China Internet & New Media

EQUITY: INTERNET & NEW MEDIA

Takeaways from an LLM expert call

Open source is not an open playbook

Research Analysts

China Internet & New Media

Jialong Shi - NIHK
Jialong.shi@nomura.com
+852 2252 1409

Rachel Guo - NIHK
rachel.guo@nomura.com
+852 2252 1400

Key highlights

The Nomura China Internet team held a group call with an expert from a Chinese AI lab affiliated with a research institution. The lab currently has its proprietary foundation models deployed across more than 100 enterprise customers.

The expert noted that persistent supply tightness and price inflation for high-end Nvidia (NVDA US, Not rated) chips are making domestic AI accelerators increasingly attractive. While Chinese model providers are broadly reducing token prices, the expert highlighted a clear bifurcation between basic and advanced models. Providers are aggressively cutting prices for basic models – which process commoditized tasks – to acquire customers, while maintaining premium pricing for their advanced models. The expert thinks that model providers can enhance enterprise customer loyalty by embedding their models deeply into corporate workflows. This integration makes switching models highly time-consuming and financially burdensome, effectively deterring customer churn.

The expert believes DeepSeek (unlisted) possesses a distinct cost advantage driven by superior operational efficiency during inference workloads. This efficiency is highly difficult for other platforms hosting the same open-source DeepSeek model to replicate. The expert views financial services, office productivity, coding, manufacturing, legal, and healthcare as the most likely primary adopters of large language models (LLMs). Consequently, these sectors are expected to be the core future revenue contributors for LLM providers.

Hardware constraints: chip supplies remain tight

The expert's firm was an early, large-scale adopter of Huawei (unlisted) Ascend accelerators and began adapting domestic hardware and software years before the recent surge in demand for Chinese AI chips. The expert observed that prices for both Nvidia and domestic accelerators have increased due to robust demand and a tight supply of high-end training hardware. Notably, pricing for certain premium Nvidia chips has risen sharply over the past year. While domestic accelerators have also experienced price hikes, the inflation has been less pronounced. Consequently, domestic hardware offers increasingly attractive economics for inference workloads.

DeepSeek: open-source does not equate to an open playbook

DeepSeek delivers materially lower inference costs compared to many models with similar capabilities. Importantly, the expert argued that competitors cannot easily replicate this cost advantage, even when DeepSeek open-sources its model weights and segments of its software stack.

The expert attributes DeepSeek's edge to superior system-level optimization, including higher cache efficiency, lower latency, and better hardware utilization. Although much of DeepSeek's model architecture is public, critical implementation details and operational know-how remain proprietary. Consequently, when cloud providers such as Tencent (700 HK, Buy), Alibaba (BABA US, Buy), and ByteDance (unlisted) deploy the same open-source DeepSeek model on comparable hardware, DeepSeek's native deployment achieves higher operational efficiency and lower average inference costs. This cost advantage gives DeepSeek greater flexibility to undercut third-party platforms on official API token pricing.

We note similar feedback from other industry sources. While most open-source models are free to deploy—and some permit free commercial use—it remains challenging for users to match the operational performance of the original developer. Furthermore, there

Production Complete: 2026-07-13 08:00 UTC

is no guarantee that Chinese AI labs will continue to open-source their frontier models. For instance, Alibaba was an early champion of the open-source strategy, but its latest flagship model, Qwen 3.6 Max, was launched as a closed-source product. As *a recent expert from Moonshot AI* noted, open-source models are effective for building brand awareness and developer ecosystems, but closed-source remains the primary avenue for monetizing proprietary models at scale.

The bifurcation of LLM pricing

Despite rising infrastructure costs, the expert expects pricing for lower-tier models to continue declining. Vendors leverage these models as loss leaders to acquire developers and enterprise clients, particularly for general-purpose workloads with low switching costs. Because clients can easily migrate basic API workloads to competing models, vendors lack pricing power when their sole differentiator is token cost.

Conversely, the pricing outlook for advanced models serving large enterprises is more robust. These clients prioritize model reliability, task-completion quality, latency, and service stability. Once a model is deeply embedded into a production workflow, the switching costs become prohibitively high. For example, the expert's institute continues to reduce prices for basic models to drive customer acquisition, while simultaneously increasing pricing for high-performance models and customized services tailored to strategic clients.

Ultimately, falling headline token prices do not indicate industrywide deflation. Model providers are well positioned to offset commoditized workload discounts by charging premium rates for advanced, customized models that are integral to enterprise operations.

Enterprise adoption: security vs. ROI

The expert segmented enterprise AI customers into two distinct cohorts:
Government and SOEs: Discussions prioritize data security, regulatory compliance, and localized deployment. This cohort strongly prefers models deployed on-premise using domestically produced hardware. While state-backed vendors currently have a customer acquisition advantage here, the expert noted that SOEs are beginning to deploy superior models from private vendors for less sensitive workflows.

Private enterprises: ROI is the primary catalyst. Rather than accepting vague promises of long-term productivity gains, private clients expect clear visibility on recovering their initial AI investments within a 12- to 18-month timeframe.

High-potential commercialization verticals

The expert identified several industries with substantial commercialization potential:

Financial services: Viewed as one of the most attractive verticals. Financial institutions possess vast proprietary datasets and demonstrate a relatively high willingness to pay for premium AI tools.

Office productivity: Applications span document generation, meeting management, and seamless integration with ERP and CRM systems. Future enterprise agents are expected to execute routine operational processes across multiple software platforms autonomously.

Coding: Highlighted as a strategically vital market. Coding capabilities serve as the foundational layer for broader agentic AI, robotics, and automated operations. The expert views GLM-5.2 as the leading coding model in China, followed by the latest iterations from DeepSeek, Moonshot (unlisted), and Alibaba's Qwen.

Industrial manufacturing: Currently in early-stage adoption (estimated at under 10% workflow penetration). As models improve their multimodal capabilities to interpret physical processes, interaction with industrial equipment is expected to accelerate.

Healthcare and legal: Both sectors represent significant long-term growth opportunities for foundation models.

