



China Internet & New Media

Global Markets Research

6 July 2026

EQUITY: INTERNET & NEW MEDIA

Takeaways from Moonshot AI expert call

Commercial acceleration and strategic pivot

Key highlights

The Nomura China internet team recently hosted a group call with an expert from Moonshot AI (unlisted), a leading Chinese generative AI startup. Its annualized recurring revenue (ARR) is projected by management to surpass USD1bn by year-end, driven primarily by overseas API usage and robust AI coding capabilities. Beyond AI coding, the expert believes that the next growth frontiers will be AI solutions for healthcare, education, and finance industries, where high-quality proprietary data will be a critical differentiator aside from specialised AI capabilities. To protect margins and improve monetization, Moonshot is shifting toward a closed-source strategy for its flagship models, while utilizing domestic chips to lower inference costs. While experts suggest pure-play AI labs like Moonshot could outcompete large internet platforms by maintaining pure focus, we continue to believe that major internet platforms such as Alibaba (BABA US, Buy) and ByteDance (unlisted) remain formidable competitors due to their vast capital, data ecosystems, and strategic need to own state-of-the-art (SOTA) LLMs (large language models).

Who is Moonshot AI

Moonshot AI (Moonshot) is a Beijing-based AI startup focused on LLMs and AI assistant products. Founded in early 2023 by Dr. Yang Zhilin—a former Google Brain and Meta/FAIR engineer – alongside a group of his Tsinghua University alumni, the company has rapidly established a prominent position in China's generative AI landscape.

Moonshot's product and model ecosystem is built around the Kimi assistant and K-series foundation models. Its flagship foundation model, Kimi K2, is a trillion-parameter model featuring advanced coding and agentic capabilities. Subsequent iterations include Kimi K2 Thinking, Kimi K2.5, Kimi K2.6, and Kimi K2.7 Code. Moonshot has also expanded into specialized AI agent products, notably Kimi Code, Kimi Agent/Agent Swarm, and Kimi Work, with the latter positioned as a dedicated desktop AI agent for office workers.

The company is backed by a premier group of Chinese technology corporations and venture capital investors, including Alibaba, Tencent, HongShan, ZhenFund, and IDG Capital. *Chinese media* reported that Moonshot completed its latest funding round in May 2026, raising USD2bn at a post-money valuation of USD20bn. This round was led by Meituan, with participation from China Mobile, Tsinghua Capital, and other venture capital investors.

ARR likely to surpass USD1bn by the year-end

According to the expert, Moonshot's annualized ARR reached USD400-500mn by mid-2026, with monthly ARR rising 10-20% sequentially. Growth was particularly robust early in the year, expanding 2-2.5x y-y, supported by new product launches and accelerating model adoption driven by the hype around the desktop agent, OpenClaw. The expert projects that Moonshot's ARR could exceed USD1bn by the year-end.

Revenue composition is shifting materially toward API calls. According to the expert, API usage has become the primary revenue driver, followed by consumer-oriented subscriptions. Private deployment, including licensing and maintenance fees, contributes a relatively small share and has been declining as a proportion of revenue.

The expert attributed API growth primarily to continuous model iteration, increased usage by overseas developers, and the release of useful vertical capabilities such as coding. Subscription revenue is also growing, but at a slower pace than API revenue.

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

Production Complete: 2026-07-06 00:47 UTC

Geographically, the revenue mix has increasingly pivoted toward international markets. Overseas revenue contribution has risen steadily since the beginning of the year and now represents a larger share of total revenue than domestic operations. Sustained future growth remains heavily dependent on the continued expansion of overseas API calls, international subscriptions, model upgrades, and the deployment of specialized vertical capabilities, according to the expert.

The next major product catalyst is expected to be K3, which the expert suggests could be released around September 2026. However, the launch window remains flexible and may be adjusted depending on the release schedules of competitors such as MiniMax, Zhipu, DeepSeek, and Alibaba's Qwen.

K3's key areas of improvement are expected to include multimodal capability, enhanced coding performance, agent task orchestration, and enriched vertical capabilities, according to the expert.

In coding, the expert views GLM and Kimi as among the leading domestic models. DeepSeek and Qwen are also viewed by the expert as strong competitors.

Healthcare, education and finance likely to become the next battlefields for AI agents, after coding

AI coding currently represents one of the clearest commercialization paths for LLM developers, a trend heavily validated by the rapid monetization of overseas peers like Anthropic's Claude Code. However, the expert thinks that the long-term total addressable market (TAM) for AI coding may face constraints. The core addressable user base requiring sophisticated, high-frequency coding assistance is relatively price-insensitive but modest in absolute size. Further, international players like Anthropic and OpenAI have already captured a meaningful portion of high-value global demand, according to the expert.

Consequently, the expert believes that the next phase of enterprise competition is expected to shift toward larger vertical markets, specifically healthcare, education, and finance. Healthcare represents the most immediate opportunity due to the acute demand for AI capabilities in medical consultations, diagnostic assistance, imaging interpretation, automated report generation, complex case discussions, and pharmaceutical R&D. Both education and finance sectors present abundant use cases, though AI penetration in finance will likely lag education due to stringent regulatory and compliance constraints, according to the expert.

Value of high-quality data elevates for vertical AI agents

According to the expert, securing high-quality vertical data remains a key prerequisite for sharpening vertical model capabilities. Because this data is rarely accessible in the public domain due to commercial sensitivity and privacy regulations, internet platforms with proprietary transaction, operational, or user-behavior data hold a structural advantage.

The expert believes this data dynamic explains why large vertical platforms are opting to build proprietary models rather than relying on external foundation model providers. The expert cited Meituan as an example that has developed its own LLM, LongCat, specifically to prevent proprietary operational data leakage.

While the general-purpose foundation model market is expected to consolidate over time, the expert anticipates that highly viable, industry-specific solution providers will continue to emerge across the healthcare, education, travel, and finance sectors.

Unit inference costs on the decline due to increased adoption of domestic chips and improved infrastructure efficiency

On the infrastructure side, the expert said Moonshot has increased the proportion of domestic inference chips in its infrastructure, while still using NVIDIA H20 and other NVIDIA GPUs for part of its inference workload. Under the current compute mix, the expert indicated that the gross margin for K2.7-related inference workloads is approaching the ~30% level.

API price reductions are likely to remain an industry trend, driven by competition and declining inference costs. However, the expert does not believe price cuts necessarily imply significant margin compression. As domestic chip usage rises, inference concurrency improves, and resource utilization increases, unit inference costs should continue to decline which could offset the pressure from price cutting.

Pivoting towards a closed-source strategy for easier monetisation

Moonshot's open-source framework serves as a balanced mechanism for developer

outreach, public relations, and commercial monetization. Open-source models are highly effective for building brand equity and developer mindshare – particularly in overseas markets – by lowering the entry barriers for initial testing and showcasing technical capability under resource-constrained environments, according to the expert.

Conversely, unrestricted open-source distribution can dilute direct commercialization, as enterprise users may favor free local deployment over paid API integrations, dampening conversion rates. To mitigate this, the expert stated that AI labs like Moonshot are inclined to adopt a tiered model strategy over time. Under such a strategy, mid-tier and lower-tier models may be open-sourced to support developer ecosystems and market awareness, while flagship models are likely to remain closed-source and priced at a premium to protect monetization and gross margin, according to the expert. We note that Alibaba's Qwen initially favored broad open-source deployment but chose a closed-source, premium API delivery model for its latest flagship model, Qwen 3.7 Max.

AI labs likely triumph over big internet platforms

The expert believes the general-purpose LLM market would undergo structural consolidation, ultimately leaving only a select group of players. The expert thinks that pure-play LLM developers like Moonshot and DeepSeek are well-positioned to compete effectively against large internet conglomerates like Alibaba and ByteDance, arguing that big internet platforms may risk diluting their focus by expanding across the entire AI value chain. In the expert's view, big platforms should and eventually will, focus primarily on cloud and AI infrastructure services, which are lucrative and easier to commercialise.

We do not fully agree with the expert's view regarding the ultimate landscape of the Chinese LLM industry. While independent AI labs currently hold a tactical lead in specific verticals like AI coding – owing to their first-mover focus – we believe large internet platforms remain formidable competitors in the broader LLM market. They possess deep technical know-how, unmatched capital reserves, and significant data ecosystems generated by their core internet operations.

We view it as highly unlikely that major tech platforms will de-emphasize their proprietary LLM initiatives simply because they are happy with the abundant revenue being generated from underlying infrastructure services. The strategic significance of owning a proprietary LLM extends far beyond standalone commercial monetization; it serves as the foundational core framework ("the brain") powering all existing and future consumer and enterprise AI applications. This strategic necessity will continue to drive large internet platforms to invest heavily to maintain SOTA model performance, in our view.

