OPINION

Noh, Sungho

DeepSeek (深度求索) AI and its Implications for Innovation in Financial Services

The recent success of the DeepSeek (深度求索) artificial intelligence (AI) model signals a notable shift in both the competitive landscape of generative AI development and the research and development (R&D) strategies of the financial services industry. DeepSeek AI is a reasoning model, which is the most advanced version of large language models (LLMs), and has demonstrated high learning efficiency and strong reasoning capabilities at relatively low costs by integrating a range of innovative technologies. Notably, DeepSeek sets itself apart from closed-source and high-cost models by openly releasing its core technical components, thereby fostering competition among emerging AI developers and contributing to a more open and dynamic AI ecosystem. This shift is expected to shorten AI development cycles, accelerate the commercialization of innovative technologies, and facilitate the broader adoption of AI-powered services. For financial institutions, the case of DeepSeek AI implies that strategically leveraging their existing IT assets could reduce reliance on external AI solutions and support a more proactive approach to digital transformation. Over the long term, such a strategy has the potential to enhance the competitiveness of financial institutions and enhance value across the industry.

Since January of this year, the DeepSeek (深度求索) artificial intelligence (AI) model, known for its performance comparable to that of industry-leading AI models at substantially lower costs, has further intensified the race to advance generative AI (GenAI) and large language model (LLM) development. Developed by DeepSeek, a research lab founded by Chinese hedge fund High-Flyer (幻方), the model claims to match or exceed the capabilities of OpenAI's

* All opinions expressed in this paper represent the author's personal views and thus should not be interpreted as Korea Capital Market Institute's official position.

* Ph.D., Research Fellow, Financial Services Industry Department, Tel: 82-2-3771-0629, E-mail: nohs@kcmi.re.kr

leading models while significantly reducing development costs.[1] As such, DeepSeek is poised to be a disruptive force in the global AI competition landscape dominated by US-based companies.

This article examines the technical foundations of DeepSeek AI's strong performance relative to conventional LLMs and the limitations of its current capabilities. It further explores the broader implications of DeepSeek's success for dynamics in AI development, as well as its potential to reshape R&D-driven innovation and value enhance ment in the financial services industry.

The technological innovation of the DeepSeek AI model

In January, DeepSeek unveiled its base LLM, DeepSeek-V3, along with a reasoning model, DeepSeek-R1, enhanced with logical reasoning capabilities.[2] Among the two, DeepSeek-R1 has demonstrated superior performance and broader applicability, and its technological advancements can be categorized into three core components: Chain of Thought (CoT), Reinforcement Learning (RL), and Distillation (see Table 1).

Table 1. Key technological innovations of DeepSeek AI

Identified Challenge		Applied Technique		Improvement
Lack of transparency in reasoning processes and weak logical coherence in responses	→	Chain of Thought (CoT)	→	Makes reasoning steps explicit; improves the quality of responses
Insufficient quantity and quality of training data	→	Reinforcement Learning (Group Relative Policy Optimization)	→	Enhances learning efficiency using relatively limited data
Increased training time and cost due to large-scale models	→	Distillation	→	Reduces cost by downsizing the model while maintaining performance

1) DeepSeek-AI, 2025, DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning.

2) Compared to the models developed by OpenAI, the market leader in generative AI development, DeepSeek-V3 and DeepSeek-R1 compete in the same category as GPT-4 and o1, respectively.

A key limitation of LLMs is their lack of transparency during the reasoning process. The complex and probabilistic mechanism by which LLMs generate their responses remains opaque, making it difficult for users to comprehend the model's internal "thought process" leading to a given conclusion. To address this issue, DeepSeek-R1 adopts a Chain of Thought (CoT) approach, which prompts the model to articulate its reasoning in explicit steps. For example, when solving a mathematical problem, the model is guided to describe each step in the problem-solving process rather than producing a simple, direct answer.[3] By incorporating CoT, DeepSeek-R1 moves beyond merely generating correct answers and self-improves problem-solving capabilities in mathematical and logical tasks to deliver better performance relative to conventional language models.

While the CoT approach has already proven effective in advanced reasoning models such as OpenAI's o1, it is also known to require significantly larger training datasets, resulting in higher computational costs and extended training time. To mitigate these challenges, DeepSeek has introduced a reinforcement learning (RL) algorithm called Group Relative Policy Optimization (GRPO). This technique enhances reasoning performance by enabling the model to iteratively complete similar tasks and learn from the most frequent and accurate responses generated during training. This process is analogous to comparing students' answers without a pre-existing reference to arrive at the most reliable solution. Through this self-improving algorithm, DeepSeek-R1 can reduce the need for human intervention, lower training costs, and still achieve performance on par with, or even superior to, market-leading models.

Moreover, the high computational burden associated with deploying and using large-scale generative AI models has historically posed a major cost barrier. DeepSeek has addressed this limitation through a distillation technique, which transfers the knowledge learned by a large-scale model to a smaller, more efficient one. This approach enables the distilled model to retain much of the original model's performance while significantly reducing the computational resources required for operation. As a result, DeepSeek has succeeded in building lightweight versions of proven high-performance models, such as Llama-3 and Qwen, that deliver competitive performance at a lower cost to ease the financial burden of generative AI adoption.

3) Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D., 2022, Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35, 24824-24837.

DeepSeek AI's success: Opportunities and challenges

One of the most notable aspects of DeepSeek AI is its release as an open-source model, allowing external researchers to directly examine and build upon its core components.[4] This level of openness stands in stark contrast to the prevailing practice of most leading AI companies, which typically treat critical elements of their models, such as weights and training methodologies, as proprietary information. Before the emergence of DeepSeek, it was widely recognized that the highest-performing models, including OpenAI's GPT and Anthropic's Claude, were closed-source systems developed in the US, while open-source systems generally lagged in performance. The DeepSeek case illustrates how quickly the balance of power in AI development can shift.

Despite these positive aspects, it is unlikely that DeepSeek AI will maintain a leading position over the long term. The model's success was largely indebted to minimizing human intervention and managing costs by utilizing publicly available technologies, such as Meta's Llama and OpenAI's APIs, to optimize its performance. Such a strategy also suggests that the model's performance can be further improved with additional human inputs. Moreover, by releasing the model as an open-source system, DeepSeek has allowed external researchers to study and adapt its underlying mechanisms, potentially enabling the development of more efficient architectures or significantly enhanced versions of existing models.[5] Paradoxically, this openness could intensify competition in AI development and ultimately undermine DeepSeek's ability to retain its leading position over time.[6]

The dual nature of DeepSeek AI's achievement not only offers critical insights into the evolution of generative AI development but also provides valuable guidance for the financial services industry as it seeks to adopt such technologies.

---

4) AI models can be released typically as one of the following two ways: (1) full disclosure of training data and methods, and (2) disclosure of the models' trained weights. Technically, the former is more open than the latter. However, since Meta's Llama model adopted the latter approach, both have been referred to as open-source architectures. DeepSeek AI, which has not released its training data, falls into the second category.

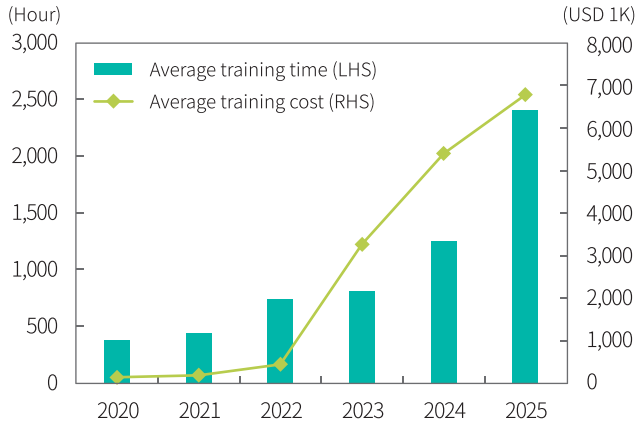
5) In practice, efforts to improve the cost efficiency of AI models are underway in both the industry (Criddle, C., 2025. 3. 2, AI companies race to use 'distillation' to produce cheaper models, Financial Times.) and academia (Muennighoff et. al., 2025, s1: Simple test-time scaling, *arXiv preprint arXiv*:2501.19393).

6) Notably, the generative AI model EXAONE, released by LG AI Research in March 2025, has attracted attention for demonstrating performance comparable or even superior to that of existing reasoning models, including DeepSeek-R1. LG AI Research, 2025, EXAONE Deep: Reasoning Enhanced Language Models, *arXiv preprint arXiv*:2503.12524.

Changing competitive dynamics in the AI development landscape

The success of DeepSeek is expected to reshape the competitive dynamics of the AI industry. It was widely assumed that the generative AI market would continue to be dominated by leading players such as OpenAI and Google, which benefit from first-mover advantages and massive investments.[7] However, DeepSeek has demonstrated that it can achieve high performance with relatively limited resources. Notably, the DeepSeek research team devised a cost-efficient method for training LLMs, even with restricted access to cutting-edge GPUs, which could restrain the recent trend of exponentially rising development costs (see Figure 1). DeepSeek-R1's high performance relative to its cost appears to have placed downward pressure on the pricing of subsequent models (see Figure 2).

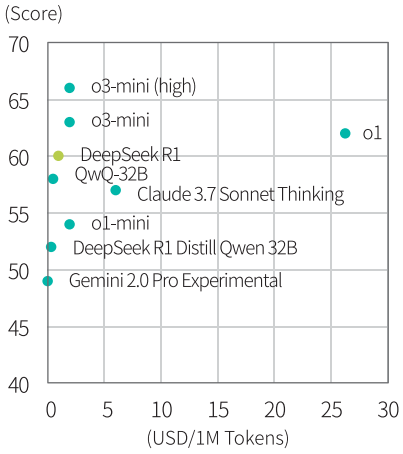
Figure 1. Training costs of generative AI models



Note: This figure represents average training time and costs for all generative AI models released in the corresponding year. Training costs are adjusted to the 2023 US dollar value.

Source: Epoch AI

Figure 2. Price vs. Performance



Note: As of March 18, 2025, this figure represents the top 10 AI models based on performance evaluation scores, with price per token (the horizontal axis) and the performance score (the vertical axis).

Source: Artificial Analysis

Moreover, DeepSeek AI's ability to deliver competitive performance through an open-source framework may accelerate the overall pace of technological advancement. The model's openness has enabled AI researchers to analyze and improve upon its architecture, potentially narrowing the performance gap between open-source and closed-source models.

7) Korinek, A. & Vipra, J., 2024, Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence, *NBER Working Paper* no. 33139.

This development holds particular significance for late entrants in the AI race, such as Korea. In terms of follow-up research, the value of leading AI technologies will likely be assessed not only by performance metrics but also by their capacity to support collaboration.

R&D and innovation strategy in the financial services industry

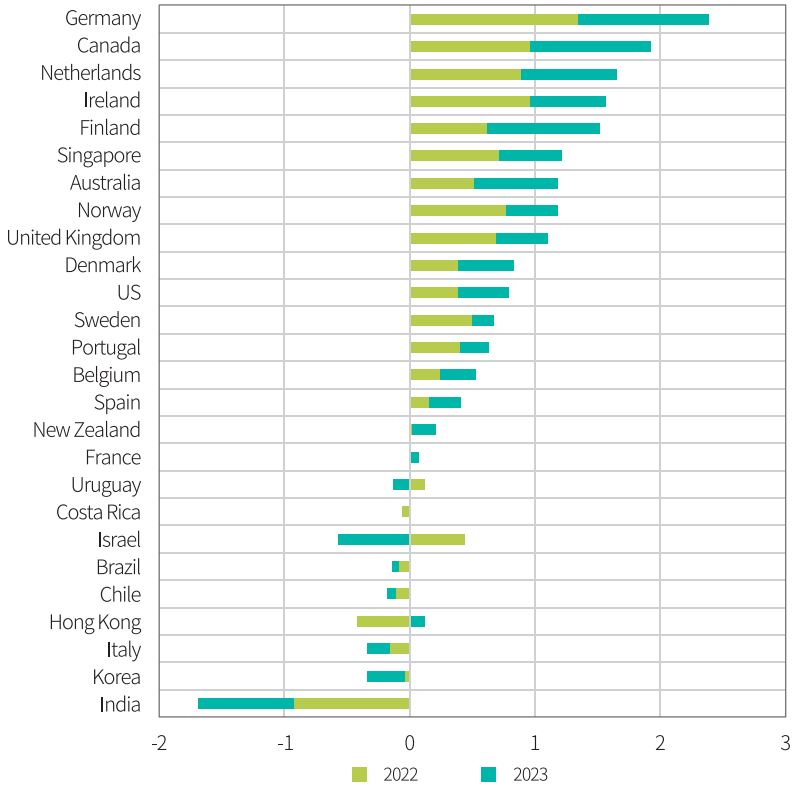
The case of DeepSeek AI underscores the need for a shift in how financial institutions approach strategic R&D. DeepSeek leveraged GPU resources secured by its parent company, High-Flyer,[8] to develop an advanced AI model. Despite limited access to high-performance GPUs as a result of US export restrictions, DeepSeek was able to build a competitive model by employing innovative techniques, as discussed earlier. Such an achievement was made possible by top-level research talent[9] and a long-term commitment to R&D over short-term profitability.[10] This serves as a compelling example of how financial institutions can generate industry-wide impact by effectively utilizing internal IT infrastructure and research capabilities. In the age of digital transformation, DeepSeek's success suggests that financial institutions should move beyond reliance on third-party IT solutions. By developing tailored services that are both efficient and secure, underpinned by their own technological capabilities, they can play a more proactive role in driving market innovation and enhancing corporate value.

8) Patel, D., Kourabi, A.J., O'Laughlin, D., Knuhtsen, R., 2025. 1. 31, DeepSeek Debate: Chinese Leadership on Cost, True Training Cost, Closed Model Margin Impacts, SemiAnalysis.

9) Chen, C., How a top Chinese AI model overcame US sanctions, MIT Technology Review.

10) Wu, Z., DeepSeek focuses on research over revenue in contrast to Silicon Valley, Financial Times.

Figure 3. Net AI talent migration by country



Note: This figure shows the net annual migration of professionals with AI skills per 10,000 LinkedIn members. Positive values indicate a net inflow, while negative values represent a net outflow.

Source: AI Index (2024), Stanford HAI

To enhance its value in the longer term through R&D, financial institutions must establish clear objectives and build infrastructure. It is essential to identify the business areas where AI can make the largest impact and set concrete goals that reflect both challenges and corresponding solutions. In parallel, they should focus on securing AI talent and enhancing technical capabilities through collaboration with startups and research organizations. Although investment in IT infrastructure and human capital within Korea's financial services sector has been on the rise, it remains below the levels observed in advanced economies.[11] Notably, while the US, Canada, and Europe have seen a net inflow of AI professionals in recent years, Korea has been experiencing a net outflow of AI talent (see Figure 3).[12] Reversing this trend will require sustained, long-term strategic investment in R&D resources.

11) Noh, S.H., 2023, Generative AI-driven Productivity Innovation and the Financial Services Industry's Responses, Korea Capital Market Institute Capital Market Focus 2023-18.

12) It should be noted that the data were obtained via LinkedIn, a network platform with relatively low penetration in Korea, raising the possibility that the outflow of AI talent may have been overestimated relative to the inflow.