

Indian companies look to Chinese LLMs as AI costs bite

asia.nikkei.com/business/technology/artificial-intelligence/indian-companies-look-to-chinese-llms-as-ai-costs-bite

SAYAN CHAKRABORTY

July 13, 2026



Indian companies are turning to Chinese artificial intelligence models amid cost pressures, raising questions about the South Asian nation's ambitions around AI sovereignty. (Nikkei montage/Source photos by Reuters)

BENGALURU -- Indian companies are increasingly leaning on Chinese large language models (LLMs) developed by DeepSeek, Alibaba and Moonshot AI to contain their artificial intelligence spends, extending India's reliance on China for cutting-edge technologies despite a long history of standoffs between the neighbors.

Puneet Kumar, CEO at Mirae Asset Venture Investments India, said that several consumer technology startups that he has met since mid-2025 use such Chinese open-weight LLMs -- those that rely on publicly accessible parameters -- which help drive down costs by an "order of magnitude."

Because these parameters are publicly available, users can download and modify them on their own computers, in contrast to the proprietary offerings by U.S. frontier labs such as OpenAI and Anthropic. The Chinese open-weight LLMs can be accessed in India through service providers such as Microsoft at a fraction of the price of their American counterparts, thanks to their low cost of development.

"The U.S. models are expensive, and for a lot of basic things, you don't need them," Gupta said. "It's overkill, like trying to drive a sports car on a crowded city road."

For the DeepSeek models that Microsoft makes available in southern India through its Foundry platform, charges range between 19 cents and \$1.74 per million input tokens -- the tiny units of data that AI models process -- while the price per 1 million output tokens varies from 51 cents to \$5.40. Input costs for Moonshot's Kimi go up to 95 cents and output costs up to \$4.

In comparison, input costs for OpenAI's GPT 5.5 series range from \$5 to \$12 per million input tokens, while output costs hover between \$30 and \$54.

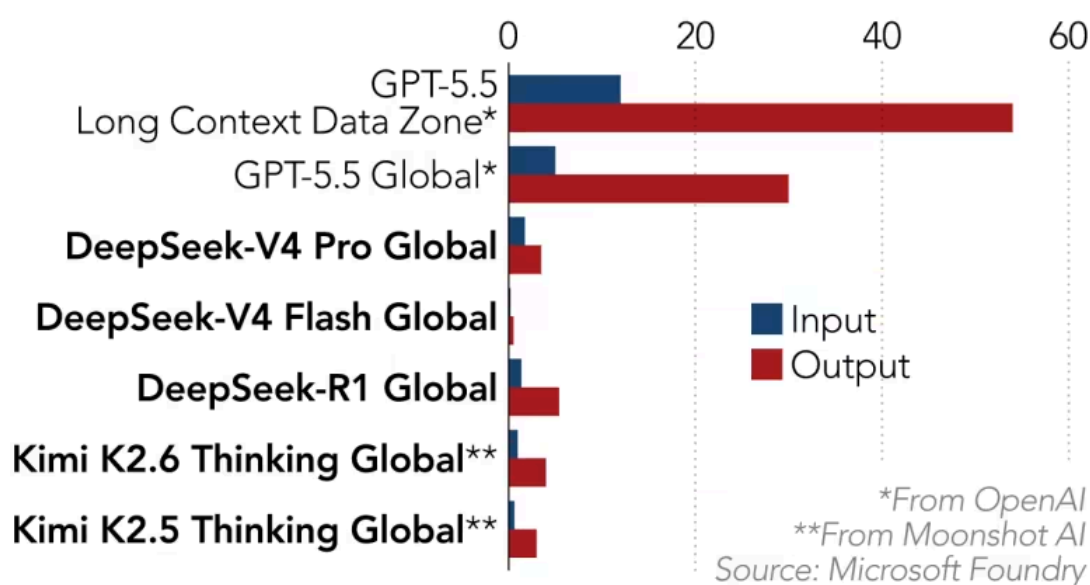
The adoption of Chinese LLMs comes at a time when the likes of OpenAI and Anthropic are opening offices and expanding their offerings in the South Asian nation. Both OpenAI CEO Sam Altman and Anthropic CEO Dario Amodei showed up at the India AI Summit earlier this year, underscoring the importance of India as a market, thanks to its large number of internet users. Anthropic has bagged Tata Group-owned airline Air India, software maker Cognizant and payment processor Razorpay as customers, while OpenAI counts India's largest software maker, Tata Consultancy Services, as a client.

The cost pressures driving the soaring popularity of Chinese LLMs in India are mirrored globally, with major companies including Tesla, Amazon, Uber and Walmart capping AI usage to arrest soaring technology spends as focus shifts from indiscriminate usage, called tokenmaxxing, to return on investment. Chinese LLMs are [emerging as winners](#) with aggressive pricing, with their usage more than doubling to 25 trillion tokens in the final week of June from the end of May, according to Open Router, a marketplace for AI models. That was 78% more than U.S. models, a sharp reversal in fortunes from the start of the year, when usage was less than half that of their American counterparts.

Companies like Coinbase, DoorDash and Airbnb have publicly said that they have begun using Chinese models.

Chinese AI models are much cheaper in India than U.S. models

(Costs per million tokens, in dollars; for southern India)



Vidya Madhavan, founder of Elevation Capital-backed dating app Schmooze, said that possibilities of "substantial" savings, coupled with wider uptake of open-weight LLMs globally and their ability to deliver a satisfactory performance in comparison with their American peers, encouraged her to deploy Alibaba's Qwen models after some initial hesitation.

"Our approach has been to use solutions from Google, OpenAI, Anthropic, ElevenLabs, etc., to start with, so that we have a sense of what great looks like, and then use a combination of open source plus our tuning to achieve the same outcome and save money wherever applicable," she said.

Nikhil Narendran, a partner at law firm Trilegal, sees startups and developers as early adopters in India, though larger firms are also deliberating over the deployment of Chinese LLMs.

"The token bills are a serious issue -- it is increasingly becoming unsustainable," he said.

Adding to the appeal of open-weight LLMs is that, by virtue of being locally hosted, data stays in India, Narendran said, although "there might be unverified deployment artifacts such as malware or trojans, which is a concern."

"Unless the token-intensive nature of the U.S. frontier AI models changes, the Chinese are likely to take a significant lead over the Americans," he said. "Since it's mainly the trust factor that goes against them, I am sure Chinese developers understand that risk, and hence are likely to be extra careful."

The incursion of Chinese models into India is raising questions about the South Asian nation's ambitions around AI sovereignty, even as companies such as Sarvam and Gnani build LLMs in [Indic languages](#).

In 2024, India earmarked about 104 billion rupees (about \$1.1 billion at current exchange rates) for the technology over a five-year period, but this pales in comparison to China's public investments "running into tens of billions of dollars annually," estimates research firm Bernstein. Moreover, disbursements have been patchy, with actual spending in the fiscal year ending March 2025 totaling 190 million rupees, against an allocation of 5.52 billion rupees, while expenditure for the following fiscal year stood at 3.79 billion rupees, versus an allocation of 20 billion rupees. Government officials have said that disbursements will increase in sync with use of graphics processing units, which have taken time to procure but are crucial to developing AI models.

The risk of being cut off from foreign AI models amid a fragile geopolitical situation has increased the pressure on India to develop its own capabilities. Washington has already prevented Anthropic from giving foreign entities access to its latest models, Fable and Mythos, though the restrictions were lifted late last month. In addition, Reuters reported earlier this week that Beijing is now considering restricting overseas access to advanced Chinese AI models.

Analysts warned that India's dependence on China -- which is also evident in other advanced technologies like [electric vehicles](#) and lithium-ion cells -- is risky, given the two nations' recent history of tensions, particularly over a territorial dispute in the Himalayas. In 2020, New Delhi tightened restrictions on Chinese investments after an outbreak of fighting there, although these were eased in March this year.

Sameer Patil, director at the Centre for Security, Strategy and Technology at the Observer Research Foundation think tank, said India's ambitions with sovereign LLMs are restricted to domestic usage, unlike the U.S. and China, which are eyeing global dominance.

"Deepening the dependence on the foreign tech, whether American or Chinese, is a concern because access can be shut down overnight and you will be left in the lurch," Patil said. "Therefore, we have to develop that kind of resilience."