

Harvard Business Review

Why DeepSeek Shouldn't Have Been a Surprise

by Prithwiraj (Raj) Choudhury, Natarajan Balasubramanian and Mingtao Xu

January 31, 2025



Anadolu/Getty Images

Summary

The Chinese startup DeepSeek shocked many when its new model challenged established American AI companies despite being smaller, more efficient, and significantly cheaper. However, management theory — specifically disruption theory — could have predicted that a challenger like this would inevitably come along. After all, disruptive innovation is all about low-cost alternatives that aren't cutting-edge but perform adequately for many users. This, it seems, is exactly how DeepSeek has created the shockwave that it has. DeepSeek isn't sui generis. It's operating along similar lines to many other Chinese, which differ from their American counterparts in two significant ways: 1) They often use cheaper hardware and leverage an open (and therefore cheaper) architecture to reduce cost, and 2) many Chinese LLMs are customized for domain-specific (narrower) applications and not generic tasks. Management theory also offers insight on how companies should proceed from here. On the one hand, a benefit of having multiple LLM models deployed within an organization is diversification of risk. On the other hand, a benefit of working with a single supplier is reduced administrative costs and better understanding of capabilities on both sides of the partnership. Prior management theories also suggest a third possibility: plural governance, which involves using a combination of external suppliers and internal developers to leverage an emerging technology.

The Chinese AI startup DeepSeek caught a lot of people by surprise this month. Its new model, [released on January 20](#), competes with models from leading American AI companies such as OpenAI and Meta despite being [smaller, more efficient, and much, much cheaper](#) to both train and run.

Yet, the Chinese company's success could likely have been predicted by management theory — specifically, the theory of disruptive innovation. After all, disruptive innovation is all about low-cost alternatives that aren't cutting-edge but perform adequately for many users. This, it seems, is exactly how DeepSeek has created the shockwave that has challenged some of the assumptions of the American AI industry and [sent tech and energy stocks tumbling](#) as a result.

If management theory can help explain what just happened, it also offers insight as to where we might go from here. Drawing on theories of technological change, we highlight implications for what this disruption means for global firms, as their leaders grapple with whether to license Chinese or American large language models (LLMs) or keep their options open.

The Differences Between Chinese and American LLMs

It is important to first point out that the Chinese LLMs differ from their American counterparts in two significant ways: 1) They often use cheaper hardware and leverage an open (and therefore cheaper) architecture to reduce cost, and 2) many Chinese LLMs are customized for [domain-specific](#) (narrower) applications and not generic tasks. However, models like DeepSeek-R1 are emerging as more general-purpose reasoning models.

American LLM models are typically trained on cutting-edge GPU clusters that include tens of thousands of NVIDIA's most advanced chips and require enormous capital investment and cloud infrastructure. In contrast, at least partly because of export controls on advanced chips, most Chinese LLMs rely on [distributed training](#) across multiple, less-powerful GPUs. Yet they achieve competitive — although not necessarily cutting-edge — performance through [more efficient architecture](#). For example, [DeepSeek's Multi-Head Latent Attention \(MLA\) and Mixture of Experts \(MOE\) architecture](#) are designed to reduce memory usage, allowing for more efficient utilization of computing resources.

The embrace of open-source codebases also plays a crucial role in Chinese LLM development. DeepSeek-V3, the foundation model powering its latest reasoning system, and DeepSeek-R1 have both been released under the [MIT open-source license](#). This permissive license encourages widespread adoption by allowing users to freely use, modify, and distribute the software, including for commercial purposes, with minimal restrictions. The advantage of this efficient architecture and open-source approach is most evident when comparing training costs: DeepSeek's reported [\\$5.6 million](#) (for V3) compared to the [\\$40 million to \\$200 million](#) U.S. AI companies such as OpenAI and Alphabet have reported spending on their LLMs.

In addition, while U.S. models prioritize general-purpose queries trained on vast, globally sourced datasets, many Chinese LLMs are also engineered for [domain-specific precision](#). Chinese tech giants, such as Alibaba, Tencent, Baidu, and ByteDance, as well as emerging startups like DeepSeek, offer [industry-specific applications](#) powered by their LLMs that are deeply integrated into China's digital ecosystems.

In summary, Chinese LLMs rely on less advanced hardware and initially focus on lower end — more specific, less general-purpose — applications that require less

computational power. This also means many Chinese LLMs are priced at [the lower end](#). For instance, Alibaba's [Qwen plus](#) and ByteDance's [Doubao 1.5-pro](#) costs less \$0.30 per 1 million tokens of output compared with more than \$60 for [OpenAIo1](#) and Anthropic's [Claude 3.5 Opus](#).

This is classic [disruption theory](#) in play. It is a repeat of how mini-mills disrupted integrated steel plants decades ago. Disruption theory predicts that an inferior technology at its inception (such as the electric arc furnace) customized to specific low-end tasks (such as producing steel for lower quality rebars) emerge as a threat to higher-end producers (such as integrated steel plants) whose sole focus is higher-end customers offering greater margins (such as customers of high-end sheet steel). Slowly and steadily, the disruptor enhances the quality of their offering, and the incumbent cedes market share in segment after segment to the disruptor.

Disruption theory predicts the emergence and evolution of DeepSeek and its ilk. In fact, it wouldn't be surprising if other disruptors emerge over the next few months. In particular, small language models (SLMs), which utilize less data and fewer resources and yield lower quality content, could be yet another technology that challenges both American and Chinese LLMs in the months to come.

Where Do We Go from Here?

The emergence of DeepSeek raises a question for boardrooms across the globe: Should companies invest in licensing American LLMs or Chinese LLMs? Or both? Here, too, prior management insights — especially around navigating technological diversification — come in handy.

A benefit of having multiple LLM models deployed within an organization is diversification of risk. With LLMs, this translates to mitigating the effects of downtime at the provider end. For instance, if OpenAI service were to be affected for some reason, the business can continue to function using another provider's model.

Another benefit of using multiple models comes from the benefits of aggregation. Different models use different algorithms and thus provide different answers to the same question. Studies have found that aggregating across multiple models and multiple sources of predictions — an approach that researchers have termed [“ensembling”](#) — often yields better quality outcomes, particularly with complex, ambiguous tasks. Indeed, platforms like Openrouter, a recently-founded U.S.-based AI model aggregator, already offer an integrated interface that allows users to compare the performance and cost of more than 180 models in real time for a small fee.

On the other hand, a benefit of working with a single supplier is reduced administrative costs and better understanding of capabilities on both sides of the partnership. Using multiple models increases data privacy and security risks, as data might have to be shared with multiple providers. Although many of these [concerns](#) pervade all LLMs, including U.S. ones, data access and use data across countries — say, between U.S. and China — each with its own regulatory framework, will add another layer of complexity. This can be particularly problematic, especially in sensitive in applications such as healthcare.

Prior management theories on technological change and diversification also suggest a third possibility beyond single- or multi-sourcing: [plural governance](#). Plural governance involves using a combination of external suppliers and internal developers to leverage an emerging technology. In fact, prior research in economics has long argued that companies that internally develop [vintage-specific human capital](#) are most likely to benefit from the emergence of new technologies. In the case of language models this might entail using American LLMs for general-purpose tasks (such as developing a bot that aids research for consultants or lawyers at a professional services firm) and leveraging Chinese LLMs for company-specific tasks (such as an HR training bot that helps onboard new workers).

Going further, a lower-cost, open-source LLM model with smaller training data requirements, even one with lesser capabilities than a closed-source one, will allow companies to develop company-specific models suited to their context. Over time, however, these lower-cost and lower-quality models will likely disrupt the higher-cost models, just like mini-mills disrupted integrated steel plants across all market segments.

Even with data privacy and security concerns — and notwithstanding [the recent TikTok episode](#) — American LLMs will ignore the Chinese LLM disruption threat at their own peril. If nothing else, they should fear the emergence of American disruptors that use SLMs, among other approaches. Large American AI companies could also attempt to disrupt themselves (e.g., GE developed its own hand-held ultrasound device to disrupt the more expensive ultrasound business), though research suggests that [self-disruption](#) is incredibly hard. In particular, the sunk-cost fallacy related to prior investments in expensive chips, hardware and training data (which are partly sunk costs at this point) and incentives to sell high-margin solutions might tether most American AI companies to their high-end LLMs rather than investing in cheaper but “good enough” LLMs.

For global companies using LLMs, disruption in the LLM space opens the gates to investing in internal skills and developing company-specific models that might lead to more targeted use cases, lower costs, and higher ROI.

PC

Prithwiraj (Raj) Choudhury is Lumry Family Associate Professor at the Harvard Business School and Associate Editor at Management Science. He studies the Future of Work and is a Forbes Future of Work-50 and TIME-Charter Future of Work-30 awardee.

NB

Natarajan Balasubramanian is the Albert & Betty Hill Endowed Professor at the Whitman School of Management at Syracuse University. He studies how technology, human capital, organizational learning, and innovation contribute to business value creation.

MX

Mingtao Xu is an Associate Professor in the Department of Innovation, Entrepreneurship, and Strategy at Tsinghua University School of Economics and Management. His research focuses on property rights in innovation, as well as the strategic implications of Artificial Intelligence (AI).