DeepSeek Business Model

Idea In Short

DeepSeek, a Hangzhou-based Artificial Intelligence [AI] lab founded in 2023, built large language models [LLMs] that match the performance of systems from OpenAI and Google at a fraction of the reported training cost, then released the model weights for free. Its parent, quantitative hedge fund High-Flyer, funds the lab entirely, which removes the pressure to chase near-term revenue that venture-backed rivals face. That financial cushion lets DeepSeek treat open releases as a distribution channel rather than a cost center, building a global developer base through free downloads while charging for API access, reasoning-heavy inference and enterprise deployments. In January 2025, DeepSeek's chatbot app overtook ChatGPT on the U.S. iOS App Store, a milestone that erased roughly $600 billion from Nvidia's market value in a single session and pushed every major AI lab to defend its own spending on compute.

-

From quant trading to AI research

Liang Wenfeng graduated from Zhejiang University and started trading financial markets during the 2008 financial crisis. He and his classmates tested early machine learning models, years before deep learning became common in quantitative finance. He built on that experimentation and co-founded High-Flyer in February 2016. The hedge fund used AI-driven algorithms to trade stocks and grew into one of China's largest quantitative funds.1 By the end of 2017, the fund had shifted nearly all of its trading to AI-driven models. It also built its own GPU computing clusters, years before most AI labs existed in China. That infrastructure work gave Liang a team fluent in large-scale computing long before DeepSeek needed one.

A hedge fund's calculated bet

In 2021, most of the trading industry focused on returns. Liang instead began buying

thousands of Nvidia graphics processing units [GPUs] for an AI project unrelated to High-Flyer's financial business, a move that puzzled competitors at the time. That stockpile reportedly included 10,000 Nvidia A100 chips, acquired before Washington restricted chip sales to China. It became the computing backbone for an AI research lab that High-Flyer announced in April 2023 and spun off as an independent company three months later. Liang personally holds an 84% stake in DeepSeek through two holding entities, and High-Flyer remains its sole financial backer.2 That structure frees the lab from outside investors demanding a fast path to profit.

The open-weight distribution strategy

DeepSeek's central strategic choice is publishing its model weights instead of keeping them proprietary. Since the DeepSeek-R1 release in January 2025, most new models have shipped under the MIT License. That permissive term allows free commercial use, modification and redistribution. The company disclosed that the pretraining run behind R1 cost roughly $5.6 million in GPU time, far below the hundreds of millions typically associated with training comparable Western systems.3 That figure covers only part of total development spending. Still, it captures a strategy built on releasing capable, freely downloadable models to build adoption before monetizing the infrastructure around them.

Turning export controls into an engineering edge

U.S. restrictions on advanced AI chips forced DeepSeek to train its models on Nvidia H800 processors. That is a less powerful variant built specifically for the Chinese market, after Washington banned sales of the more capable H100 and A100 chips. DeepSeek's engineers treated the constraint as a design problem rather than a ceiling. They rebuilt core software layers, including custom communication libraries and a distributed file system, to extract more throughput from a smaller GPU fleet. Mixture-of-experts architecture reinforced that efficiency push by activating only a subset of a model's parameters for each task, cutting both training and inference costs. Congressional inquiries and reporting since have questioned whether DeepSeek later gained indirect access to more advanced Nvidia chips despite the restrictions, a claim the company disputes.4

Revenue beyond the free chatbot

DeepSeek layers several revenue sources on top of its free consumer app. Developers and

businesses pay per token to access DeepSeek-V3 for general chat tasks, and DeepSeek-R1 for complex reasoning, through the company's API. Rates undercut most Western competitors. Enterprises can also commission customized deployments suited to specific workloads, adding a services layer on top of the API business. Independent estimates of DeepSeek's total hardware spending run as high as $500 million, once the computing clusters built years before its public model releases are factored in.5 That gap suggests the disclosed training costs cover only a slice of what the company has actually invested in compute.

Who actually uses DeepSeek

Three groups anchor DeepSeek's customer base. Individual users download the free chatbot app on iOS and Android for everyday tasks such as research, writing and general problem-solving. Developers and businesses integrate the API into products that need natural language processing, customer support automation or content generation. The per-token cost usually runs lower than comparable U.S. services. Academic researchers and open-source contributors use the published model weights to study, fine-tune and build on DeepSeek's architecture without paying licensing fees. That access has made the company a reference point in AI research well beyond China. Semiconductor makers including Huawei and Cambricon have also adopted DeepSeek's models to run on their own chips, extending the company's reach into hardware ecosystems built to compete with Nvidia.

Competing with OpenAI, Google and Chinese rivals

DeepSeek operates in a field crowded with better-funded competitors. OpenAI's ChatGPT and Google DeepMind's models still lead on many benchmark comparisons and enterprise trust, backed by far larger balance sheets and cloud infrastructure. Anthropic occupies the safety-focused end of the market with its Claude models. Domestic rivals Baidu, SenseTime and Tencent compete for the same Chinese enterprise and consumer customers DeepSeek targets. DeepSeek's edge is price and openness, not raw capability across every benchmark. It offers comparable performance on many tasks at a lower cost per token, and without the licensing restrictions that come with closed models. That combination is enough to win over developers who are cost-sensitive or want to run models on their own infrastructure.

Strengths, and the risks that come with them

DeepSeek's cost efficiency is a genuine structural advantage. Matching the performance of far more expensive systems while spending less on compute changes what a frontier AI lab needs to raise before it can compete. The chatbot's rapid rise offers a clear illustration. In January 2025, the app climbed past ChatGPT to the top of the U.S. iOS App Store's free chart within days of its release. The jump coincided with an 18% single-day drop in Nvidia's share price, as investors reassessed how much compute demand AI actually requires.6 That same visibility exposed the company's weaknesses. Regulators and researchers in several countries have raised concerns about DeepSeek routing user data to servers in China, and about the chatbot declining to answer or redirecting questions on topics the Chinese government treats as sensitive. Both limit its appeal in markets that weigh data sovereignty and free expression heavily.

What DeepSeek signals for the AI industry

DeepSeek's trajectory tests an assumption that had gone largely unchallenged since the release of GPT-4: that frontier AI requires enormous capital before it can compete. A lab funded by a single hedge fund, working with export-restricted chips, produced models that trade blows with systems built by companies valued in the hundreds of billions of dollars. That outcome does not eliminate the advantages of scale. It does show that engineering efficiency can substitute for some of the capital Western labs have treated as a prerequisite. For consultants and executives advising on AI strategy, DeepSeek's model is a reminder that compute budgets and model quality do not move in lockstep. A well-funded internal research group with patient backing can close a gap that looked insurmountable a few years earlier.

- 1Meet DeepSeek Founder Liang Wenfeng, A Hedge Fund Manager
- 2DeepSeek
- 3In Rare Disclosure, DeepSeek Claims R1 Model Training Cost Just $294K
- 4Nvidia Worked To Co-Design DeepSeek Model, US Lawmaker Says
- 5DeepSeek's Hardware Spend Could Be As High As $500 Million, New Report Estimates
- 6DeepSeek Hit By Cyberattack As Users Flock To Chinese AI Startup

Summary

DeepSeek's rise rests on a financing structure few AI labs can replicate. A profitable hedge fund covers the bills, so the company doesn't need to chase revenue the way venture-backed labs do. That patience, combined with real gains in training efficiency, let DeepSeek ship models that rival GPT-4-class systems while spending far less than OpenAI or Google reportedly spend on comparable work. The open-weight approach turns that cost advantage into distribution: developers who download DeepSeek's models for free become a funnel toward paid API usage and custom enterprise deployments. Whether this model keeps working is an open question. Export controls, data-privacy scrutiny in Western markets and rivals copying the same efficiency techniques could narrow DeepSeek's advantage over time. For now, the company has forced competitors across the industry to rethink what frontier AI actually costs to build.