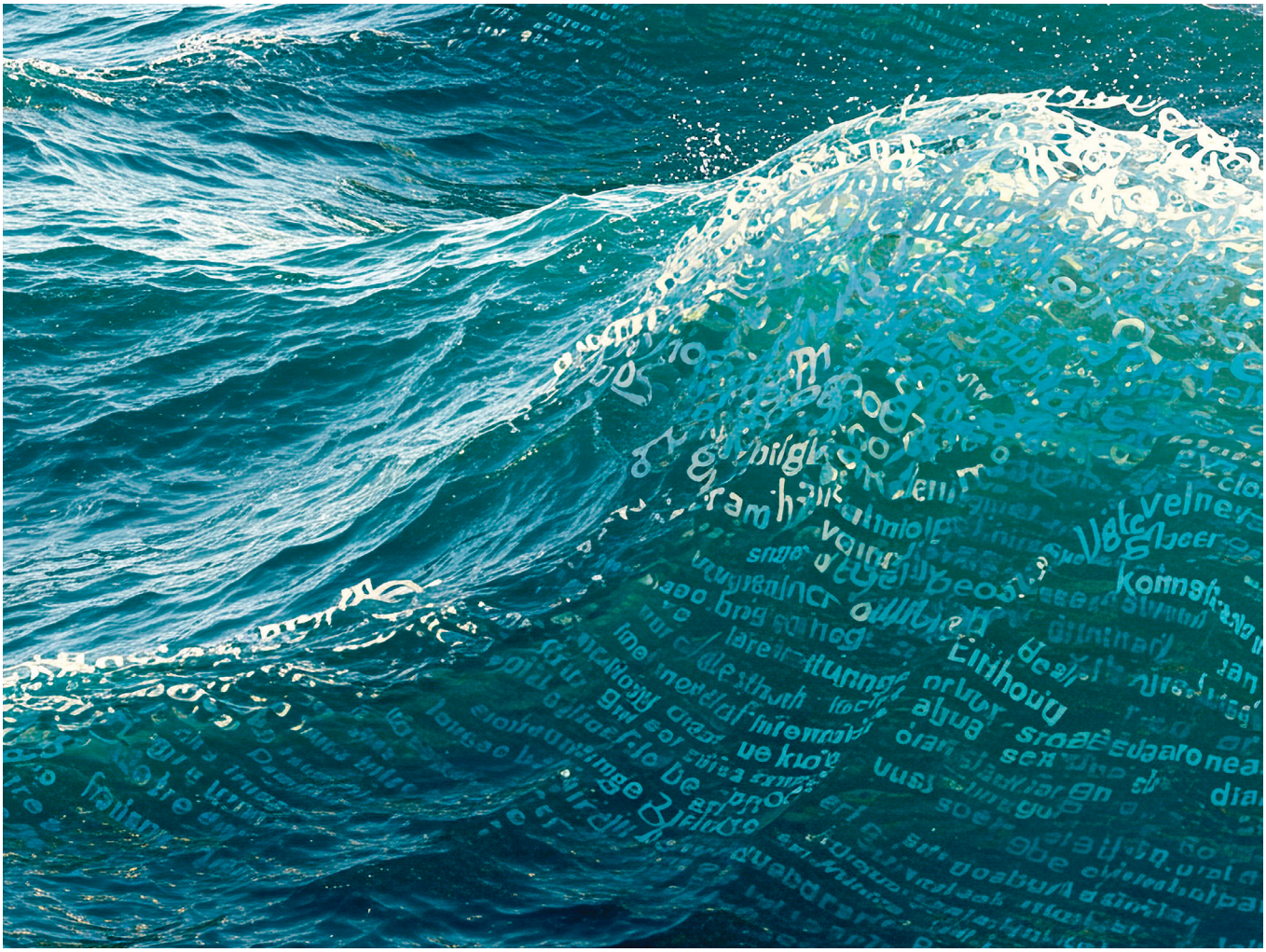

IMAGE BY ANDRIJ BORYS ASSOCIATES WITH ASSISTANCE OF SHUTTERSTOCK.AI

focused on consumer AI.[9,25] In 2023, High-Flyer spun off this second team as DeepSeek (see the accompanying table). One year later, the new company released products and reports that attracted worldwide attention. (In March 2025, the Chinese government also named DeepSeek a "national treasure" and placed travel restrictions on key employees.[26])

Low Cost, High Performance

What made headlines was DeepSeek's claim, based on a technical report published on its website dated Dec. 27, 2024, that the company spent merely $5.6 million training V3. DeepSeek also claimed V3 approximated or exceeded the performance of other leading LLMs (Qwen 2.5, Alibaba; Llama 3.1, Meta; GPT-4o, OpenAI; Claude 3.5, Anthropic). The cost was so low because V3 required only 2.788 million hours of pre-training and post-training, done on a cluster of 2,000 Nvidia H800 GPUs (the lower-bandwidth version of H100, permissible for export to China), at an estimated cost of $2/hour. By contrast, to train GPT-4, it is reported that OpenAI used a cluster of 16,000 Nvidia H100 GPUs and spent at least $100 million.[12,29] It had generally been accepted that companies need to spend $100 million to $1 billion to train their LLMs.[11] The higher number covers variations, such as the nine different versions of Meta's Lama 3.[d]

Table. DeepSeek development evolution.

Nov. 2023	DeepSeek Coder
Dec. 2023	DeepSeek-V1 ("General" LLM)
May 2024	DeepSeek-V2
Dec. 27, 2024	DeepSeek-V3 Technical Report
Jan. 10, 2025	DeepSeek-V3
Jan. 20, 2025	DeepSeek-R1 ("Reasoning" LLM)
Jan. 22, 2025	DeepSeek-R1 White Paper

It probably did not cost much to develop R1 once DeepSeek had V3.[1] However, DeepSeek has not revealed the separate development costs for R1 or its predecessors and variants. It is also unknown how much hardware High-Flyer and DeepSeek had access to in addition to the supercomputer. The research firm SemiAnalysis estimates DeepSeek had access to as many as 50,000 Hopper-class Nvidia GPUs and, for V3/R1, spent more than $500 million on GPU hardware and about $1.3 billion in development costs.[9] These estimates, which seem to be approximate guesses made in hindsight, would be comparable to development costs for the reasoning GPT models from OpenAI and the Llama 3 series from Meta.

The December 2024 technical report also states that, "During pre-training, we train [sic] DeepSeek-V3 on 14.8T high-quality and diverse tokens."[8] One token equals approximately four characters or 0.75 English words. Fifteen billion tokens are approximately the

d Hugging Face, Meta Llama: The Llama Family; https://bit.ly/4dhLlue

Coming Next Month in COMMUNICATIONS

Descriptive, Normative, and Action AI Ethics: The Case of X

Two Types of Data Privacy Controls

The Future of Professional Ethics in Computing

AI and Trust

Thinking Fast and Slow in Human and Machine Intelligence

It Takes a Village: Bridging the Gaps between Current and Formal Specifications for Protocols

Understanding Mobile App Reviews to Guide Misuse

SRAM Has No Chill

Plus, the latest news about zero-knowledge proofs, quantum leaps, and a time zone for the moon.

amount of high-quality text available on the English-language Internet. It is also the estimate for pre-training tokens Meta used for Llama 3 and twice the estimate for OpenAI's GPT-4 of 6.5 trillion tokens.[10] Training on DeepSeek-V3 reportedly took many fewer hours and much less money than comparable LLMs but covered a similar or much higher level of content. Sasha Rush concludes as well that V3's performance is comparable to OpenAI GPT-4o but with substantially less training costs, while R1 resembles an open source version of OpenAI-o1, which had been the market leader for reasoning LLMs.[22]

Reasoning LLMs generate better answers if they run more computations in the inference (question-answering) stage. The reasoning approach works particularly well when there are correct answers the system can check against and provide rewards for, like in mathematical modeling, coding, or logical reasoning.[22] DeepSeek built R1 as a reasoning version of V3 and released two smaller versions of R1 based on distillations of the open source Llama 3 (Meta) and Qwen 2.5 (Alibaba).[7,21,e] V3 consumed the bulk of the engineering work but R1 is important because it showed DeepSeek that "reasoning emerges as a behavior from pure reinforcement learning."[19]

Here we see a difference in training strategy. OpenAI built its reasoning models by starting with human-supervised fine-tuning and then using reinforcement learning, followed again by extensive human fine-tuning. By contrast, DeepSeek skipped human pre-training and allowed the model to improve its reasoning capabilities by learning on its own (that is, reinforcement learning with rewards) and then a few hours of human post-training. As it built R1, DeepSeek also enhanced V3 through rounds of reinforcement learning, with rewards and verifications of answers. V3 and R1 used the same training data, and the V3-base model is inside R1. However, R1 benefited from this "distillation" of learning from V3. Moreover, designing R1 to show the reasoning process made it easier to fine-tune.[22]

e Hugging Face, DeepSeek-R1-Distill-Llama-8B; https://bit.ly/4dhLlue

Technology Essentials

Part of the explanation for the low-cost, high-performance of V3/R1 is DeepSeek's follower position and access to open source technology.[20] But equally important seems to be DeepSeek's technology and training strategy. Most experts cite the clever architecture and multistage training, with human fine-tuning only in a post-training phase. In particular, three complementary approaches and supporting techniques balanced engineering cost and system performance: mixture of experts, reinforcement learning, and distillation.

Mixture of experts (MoE) and supporting techniques. MoE architectures use several small specialized language models to spread out the data analysis and training load. Different prompts or "weights" call up specialized experts that, of the total parameters available, access only a small subset, such as 37 billion out of 671 billion parameters in R1. Accessing a subset of parameters, called *Sparse Neural Network Training*, avoids redundant or unnecessary calculations. By contrast, GPT-4 appears as a "dense" neural network with access to 1.8 trillion parameters for every operation.[16] The MoE concept is not new but has been difficult to scale.[22] The French startup Mistral AI did release in December 2023 the "open-weight" MoE LLM 8x7B, with eight experts (two operational at any given time), each accessing a total of 166 billion parameters and resembling a "scaled-down GPT-4." Apparently, it was inno-

As it built R1, DeepSeek also enhanced V3 through rounds of reinforcement learning, with rewards and verification of answers.

vative and efficient but did not garner the same attention as the efficient and low-cost DeepSeek V3/R1.[13]

DeepSeek relied on specific techniques to minimize compute resources needed to share data across the smaller experts. It used parallelization to coordinate the computational tasks running on different clusters of networked computers (GPUs). It used quantization to improve computational efficiency by limiting the precision of calculations (decimal points) to the lowest level necessary for a specific computation.[22] Examples would be converting 32-bit floating-point numbers to 16-bit, 8-bit, or even integer (whole-number) formats.

Reinforcement learning (RL) and supporting techniques. The RL approach focuses on ML via trial and error with rewards for particular outcomes, in R1's case, accuracy and format of the answers. OpenAI and other reasoning LLMs also use reinforcement learning and rewards.[18] However, OpenAI has relied heavily on supervised training by human experts that requires enormous amounts of data and time for pre-labeling the data sets and feeding the system questions and correct answers.[16] DeepSeek says it used human supervision only in post-training to fine-tune its models and spent very few hours.[8] In general, the analogy would be to train a computer to play chess by giving it the rules on how to move pieces and then letting the machine learn mostly by playing matches rather than pretraining with insights from a human grandmaster.[11]

Other techniques complemented RL. Chain-of-thought (CoT) reasoning divides a large, complex problem into intermediate steps that are easier to analyze and refine.[17] DeepSeek used RL and CoT training to turn V3 from a relatively simple LLM into the more sophisticated reasoning R1 by asking it to show the steps used to analyze and answer questions.[21] Another complementary technique is group relative policy optimization (GRPO). Many LLMs incorporate a slow and expensive training approach with a "policy model" (rules or guidelines) to decide on actions and then a separate "critic model" to evaluate those actions. GRPO skips the critic model step by taking groups of scores (data from several actions) and learning by comparing the scores.[29] In the R1 white paper, DeepSeek reported: "we use DeepSeek-V3-Bases as the base model and employ GRPO as the RL framework to improve model performance in reasoning ... We directly apply RL to the base model without relying on supervised fine-tuning (SFT) as a preliminary step. This approach allows the model to explore chain-of-thought (CoT) for solving complex problems."[7] DeepSeek also used multi-token prediction (MTP) during V3 pretraining to improve data efficiency, noting: "Instead of predicting just the next single token, DeepSeek-V3 predicts the next two tokens through the MTP technique ... which we have observed to enhance the overall performance on evaluation benchmarks."[8]

There has been concern that DeepSeek's low prices could commoditize GenAI techology and lead to steep drops in company sales.

Distillation. This approach uses a large "teacher" model to transfer knowledge to train or fine-tune a smaller "student" model. The smaller model requires much less memory and compute time to operate, and can be faster during the inference stage. For example, the teacher model can transfer prediction outputs that serve as instructions for how to answer specific types of questions, saving the student model from the extensive training and computations normally required to arrive at those instructions.[17] There have been suggestions that DeepSeek distilled outputs or acquired proprietary data from OpenAI models, but there is no evidence of this.[22,23] DeepSeek used a distillation of R1 for post-training of V3 to add some reasoning capability, and it used V3 to help train R1.[8] DeepSeek also experimented with small versions of R1 trained on distillations of the open source Qwen 2.5 and Llama 3. These distilled versions performed better than a preliminary model (DeepSeek-R1-Zero), despite being much smaller, though they did not perform as well as R1.[19] The DeepSeek team offered two conclusions regarding this combination of distillation and reinforcement learning as well as the limitations of their approach: "First, distilling more powerful models into smaller ones yields excellent results, whereas smaller models relying on the large-scale RL ... require enormous computational power and may not even achieve the performance of distillation. Second, while distillation strategies are both economical and effective, advancing beyond the boundaries of intelligence may still require more powerful base models and larger-scale reinforcement learning."[7]

Market Impact

DeepSeek optimized V3 and R1 for different tasks. V3 does very well on standard benchmarking tests for logic problems and code writing. R1 is better at deeper reasoning and math.[12,16] Both V3 and R1 are said to be comparable or better than LLMs from Anthropic (Claude) and xAI (Grok), though still behind Google (Gemini).[11] Chatbot Arena at University of California Berkeley publishes a ranking where R1 came in third, matching the performance of ChatGPT-4o.[3]

The cost to access V3 and R1 application programming interfaces (APIs) is so low that DeepSeek has disrupted the economics of the LLM API market.[4] In early 2025, DeepSeek priced V3 at $0.07 per million input tokens and $0.11 per million output tokens, and R1 at $0.55 for input and $2.19 per million output tokens. By contrast, OpenAI priced GPT-4o at $15 per million input and $60 per million output tokens (prices that were approximately half of what OpenAI had charged earlier).[29]

Price-performance has made DeepSeek very popular among open source developers, and all major cloud service providers support access. In China, car manufacturers, local governments,

hospitals, state-owned enterprises, and private companies have embedded DeepSeek into their applications. Tencent has incorporated DeepSeek into its search engine for WeChat and the Chinese version Weixin. Baidu uses DeepSeek in its search engine. Other major companies such as Huawei, NetEase, Bytedance/TikTok are also users.[15] In the U.S., AMD has integrated DeepSeek-V3 into its Instinct MI300X GPUs to optimize AI inference and GPU performance.[28] However, some countries (the U.S., Australia, South Korea, India, Italy, and Taiwan) have warned about Chinese ownership or prohibited usage on government devices.[27]

There has been concern that DeepSeek's low prices could commoditize GenAI technology and lead to steep drops in company sales. But another potential outcome is Jevons Paradox—the observation that improvements in efficiency and lower prices do not necessarily lead to falling revenues.[3] If demand is sensitive to price, then usage and sales could increase dramatically as more people access the less expensive product or service. So far, Jevons Paradox seems to be holding up: Demand for GenAI software and GPU hardware continues to rise. Reasoning LLMs are improving and likely to find many more applications.

Debates continue around the significance of DeepSeek's innovations. Dario Amodei, cofounder of Anthropic and former VP of Research at OpenAI, describes DeepSeek's achievement as mainly an important advance in engineering efficiency. He sees V3/R1 as part of a shift already under way from a focus during 2020–2023 on expensive and time-consuming pretraining, running software on ever-more-powerful and expensive hardware, to more economical machine-based reinforcement learning, sometimes with less than state-of-the-art hardware.[1]

However we view DeepSeek, it seems noteworthy that a Chinese start-up, operating under American export constraints, has demonstrated clever ways to balance cost and performance in LLM development. DeepSeek's open source code and sharing of detailed information has already made GenAI technology more accessible and understandable. C

References

1. Amodei, D. *On DeepSeek Export Controls* (Jan. 2025); https://bit.ly/4iPmaR6
2. Baek, D.H. *How Many Employees Does DeepSeek AI Have* (Jan. 28. 2025); https://bit.ly/4kpCyJl
3. Cohen, B. DeepSeek arrived. America freaked. What happens now? *Wall Street J.* (Jan. 31, 2025); https://bit.ly/4m1OdQ1
4. Cusumano, M.A. et al. Generative AI as a new platform for applications development. *An MIT Exploration of Generative AI: From Novel Chemicals to Opera*. D. Huttenlocher and A. Ozdaglar, Eds. MIT Press (2024).
5. Cusumano, M.A. Generative AI as a new innovation platform. *Commun. ACM 66*, 10 (Oct. 2023).
6. Cusumano, M.A. Nvidia at the center of the genAI ecosystem—for now. *Commun. ACM 67*, 1 (Jan. 2024).
7. DeepSeek-AI. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* (Jan. 22, 2025).
8. DeepSeek-AI. *DeepSeek-V3 Technical Report* (Dec. 27, 2024).
9. Dou, E. et al. China's new AI star DeepSeek grew on more than a shoestring budget. *Washington Post* (Jan. 31, 2025).
10. Educating Silicon *How Much LLM Training Data Is There, In the Limit?* (May 9, 2024); https://bit.ly/44qBboO
11. Huang, R. Silicon Valley is raving about a made-in-China AI model. *Wall Street J.* (Jan. 27, 2025).
12. Metz, C. What to know about DeepSeek and how it is upending AI. *New York Times* (Jan. 27, 2025).
13. Mittal, A. *Mistral AI's Latest Mixture of Experts (MoE) 8x7B Model* (Dec. 15, 2023); https://bit.ly/3GHfgjl
14. Mortensen, O. How many people work at OpenAI? *Statistics and Facts* (Dec. 2, 2025); https://bit.ly/4kpCyJl
15. Olcott, E. and Ding, W. DeepSeek spreads across China with Beijing's backing. *Financial Times* (Feb. 27, 2025).
16. Pipe, A. and Rattner, N. How DeepSeek's lower-power, less-data model stacks up. *Wall Street J.* (Feb. 16, 2025).
17. Pure AI Editors. *Understanding LLM Distillation, Enabling Revolutionary DeepSeek R1 Model* (Feb. 3, 2025); https://bit.ly/4jLEITu
18. Ramakrishnan, R. *The Road to ChatGPT: An Informal Explainer on How ChatGPT Was Built*. MIT Sloan School of Management, (Mar. 5, 2023); https://bit.ly/3GC8xaB
19. Raschka, S. *Understanding Reasoning LLMs* (Feb. 5, 2025); https://bit.ly/3RNoqgH
20. Romero, L.E. ChatGPT, DeepSeek, or Llama? Meta's LeCun says open source is the key. *Forbes.com* (Jan. 27. 2025); https://bit.ly/4m2gx4N
21. Roose, K. Why DeepSeek could change what Silicon Valley believes about AI. *New York Times* (Jan. 28, 2025).
22. Rush, S. *How DeepSeek Changes the LLM Story* (Feb. 3, 2024); https://bit.ly/4m7Qqti
23. Satoh, R. What is AI distillation and what does it mean for OpenAI? *Nikkei Asia* (Jan. 30, 2025).
24. Swanson, A. and Tobin, M. Do China's AI advances mean U.S. technology controls have failed? *New York Times* (Jan. 28, 2025).
25. Tobin, M., Mozur, P., and Stevenson, A. DeepSeek's rise: How a Chinese start-up went from stock trader to AI star. *New York Times* (Jan. 28, 2025).
26. TOI Tech Desk DeepSeek gets 'National Treasure in China' status. *TimesofIndia.com* (Mar. 19, 2025).
27. Yu, Y. Is DeepSeek next in line for a TikTok-like U.S. ban? *Nikkei Asia* (Feb. 12, 2025).
28. Zhou, C. and Truitt, J.S. DeepSeek crashes Nvidia stock with prospect of cheaper AI. *Nikkei Asia* (Jan. 28,2025).
29. Zhou, C. and Cheng, T.-F. Is China's DeepSeek an Nvidia killer or overhyped? 4 things to know. *Nikkei Asia* (Jan. 28, 2025).

Michael A. Cusumano (cusumano@mit.edu) is the SMR Distinguished Professor and former Deputy Dean at the Massachusetts Institute of Technology Sloan School of Management, Cambridge, MA, USA, and coauthor of *The Business of Platforms* (2019). The author thanks Rama Ramakrishnan, Imran Sayeed, and Nagarjuna Venna for their comments and help developing this Opinion column.