

 Handcrafted By **Our Editors**

Claude and ChatGPT too expensive, Chinese AI models surge in use due to low cost

As AI gets more expensive, popularity of Chinese AI models like DeepSeek is growing fast, even though these models are not as good as ChatGPT, Claude and Gemini. Industry experts believe the cost pressure is real and most companies will move to a hybrid model for AI use.



Companies are increasingly choosing good enough Chinese models to save costs. (Representational image made with AI by Divya Bhati)

In the world of tech, at least outside China, we do not hear that much about Kimi, MiniMax and Qwen. Instead, the buzz is mostly around ChatGPT, Gemini and Claude. As in Claude this and Claude that. Yet, in the first half of 2026, a surprising trend is underway. The popularity of Chinese AI models like Kimi, DeepSeek, Qwen and MiniMax [seem to be surging across the world](#). And it is not because these are the best AI models. But it is because, as an Industry expert puts it, they are "good enough."

With AI models like OpenAI Codex and Claude Code fast becoming [prohibitively expensive even for tech giants like Uber and Microsoft](#), companies across the world are trying to get their AI fix through Chinese models that are significantly cheaper. The idea is that even though the Chinese models are not exactly in Claude-class, they are also not that far behind. At least for mundane tasks, such as customer service, they seem to be more than serviceable. This is leading organisations, according to AI company Sekond Brain founder Sachin Dev Duggal, to ask "why are we using a Ferrari to go to the supermarket?"

At least data from the early months of 2026 is clear. As OpenAI, Google and Anthropic switch to token-based pricing, which is extremely expensive, usage of Kimi and DeepSeek has gone up significantly.

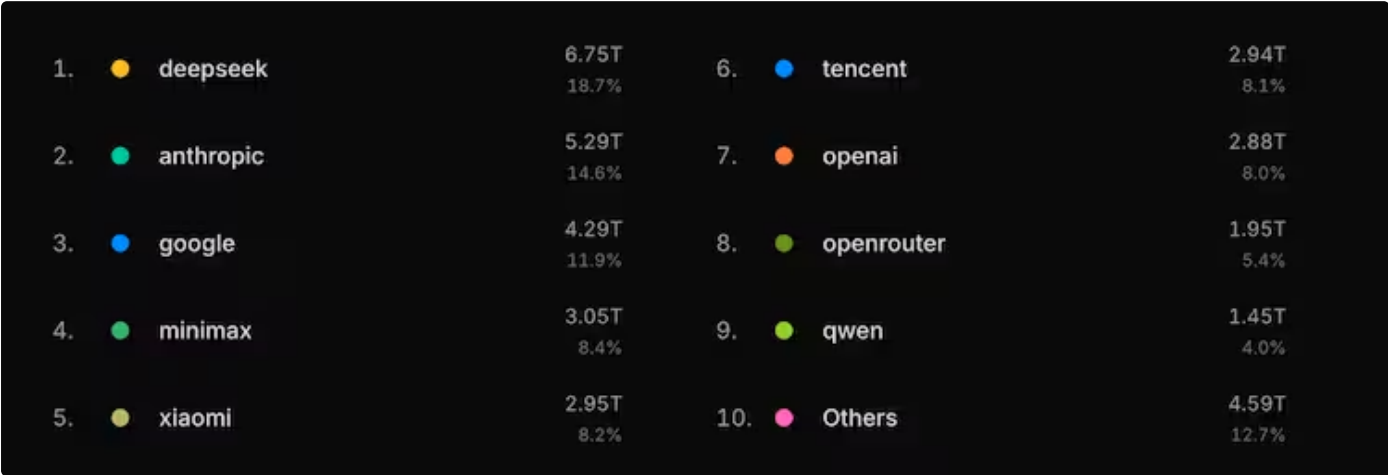
According to the latest data from OpenRouter, a service that allows organisations to hook to APIs of popular AI models, Chinese AI providers have found favour among users this year. The data suggests that DeepSeek now has a share of around 17.6 per cent in AI use through OpenRouter. The data also reveals that in February this year, Chinese AI models reportedly delivered 4.12 trillion tokens against 2.94 trillion from American models. Overall, it seems that the Chinese models have gone from roughly 1 per cent of developer usage in 2024 to over 60 per cent by May 2026.

1.	 DeepSeek V4 Flash by deepseek	4.23T tokens ↑55%	11.	 Gemini 3 Flash Preview by google	1.06T tokens ↑9%
2.	 Hy3 preview by tencent	3.53T tokens ↑25%	12.	 Step 3.7 Flash by stepfun	855B tokens ↑195%
3.	 MiniMax M3 by minimax	3.22T tokens ↑>999%	13.	 Gemini 2.5 Flash by google	628B tokens ↑8%
4.	 MiMo - V2.5 by xiaomi	2.73T tokens ↑40%	14.	 Laguna M.1 (free) by poolside	599B tokens ↑6%
5.	 DeepSeek V4 Pro by deepseek	2.03T tokens ↑59%	15.	 Gemini 2.5 Flash Lite by google	594B tokens ↑3%
6.	 Owl Alpha by openrouter	2.01T tokens ↑9%	16.	 Nemotron 3 Ultra (free) by nvidia new	534B tokens
7.	 Claude Sonnet 4.6 by anthropic	1.96T tokens ↑5%	17.	 Gemini 3.5 Flash by google	472B tokens ↓10%
8.	 Claude Opus 4.7 by anthropic	1.61T tokens ↓4%	18.	 MiMo - V2.5 - Pro by xiaomi	471B tokens ↓65%
9.	 Claude Opus 4.8 by anthropic	1.33T tokens ↑140%	19.	 Nemotron 3 Super (free) by nvidia	466B tokens ↓29%
10.	 DeepSeek V3.2 by deepseek	1.18T tokens ↑7%	20.	 GPT-5.5 by openai	456B tokens ↑2%

In terms of output token, the leaderboard at OpenRouter is dominated by Chinese AI models.

The reasons are easy to see even though the Chinese AI models lag behind the American AI models in capability and refinement in highly complex tasks. Models like ChatGPT 5.5 and Claude Opus 4.8 are miles ahead of their Chinese counterparts in terms of the quality of their output. But this quality may not be needed in most instances and "good enough" models can do a lot of mundane work at a relatively low cost.

A glance at the pricing of different AI models highlights the gap between the Chinese and American AI. Chinese models like MiniMax and Moonshot charge around \$2–\$3 per million output tokens. DeepSeek, with its latest Pro model, is even cheaper with a pricing of less than \$1 for one million output tokens. On the other hand, Anthropic's Claude Sonnet runs at about \$15, while the latest Opus 4.8 can cost as much as \$25. The cost of the latest ChatGPT and Gemini is slightly lower compared to Claude, but compared to prices of Chinese models it is still significantly higher.



On June 1, out of 10 top AI models, five were from Chinese companies. DeepSeek was No 1 in terms of usage. Source: OpenRouter

Cost is one factor that has made SimplifyGenAI, a global firm that helps companies integrate custom AI and automation tools, adopt Chinese models. "We never used (OpenAI) Sora at all," Daksh Sharma, co-founder of SimplifyGenAI told India Today Tech. "We now rely on Chinese models like Seedance and Kling, which deliver outstanding output at a fraction of the cost."

Sharma says he is seeing a similar pattern emerge in the LLM market as well. "For most commercial use cases, good enough at a fraction of the price simply wins," he says.

Companies will explore hybrid AI mode

Industry experts believe the surge in Chinese AI use is the beginning of a trend where companies will use AI while keeping an eye on their usage pattern, benefit and cost. It does not mean an outright win for Chinese models but instead will force everyone in the industry, from users to AI firms creating and selling models, to become more aware of the economics of AI. And this change, industry experts suggest, will lead to a hybrid mode of AI use.

Sachin Dev Duggal, following his Ferrari and supermarket example, suggests that currently supermarket trips make up the vast majority of AI workloads at most companies. "For classification, summarisation, first-draft, internal automation, repetitive support or high-volume outbound, it is very hard to justify using the most expensive frontier model," he says. "In future, cheaper models will do commodity cognition. Frontier models will do high-stakes reasoning."

In other words, Duggal says, in future, both Chinese AI models and top-of-the-line US AI models like Claude will co-exist within enterprises, with different kinds of work handled by different models.

Keshava Murthy, co-founder and CEO of Matters.AI, believes the hybrid model will be the way forward. "Affordable models may power internal copilots, support automation, or lightweight workflows, but mission-critical systems involving proprietary data, code generation, or autonomous agents will continue to rely on trusted frontier models," he says. "The enterprise AI future will be hybrid, not winner-takes-all."

The initial days of AI have been marked by experimentation and pushing of boundaries of what is possible. But now that companies are done with their exploration and experimentation, they are looking to cut costs, ironically at the same time when US AI models are increasing their prices.

Anurag Sahay, CTO and global head of AI and data science at Nagarro, explains the coexistence of Chinese models and the frontier AI models from the US using a train analogy. "The AI model landscape runs like a train. Frontier models lead at the front – capable, expensive, and pulling the whole industry forward. Behind them, a growing line of smaller, faster, cheaper models stretch further every month," he says. "The real insight is that enterprises don't have to pick sides. Use frontier models to build what's new. Use leaner models to run what's known."

Don't discount the trust factor

Of course, these are early days in the world of AI. The US AI models and the Chinese AI models currently seem to be chasing different goals. The likes of ChatGPT and Claude are trying to race to AGI aka superintelligence as fast as possible with their proprietary approach, irrespective of the cost. They are hoping to take stock of the economics once they have reached this superintelligence. The Chinese AI companies, hampered by export controls that the US has imposed on AI chips, are going for a good-enough approach and an open-source model to increase their adoption among users and companies.

Then there are factors like trust, privacy, data security and transparency, all of which are governed by the underlying political and legal framework under which AI companies are operating. In this aspect, at least in terms of perception, the US AI models seem to have an edge. And that too is a factor in a company choosing one AI model over another.

"When a company chooses an AI model for production deployment, cost is one variable among several. Data privacy, model reliability, regulatory compliance, audit-ability and vendor accountability all sit on the same table," says Abhishek Agarwal, president at the Judge Group, an IT services company.

Govind Rammurthy, CEO and managing director at eScan, shares a similar view. "The market will get divided: companies doing commodity AI tasks will migrate to cheaper alternatives. Companies where AI mistakes carry real costs will pay for superior models," he says.

On social media, and in the tech world, the conversation is currently framed as DeepSeek versus Claude, or China versus the US. But the industry reality is that the future is more likely to be one of coexistence rather than replacement. The experts suggest the future of AI may not belong to the cheapest model or the smartest model. It will belong to organisations that learn how to combine both for the best possible outcome.

This is how Chitranshu Mahant, co-founder and CEO of Primebook, frames it: "The objective is rarely to use the cheapest model or the most expensive one, but to achieve the best outcome for a given use case while maintaining efficiency, reliability and scalability."

- Ends ■