

How do Chinese AI companies make money?

2026-10-06

Contents

Key takeaways	1
Overview	2
Consumer-facing AI applications appear to have weak gross margins	2
Selling model access accrues limited revenue, since releasing model weights reduces pricing power	5
Enterprise and government solutions generate meaningful gross profits, but are harder to scale	7
Licensing fees could help AI companies earn revenue in theory, but we lack public data on how much this contributes in practice	11
AI could generate indirect revenue by complementing existing product lines, but only in large conglomerates	13
Summary	15
Appendix	15
About the authors	18
Related topics	20

Key takeaways

- **Chinese AI firms earn a small fraction of what US frontier labs do.** As of September 2026, China’s six leading AI firms collectively earn about 10% of OpenAI and Anthropic combined, in terms of AI-related revenue.
- **Chinese AI companies earn revenue through five streams:** (1) consumer-facing AI applications, (2) selling model access (e.g., via API), (3) consulting and customizing AI for enterprises and governments, (4) AI licensing fees, and (5) using AI to complement existing product lines. Model-focused firms earn the most revenue through API sales, though this is unclear for large AI conglomerates, who potentially earn more by complementing existing product lines.
- **Consumer apps attract users but have low gross margins** — e.g., MiniMax’s AI-native apps had a gross margin of 4.7% in the first nine months of 2025. Margins are likely low because only a small fraction of Chinese AI app users pay, and those who do only pay small amounts.
- **API margins are often limited because model weights are publicly released, reducing pricing power.** Z.ai’s share of GLM 5.3 Flash tokens on OpenRouter fell from 88% to 22% in just 20 days after releasing model weights.
- **Enterprise and government solutions generate meaningful profits, but may be hard to scale.** For example, Z.ai has been deliberately shifting away from this business strategy, citing high maintenance costs that limit economies of scale. As such, from 2025 to H1 2026, the share of their revenue from on-premise deployments fell from 73.7% to 13.5%.
- **Licensing fees are a potential revenue stream, but there is little public data on their contribution.** Moonshot AI released Kimi K3’s weights under a modified MIT license, but we lack public data on overall earnings from the associated licensing fees.

- **For large AI conglomerates, AI can indirectly generate revenue by complementing existing product lines.** Alibaba and ByteDance’s model releases support their existing compute sales and advertising revenues.

Overview

As of September 2026, China’s leading AI companies earn a small fraction of what OpenAI and Anthropic do. Understanding their current revenue sources can help us assess the Chinese frontier AI industry’s potential by showing us where demand for Chinese AI comes from, which firms have structural advantages, and how these firms are shaping their business strategies.

Revenue estimates are from different months ranging from June to September. Run rates were reported for [Anthropic](#) (July), [OpenAI](#) (August), and [DeepSeek](#) (September), [Moonshot](#) (August), [MiniMax](#) (August), [Z.ai](#) (September), [ByteDance](#) (July), and [Alibaba](#) (August). The Alibaba and ByteDance figures are specific to Model-as-a-Service revenue, including serving proprietary and third-party models.

In this article, we analyze the revenue sources, gross margins, and growth potential of six notable Chinese AI companies: Alibaba, ByteDance, Z.ai, Moonshot, DeepSeek, and MiniMax. The first two are big tech companies with existing businesses besides AI (like Google), and the latter four are model-focused companies (like OpenAI).

Concretely, we find that these Chinese AI companies mainly make money from five sources:

1. Consumer-facing AI applications (direct monetization)
2. Selling model access (direct monetization)
3. Consulting and customizing AI for enterprises and governments (direct monetization)
4. Licensing fees (direct monetization)
5. Other products that are complementary to AI, e.g., hardware and compute (indirect monetization)

Examining these more closely helps shed light on which revenue source is the most important among Chinese AI labs. For model-focused companies, the dominant revenue source is API sales. In contrast, Alibaba and ByteDance potentially earn more through indirect monetization, though this is unclear from public data. We share company-specific data on these revenue streams in the Appendix.

Consumer-facing AI applications appear to have weak gross margins

Consumer-facing AI applications are AI products and interfaces used directly by individual consumers. Examples include general-purpose AI assistants on the web, apps like ChatGPT and Claude, and more specialized products like Sora or AI companions.

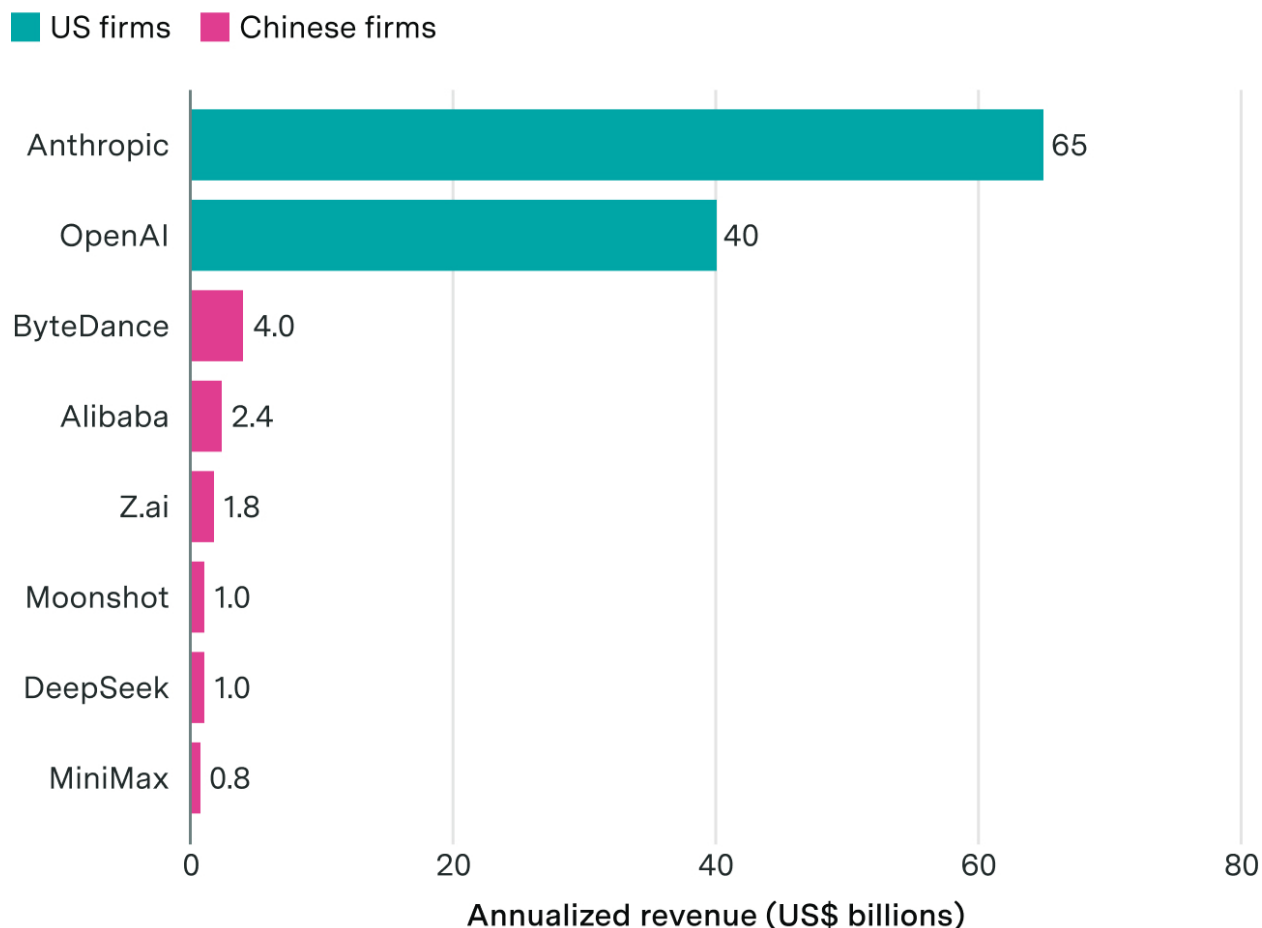
In China, ByteDance’s Doubao is the most popular AI-native app. DeepSeek, Qwen, and others are also AI-native apps with many monthly active users ([MAU](#)), [according to QuestMobile](#). Some video and audio models are also turned into AI applications, such as MiniMax’s Hailuo and Talkie.⁽¹⁾

These consumer apps make money in several ways. First, they can **display ads**. For example, advertisements inside Talkie, MiniMax’s companion app, generated 20.9% of the company’s total revenue in the first nine months of 2025.⁽²⁾

Alternatively, **they charge fees for better models or more tokens**. For example, ByteDance’s Doubao’s standard subscription [costs](#) ¥68 (\$9.6)/month.

However, this likely does not yield a large profit because consumers rarely pay for such services. MiniMax’s consumer-facing AI products, including MiniMax, Hailuo AI for video generation, MiniMax Audio for speech and audio generation, and Talkie, an AI companion app, accounted for

China's six leading AI firms together earn about 10% of OpenAI and Anthropic's combined revenue



Annualized direct AI revenue (from models and applications, excluding indirect revenue such as cloud compute and advertising) based on latest disclosures, as of September 2026.

Source: CNBC, Bloomberg, The Information, Caixin, 21st Century Business Herald, Sina Finance, Alibaba Q1 FY2027 earnings call.

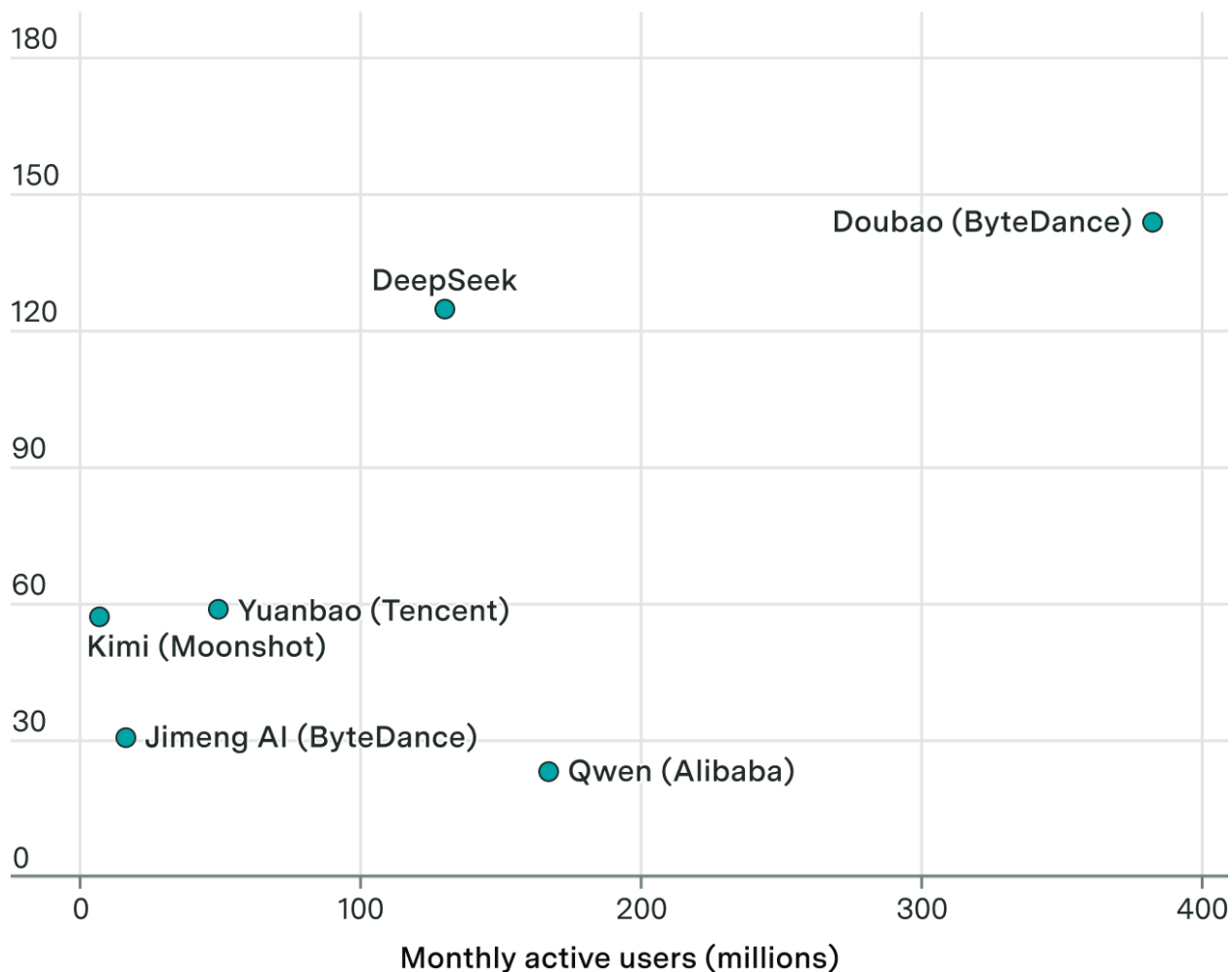
 EPOCH AI | CC-BY

epoch.ai

Figure 1: Bar chart of annualized revenue in US\$ billions: Anthropic 65, OpenAI 40, ByteDance 4.0, Alibaba 2.4, Z.ai 1.8, Moonshot 1.0, DeepSeek 1.0, MiniMax 0.8. China's six leading AI firms together earn about 10% of OpenAI and Anthropic's combined revenue.

Doubao leads Chinese AI apps on reach and usage

Minutes per user per month



Numbers based on data in June 2026. Users may be counted in more than one app.
Source: QuestMobile, 2026 H1 AI Application Market Report (July 2026).

 EPOCH AI | CC-BY

epoch.ai

Figure 2: Scatter plot of Chinese AI apps by monthly active users and minutes per user per month, June 2026. Doubao leads with about 380 million MAU and 140 minutes; DeepSeek has about 130 million MAU; Qwen about 170 million MAU but under 30 minutes.

36.6% of its total revenue in the first six months of 2026. However, the gross margin of MiniMax' s AI-native apps was only 4.7% in the first nine months of 2025. For comparison, MiniMax' s gross margin for Open Platform and other enterprise services was 69.4% in the same period.

Why are these margins so low? It is hard to know for certain without information about the compute and labor costs of deploying these consumer apps. But even still, we can identify likely contributing factors:

1. **Few users pay.** The apps averaged 27.6 million monthly active users (MAU) but only 1.77 million paid users in the first 9 months of 2025 (~6% of MAU).⁽³⁾ In another example, when ByteDance trialed a [paid membership service](#) for Doubao, the paid conversion rate was allegedly low. Low paid conversion is not unique to Chinese consumer-facing AI applications. [The Information](#) reports that only 5% of ChatGPT' s ~900 million weekly active users (WAU)⁽⁴⁾ paid in 2025, while OpenAI spent \$3.9B on inference for nonpaying users.
2. **Users who do pay don' t spend a lot.** MiniMax Hailuo' s standard package costs as little as ¥55/\$7.99, which is cheap compared to video models in the US. Another example is Z.ai' s chatbot, which is free in the US, and with a [VIP subscription](#) of \$11 (¥79)/month in China. In contrast, OpenAI' s paying users pay far more: ChatGPT Plus starts at \$20/month.

Overall, consumer-facing AI apps are not highly profitable when there are many free-tier users and paid tiers are very cheap. Ads could help monetize free-tier users, though the revenue they generate depends a lot on the size of the user base.

Selling model access accrues limited revenue, since releasing model weights reduces pricing power

AI companies can sell model access through either metered API tokens or flat-token subscription plans.⁽⁵⁾ Most revenue in US frontier AI comes from these sources. [SemiAnalysis](#) estimates that 75-85% of Anthropic' s revenue comes from selling API tokens.

Chinese AI companies also sell model access, but using a somewhat different strategy: they tend to release their model weights, which increases adoption but comes at the expense of their pricing power.

Consider Z.ai' s GLM-5 model. [As of September 16th](#), the model had eight separate providers, some of which (GMI Cloud and StreamLake) had cheaper offerings than Z.ai itself. Only about 25% of all GLM-5 tokens were served through Z.ai' s own API over the previous 24 hours, with the rest handled by other providers. Another example is DeepSeek V4 Pro. As of the same date, the [cheapest offering](#) on OpenRouter was under 87% of DeepSeek' s input token pricing. And in the last 24 hours, only 28.4% of the model' s tokens were served by DeepSeek.

More generally, there is a pattern: when open-weight model developers charge a higher premium, they capture a smaller share of their own models' tokens on OpenRouter.

Prices are weighted averages of uncached-input, output, and cached-input token rates, using the same token mix and cache-hit rate for all hosts of a given model.

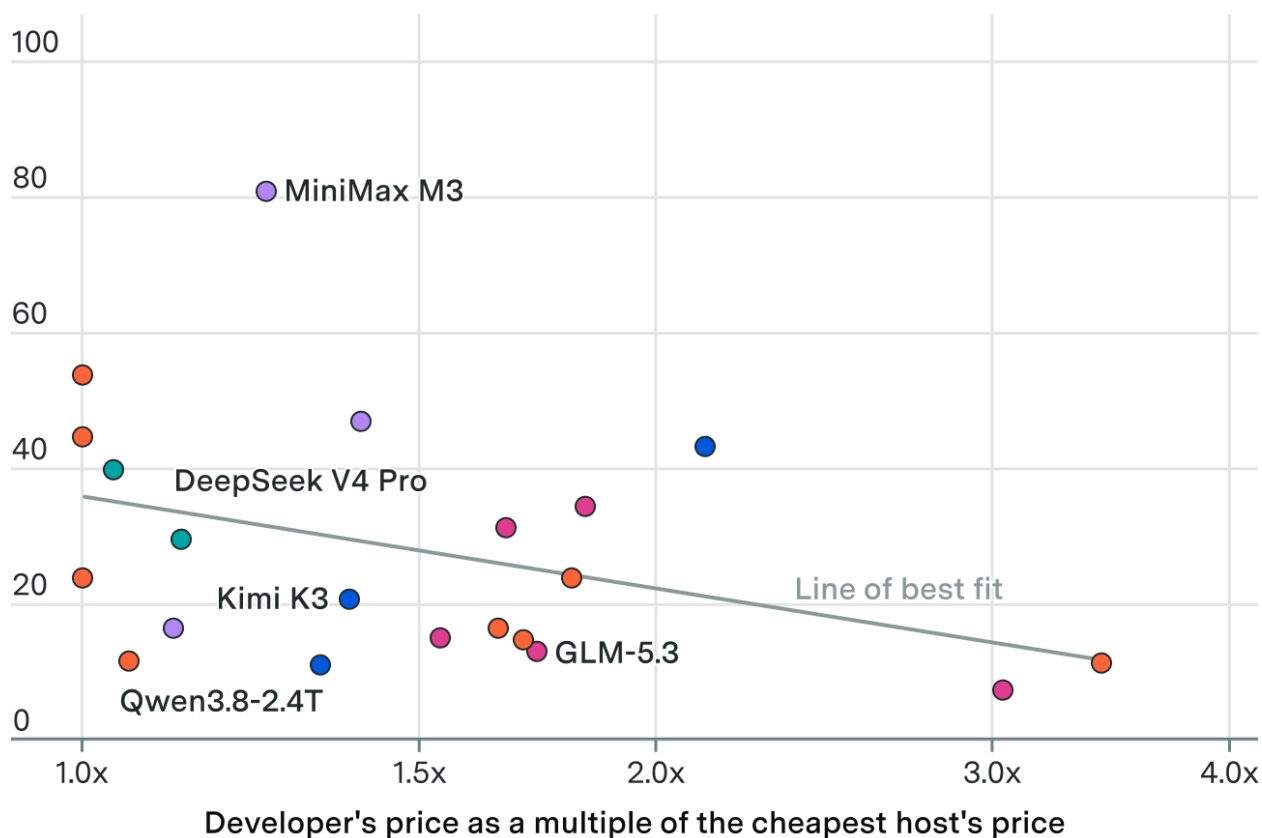
By contrast, a closed-weight model such as Anthropic' s Fable is almost always [provided](#) at the same list price of \$10 per million input tokens, whether it' s accessed from Claude' s API platform or Amazon Bedrock. Small price differences sometimes arise because of factors like geographic restrictions or service tiers, but these are not the result of providers competing to serve the model more cheaply.

We can get an even clearer sense of how releasing model weights reduces pricing power using a GLM 5.3 Flash case study. On August 26th, Z.ai listed the model on OpenRouter at a 50% promotional discount and simultaneously released its weights. Within hours, third parties began serving the model, and after just a day there were twelve competitors, two of which price-matched the promotional cost. Z.ai had already ended their discount by September 15th, but by then they found

Developers that charge a higher premium serve a smaller share of their own model's tokens

■ DeepSeek ■ Z.ai ■ Alibaba/Qwen ■ Moonshot ■ MiniMax

Developer's share of the model's OpenRouter tokens (%)



Data from 21 open-weight Chinese models, averaged over eight days between Aug 27 and Sep 19, 2026.

Source: OpenRouter provider pages.

EPOCH AI | CC-BY

epoch.ai

Figure 3: Scatter plot of 21 open-weight Chinese models: the developer's price as a multiple of the cheapest host's price against the developer's share of the model's OpenRouter tokens. A downward-sloping line of best fit shows that developers charging a higher premium serve a smaller share.

themselves competing with 26 rivals, 8 of which managed to undercut them. The consequences on pricing were massive: over those 20 days, Z.ai's largest rival, Relace, ended up serving more tokens than they did. Moreover, Z.ai's share of OpenRouter tokens collapsed from 88% (measured around 33 hours after launch) to 22%, and their daily token volume dropped 70% despite GLM 5.3 Flash's volume growing 17%.

Given this limited pricing power, what is the gross margin on API access? Unfortunately, there is very little precise data on this. According to Z.ai's annual and interim reports, the gross margins on their API and developer platform were 18.9% in 2025 and 24.6% in the first half of 2026. They attribute this growth to the "expansion of cloud-based deployment business, the launch of programming subscription packages, and enhanced inference efficiency." ⁽⁶⁾ For comparison, Opus 4.8 had 85% API-serving gross margins at the first-party list API price by mid-2026.

One outlier is DeepSeek. The API gross margin for DeepSeek V4 (both Pro and Flash) is estimated to be 70-80%. This is due to two reasons. First, Liang Wenfeng [stated](#) that DeepSeek "doesn't have that much compute" and "the primary goal of the model we built is not for others to find it easy to use, but for ourselves to find it easy to use." According to [The Information](#), DeepSeek "has continued to invest in developing new models, allocating more than 70% of its computing capacity to model training and less than 30% to inference." With limited compute resources, keeping the gross margin high helps reduce inference demand. Second, DeepSeek has managed to [keep its inference costs low by improving the efficiency of its AI infrastructure](#). Consistent with this, DeepSeek captures the smallest share of OpenRouter spending on its own models out of the five labs: only about 8% of estimated customer spending on DeepSeek models goes to DeepSeek. We see a similar pattern across other Chinese AI companies on OpenRouter, when we look at model use across eight sampled days.

Overall, releasing model weights leaves Chinese labs competing with third-party hosts to serve their own models. On OpenRouter, that competition limits what the labs can charge without losing volume and, for most of them, hands the bulk of the spending on their models to other providers.

Enterprise and government solutions generate meaningful gross profits, but are harder to scale

AI companies can monetize by customizing and deploying their models for companies or governments. This usually involves adapting the model to the customer's needs, integrating it with the customer's data, and deploying it in a dedicated cloud environment or on the customer's premises. ⁽⁷⁾

Most Chinese AI companies (except DeepSeek) do this to some extent, but Z.ai in particular is known for serving entities with sensitive information, like government agencies, state-owned enterprises, and banks. This may be due to their background at the prestigious [Tsinghua University](#), which gives the company credibility. For example, in 2025, [Z.ai worked with Hangzhou Urban Investment Group](#) to create a public transportation model, flood control agent, and multimodal large-scale road and bridge maintenance model. According to Z.ai's annual report, 73.7% of its revenue in 2025 came from on-premises deployment with a gross margin of 48.8%.

However, Z.ai is shifting away from the on-premises deployment business toward standardized cloud and API services. Notably, we see a sharp decline in on-premises deployment in H1 2026.

This may be because it is hard to scale this business strategy. As the company's [interim report](#) states, "the delivery and maintenance costs of this type of business have a certain degree of rigidity, resulting in reduced economies of scale." Unlike selling standardized API access, customized deployments require substantial engineering work to fine-tune and integrate a model into a customer's business. Z.ai's gross margin on on-premises deployment has indeed decreased over the years. ⁽⁸⁾

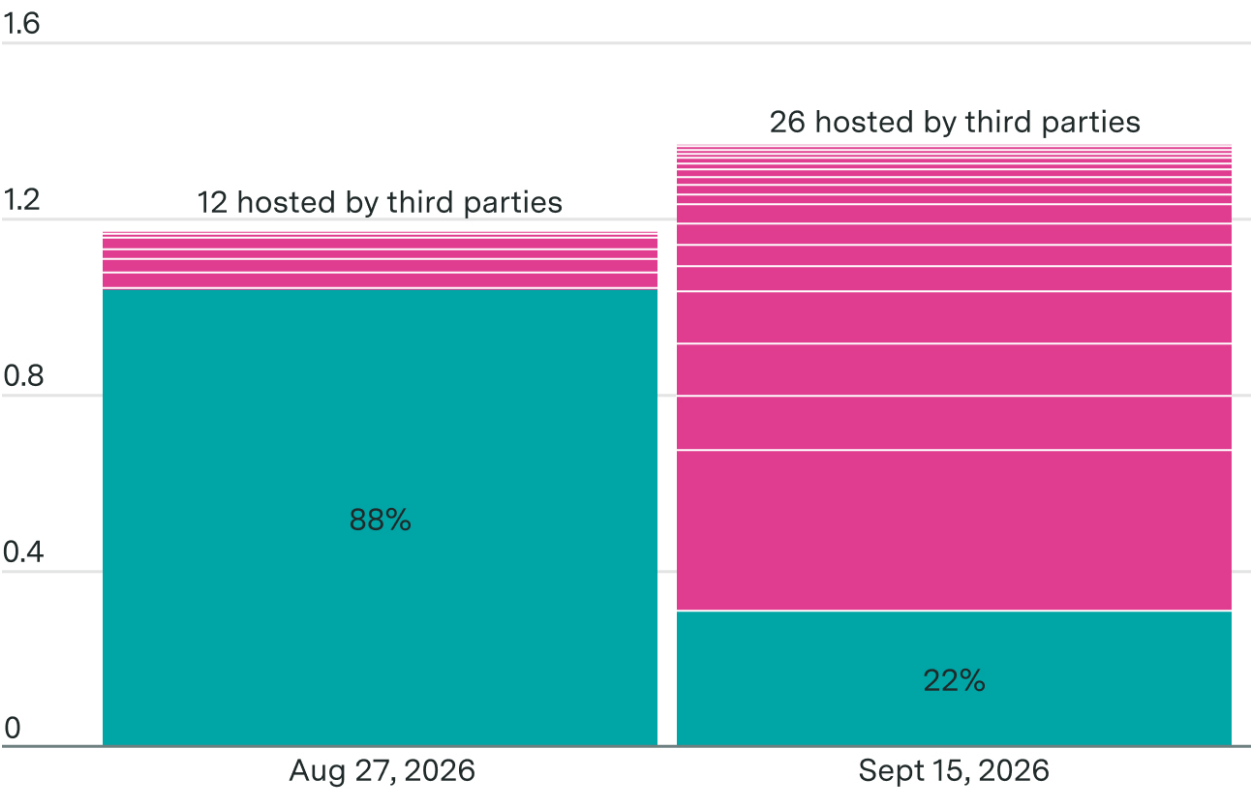
Alibaba and ByteDance also sell enterprise and government solutions. In May 2026, Alibaba's

Third parties took over serving GLM 5.3 Flash within three weeks of its open-weight release

Number of GLM 5.3 Flash tokens served by Z.ai and third-party hosts on two separate days.

■ Hosted by Z.ai ■ Hosted by third parties (one slice each)

Tokens (trillions)



Source: OpenRouter provider pages.

EPOCH AI | CC-BY

epoch.ai

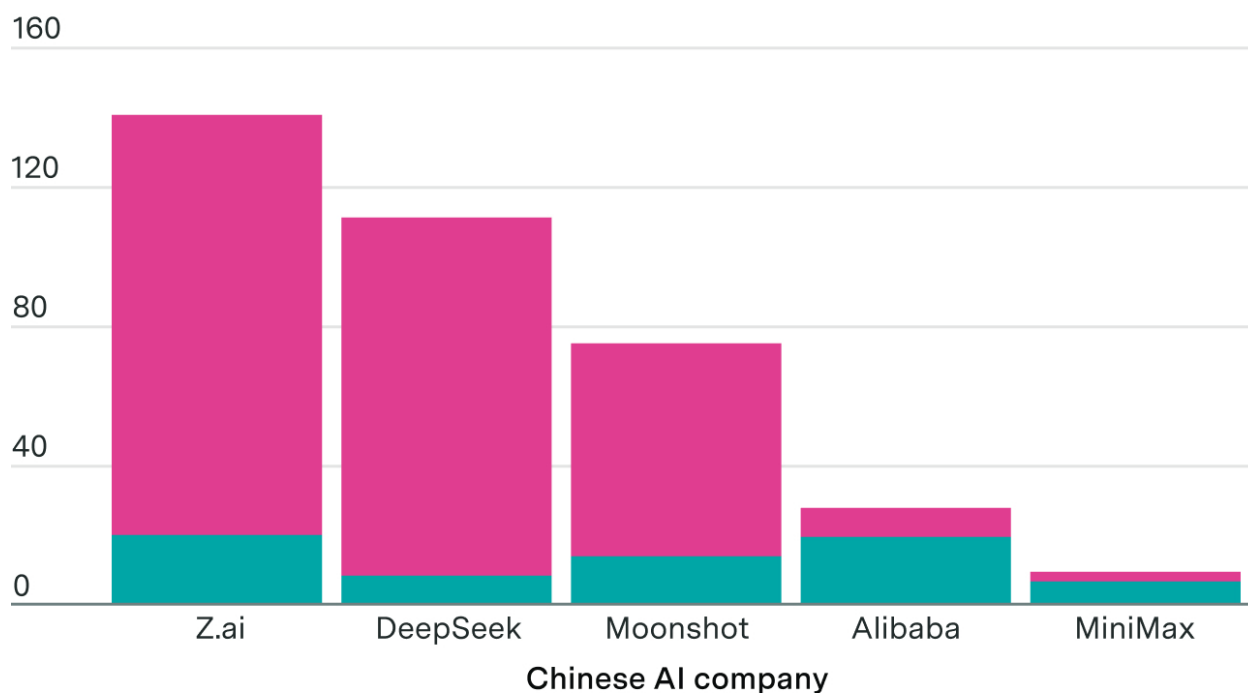
Figure 4: Stacked bar chart of GLM 5.3 Flash tokens served on OpenRouter on Aug 27, 2026 and Sept 15, 2026. Z.ai’ s share fell from 88% with 12 third-party hosts to 22% with 26 third-party hosts.

Most OpenRouter spending on Chinese models goes to third-party hosts, not to the model developers

Estimated customer spending on each developer's models on OpenRouter, annualized from eight sampled days between Aug 27 and Sep 19, 2026.

■ Hosted by developer ■ Hosted by third-party

US\$ million per year (annualized)



For each model, we estimate the average daily OpenRouter spending over eight days by multiplying two factors: (1) each host's volume of input and output tokens, and (2) their respective prices. Note that final annualized estimates could fluctuate a lot depending on which days we sample over. Source: OpenRouter provider pages.

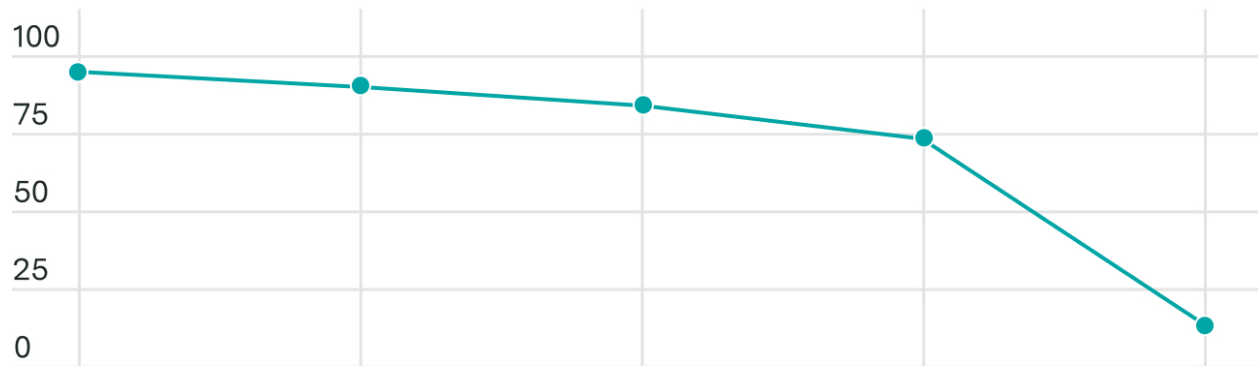
≡ EPOCH AI | CC-BY

epoch.ai

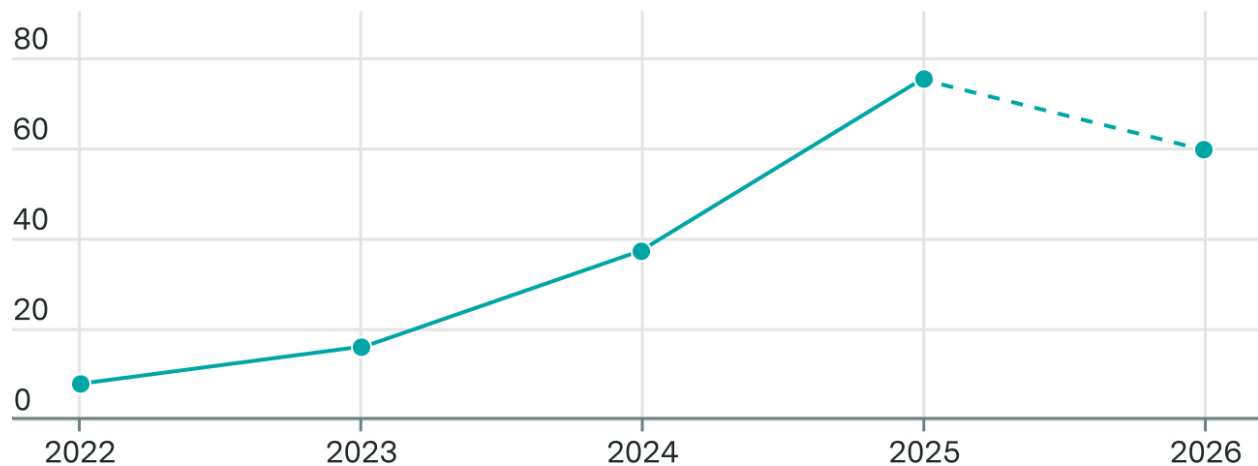
Figure 5: Stacked bar chart of estimated annualized OpenRouter spending on each Chinese developer's models, split between the developer and third-party hosts. For Z.ai, DeepSeek and Moonshot most spending goes to third parties; Alibaba retains the largest share.

Z.ai's revenue share from on-premises deployments fell sharply in 2026

Share of Z.ai's revenue (%)



Revenue (US\$ million)



Source: Z.ai HKEX listing prospectus, 2025 annual results announcement, and H1 2026 interim report. The 2026 share is for H1 2026, as reported. The 2026 revenue is a full-year estimate, annualized from H1 2026 data assuming 2025's seasonality

Figure 6: Two line charts of Z.ai's on-premises deployment business from 2022 to 2026: its share of Z.ai's revenue falls from over 90% to about 13.5% in H1 2026, while on-premises revenue in US\$ million rises to about 75.2 in 2025 before an estimated decline in 2026.

[Cloud Intelligence Group](#) collaborated with Tongji Hospital to work on medical imaging and support the hospital's research model training, among other initiatives. Similarly, ByteDance's enterprise cloud branch collaborated with the investment bank [China International Capital Corporation](#) to develop a customized financial model.

Unlike independent model startups, both Alibaba and ByteDance have more enterprise reach and cloud infrastructure, which make it easier for them to sustain an enterprise deployment strategy. We have not been able to find public information about enterprise solution gross margins for these two firms.

Licensing fees could help AI companies earn revenue in theory, but we lack public data on how much this contributes in practice

AI companies could in theory make money from licensing fees, but it is unclear whether developers could actually enforce these fees. At the moment, only a few flagship model releases from China's top AI labs are open under a restrictive license.

On July 27, 2026, Moonshot released Kimi K3's weights under a modified MIT license. Section 2 of the [license](#) reads:

“If the Licensee or any of its affiliates operates a Model as a Service business, and the aggregate revenue of the Licensee and its affiliates exceeds 20 million US dollars (or the equivalent in other currencies) in total over any consecutive 12 months, the Licensee must enter into a separate agreement with Moonshot AI before using the Software or its derivative works for any commercial purpose.”

The statement has several caveats. First, Section 4 of the license says that the requirements do not apply to “(a) internal use of the Software, defined as any use that does not make the Software, its outputs, or its underlying capabilities available to third parties.” For example, if a firm downloads Kimi K3 and runs it on its own servers for employees, that would constitute internal use, and the firm would not need to enter a separate agreement with Moonshot.

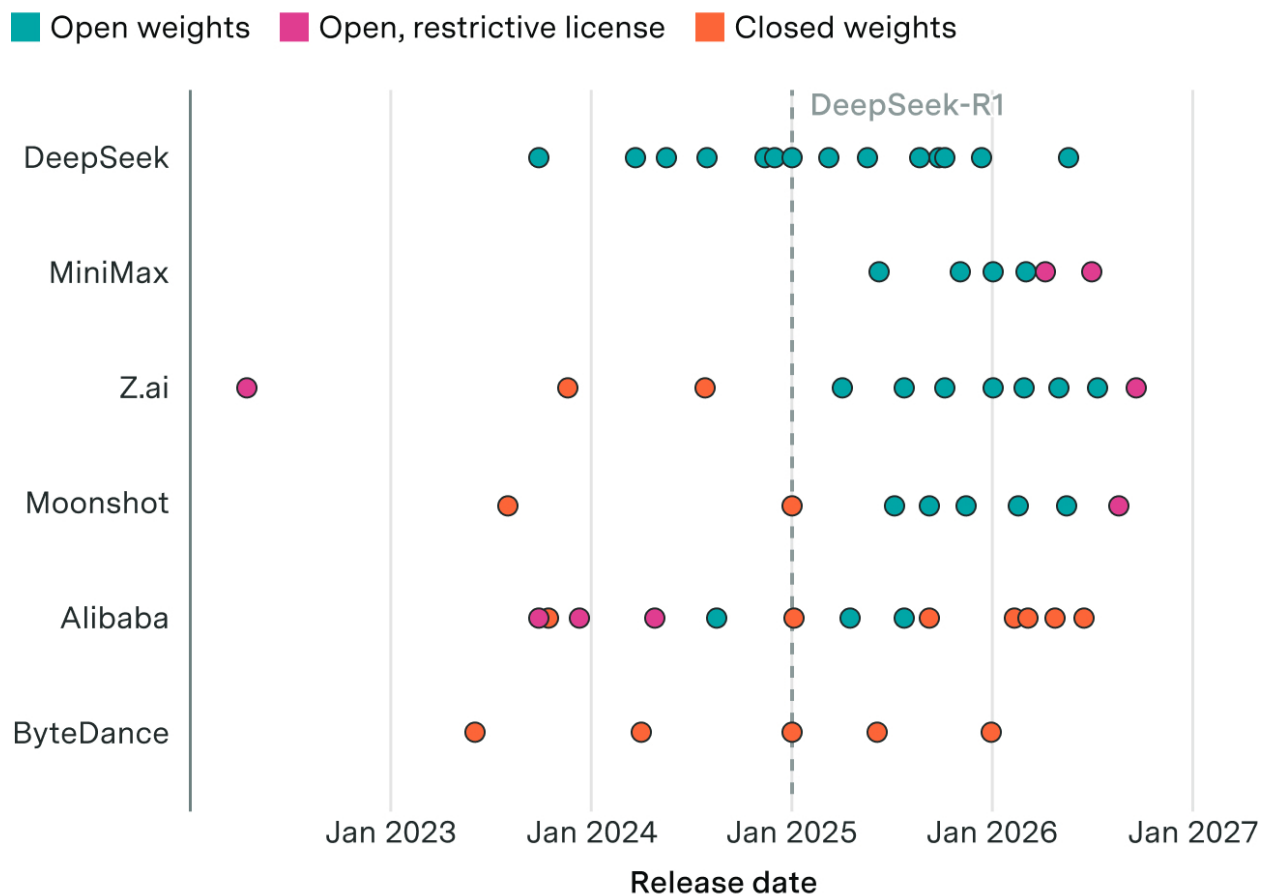
Second, the copyright status of model weights is unsettled. US copyright law generally requires human authorship, but model weights are numerical parameters produced by an automated training process. As such, enforcing these restrictions through copyright law may not be straightforward, and Moonshot might instead have to rely more on contract law.

Third, it could be difficult to attribute use to a specific model and determine the licensee's revenue. In practice, the license might therefore apply primarily to major cloud or inference providers and public companies whose revenue clearly exceeds the \$20 million threshold.

Ultimately, this license likely targets large inference providers who serve Kimi K3. These include (but are not limited to):

1. Chinasoft, a large public Chinese IT services company, [entered](#) a joint innovation cooperation with Moonshot to integrate Kimi for enterprise customers. The two companies share revenue based on token consumption.
2. Cursor, who built its own model on top of Kimi K2.5 through continued pretraining and reinforcement learning (RL), with RL sampling and inference run on Fireworks as part of an authorized [commercial partnership](#).
3. Together AI, a cloud platform and API provider, [entered](#) a strategic partnership with Moonshot AI.
4. According to [Reuters](#), Moonshot is seeking a share of up to 30% of revenue from K3-related services on Microsoft's Azure, AWS, and Google Cloud. (For comparison, [SemiAnalysis](#) estimates Anthropic's revenue share is about 75%, though after paying Amazon's infrastructure fee, its gross margin on Bedrock sales was only about 45% in early 2026).

Few flagship releases by China's leading AI labs are open under a restrictive license



Each dot is a flagship language model released by one of China's six leading AI developers. The sample only covers each firm's flagship line (Qwen, Doubao, GLM, DeepSeek-V/R, Kimi, and MiniMax-M), excluding small, specialty, and preview releases as well as non-language models. "Open, restrictive license" means that a model's weights are downloadable but commercial use is subject to conditions.

Figure 7: Timeline of flagship language model releases by DeepSeek, MiniMax, Z.ai, Moonshot, Alibaba and ByteDance from 2022 to 2026, colored by open weights, open with a restrictive license, or closed weights. A dashed line marks the DeepSeek-R1 release in January 2025. Only a handful of recent releases use a restrictive license.

Z.ai has also established revenue-sharing arrangements with cloud providers. According to China Star Market, Amazon has integrated Z.ai's GLM-5.3 into [Bedrock](#) in Oct 2026 and will share revenue with Z.ai. Z.ai has also [signed revenue-sharing agreements](#) with several leading domestic and international cloud service providers. However, the terms of the revenue-sharing arrangement, including the share received by Z.ai, have not been disclosed.

We do not have reliable figures for how much money is made in this way, though we do know that some American companies consider this a viable strategy. In particular, Meta has pursued a similar licensing strategy with Llama. Meta has [revenue sharing agreements](#) with Llama AI model hosts, and Llama 4's community license states the following:

“If, on the Llama 4 version release date, the monthly active users of the products or services made available by or for Licensee, or Licensee's affiliates, is greater than 700 million monthly active users in the preceding calendar month, you must request a license from Meta.”

AI could generate indirect revenue by complementing existing product lines, but only in large conglomerates

Up to this point, we have only considered direct sources of AI company revenue — that is, scenarios where the AI model itself is the product. But AI models can also produce revenue indirectly, such as by increasing sales of complementary products. This is only possible at large conglomerates with many product lines, like Alibaba and ByteDance.

Broad adoption of AI increases demand for the compute needed to run the models. Alibaba and ByteDance can thus earn revenue from enterprise AI projects not only by providing models and integration, but also by selling the associated compute. For example, the aforementioned Alibaba-Tongji Hospital collaboration includes Alibaba Cloud providing cloud computing infrastructure.

Collaborations like these may have contributed substantially to Alibaba's AI revenue. For Q2 2026, Alibaba's [revenue](#) from AI cloud and compute services was \$7.1 billion, with year-over-year revenue growth accelerating to 45%, allegedly driven by “the increasing adoption of AI-related products.” AI-related product revenue specifically reached \$1.8 billion in Q2 2026, the twelfth consecutive quarter in which this product line saw triple-digit year-over-year growth. Note that in annualized terms, these are higher than the \$2.4 billion we saw in the first figure of the report, which comes from MaaS sales specifically.

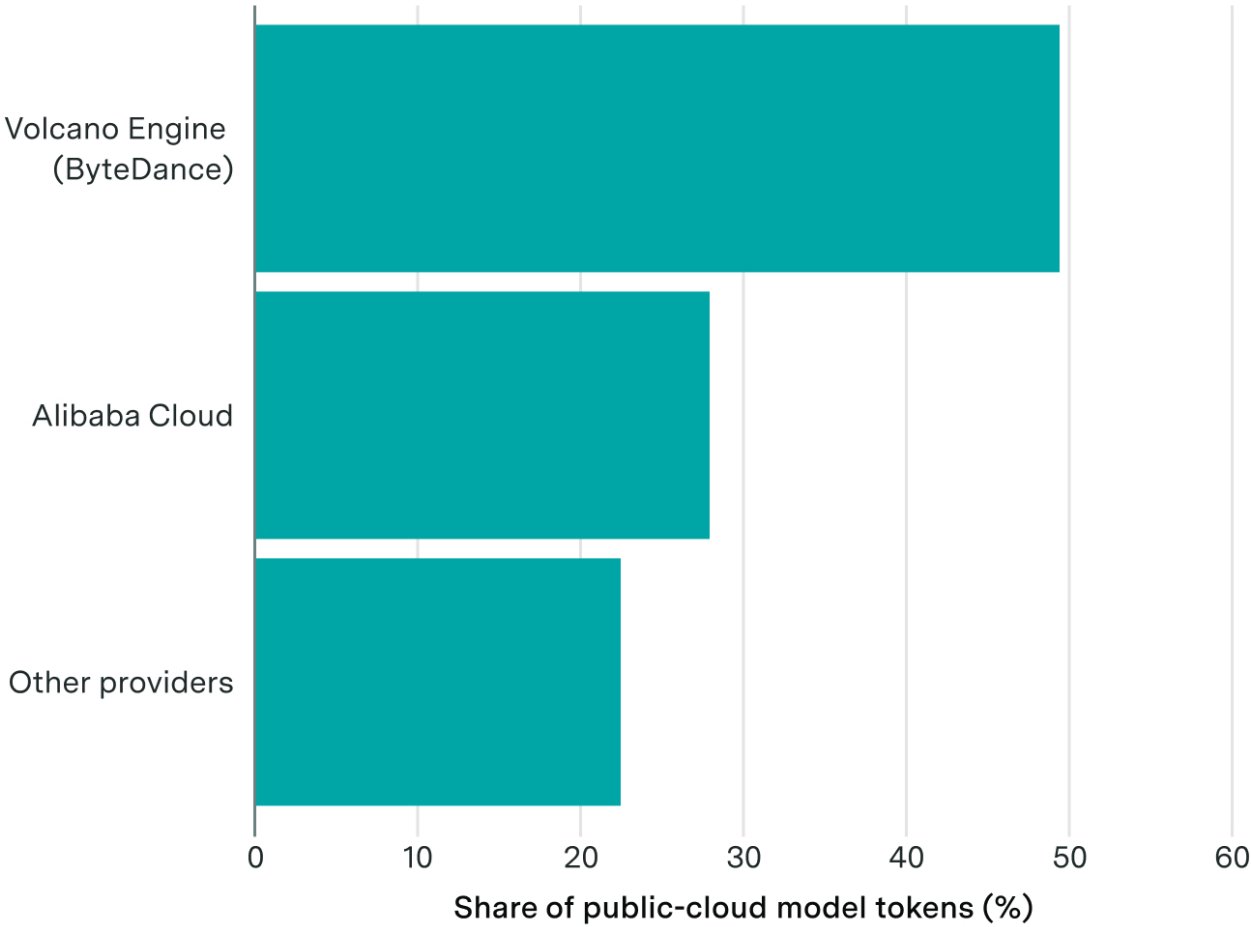
Similarly, ByteDance's cloud platform Volcano Engine handled 49.5% of China's public-cloud large-model token calls in 2025, according to the [IDC](#). Its model-as-a-service (MaaS) [revenue](#) was around \$221 million in 2025, and ByteDance recently raised its 2026 target to [\\$4 billion](#).

Alibaba and ByteDance's compute businesses also benefit from open-weight models from other companies. Alibaba's Model Studio (Bailian) [lists](#) DeepSeek, Z.ai, Moonshot, and MiniMax's models alongside Qwen. The CEO of Alibaba Group, Yongming Wu, [claimed](#) that “a thriving open source model ecosystem drives greater demand for our cloud computing services, creating a virtuous cycle.” Similarly, ByteDance's Volcano Engine has a [coding plan](#) that allows users to use DeepSeek, Moonshot, and Z.ai models.

A second source of indirect AI revenue is advertisements. As AI helps produce content, creators have more content to advertise on platforms, and thus spend more on ads to attract viewers. One interesting example is Seedance, which complements Douyin's (Chinese TikTok) advertisement business. As Seedance made creating videos cheaper, short drama production companies produced more episodes. In Q1 2026 alone, about 128,000 micro-dramas launched—over 95% of them [AI-generated](#)—compared to ~33,000 micro-dramas launched in all of [2025](#).⁽⁹⁾

After production, short dramas gain viewership by advertising on Douyin. Short-drama producers and distributors devote substantial resources to paid distribution. According to Wang Xiaoshu

ByteDance's Volcano Engine handled half of China's public-cloud model tokens in 2025



Shares of external customers' token usage on public-cloud model services in China, 2025. Model token counts exclude usage by providers' own products, such as ByteDance's Douyin, Doubao, and Jimeng.
Source: IDC, China AI Software Semiannual Tracker, 2025H2 (May 2026 summary).

Figure 8: Bar chart of 2025 shares of China's public-cloud model tokens: Volcano Engine (ByteDance) about 49.5%, Alibaba Cloud about 28%, other providers about 22%.

(founder of Jiashu Technology, a business distributing short dramas),⁽¹⁰⁾ his company’s spending on advertising has grown [tenfold](#) over the last year or so.⁽¹¹⁾ Li Tao (head of Fengxing Culture, the company behind many top short dramas) also emphasized that “for a short drama, 90% of the revenue is spent on online distribution.”

This ad spending is a big part of ByteDance’s business. [An industry insider](#) said that AI short-drama advertisers’ daily ad spending on ByteDance’s platform peaked at roughly ¥120 million. If this is true, and average daily spending was even a quarter of that peak, AI short-drama advertisers would be spending roughly \$1.6 billion per year on Douyin ads.⁽¹²⁾ That is roughly two-thirds of the API revenue for DeepSeek, Moonshot, MiniMax, and Z.ai combined.

AI can also increase advertising revenue more directly by improving and powering the tools used to create advertisements, such as Alimama’s multimodal creative platform Wanxiang Creation.⁽¹³⁾ These kinds of tools have produced quantitative results. Alibaba [reported](#) in 2025 that Chinese e-commerce customer management revenue “rose 10% year over year to ¥89.3bn (\$12.5 billion)”, and explicitly attributed the improvement in part to “the growing adoption of the AI-powered marketing tool Quanzhantui.”

Overall, this strategy of supporting complementary products seems fairly lucrative among large Chinese AI conglomerates. It is not surprising that we see similar strategies among US companies. Google [says](#) that “with the launch of AI Overviews, the volume of commercial queries has increased,” meaning that more ads can be served to users. Similarly, [Meta’s second quarter 2026 results](#) conference call says “9 million small businesses on [Meta’s] platforms are now using at least one of [their] AI ad creative tools.”

Summary

To some extent, Chinese AI companies face similar business problems to those faced by their US counterparts. For example, consumer AI products can attract large numbers of users who are expensive to serve but difficult to convert into paying customers.

However, Chinese companies face those constraints from a different starting point. Their models are less capable and their businesses are younger. This might have motivated strategies that trade short-term monetization for broader model diffusion, such as by releasing their model weights.

This does not mean Chinese companies will not capture the value they create. They are experimenting with new ways of reconciling diffusion with monetization, such as with modified licenses. Alibaba and ByteDance have an additional solution where they can benefit from widespread AI adoption through cloud computing, ads, and other existing businesses.

As Chinese AI models become more capable, it will be important to keep an eye on how Chinese companies try to monetize their models. The strategies they employ have significant implications for their future investments and funding, and it is crucial to watch how they evolve.

All translations are Cheryl’s, unless we specify otherwise.

We would like to thank Alan Chan, Campbell Hutcheson, Zi Cheng Huang, Mary Clare McMahon, Zilan Qian, Konstantin Pilz, Lisa Soder, Jean-Stanislas Denain, Karthik Tadepalli, Elias Groll for their helpful feedback on this article.

Appendix

Appendix: Summary by company

Below is a company-by-company summary of the revenue sources discussed in this article. “Main AI revenue source” is our estimate of each firm’s main source of AI-related revenue, based on

company disclosure and other public reporting. Estimates for MiniMax and Z.ai are from their interim reports, annualized assuming 2025' s seasonality.

DeepSeek

Main AI revenue source: Selling model access.

- Consumer-facing apps: Free [app/web](#); no ads.
- Selling model access: [Yes](#). Estimate: [\\$1B](#), assuming most of DeepSeek' s revenue comes from API sales.
- Enterprise/government solutions: Not publicly identified.
- Licensing fees: Not publicly identified.
- Other businesses complemented by AI: Not publicly identified.

Moonshot

Main AI revenue source: Selling model access.

- Consumer-facing apps: [Subscription](#). Estimate: unknown, but probably low.
- Selling model access: [Yes](#). Estimate: [70%](#) of total revenue as of June 2026.
- Enterprise/government solutions: [Some](#). Estimate: unknown.
- Licensing fees: [Yes](#), for K3. Starting to partner with forward-deployed engineers such as AsiaInfo and let them serve private-deployment and customization projects.
- Other businesses complemented by AI: Not publicly identified.

MiniMax

Main AI revenue source: Selling model access, together with enterprise services. The two are reported as one figure in the interim report. Note that to categorize these revenues accurately, we use the numbers from the interim report (H1 2026), which differ from the number in the first figure of this report.

- Consumer-facing apps: [Subscription, credits, and ads, mainly on Talkie](#). Estimated 2026 revenue: \$107M (based on H1 results and 2025 H1/FY seasonality).
- Selling model access: [Yes](#). Estimated 2026 revenue: ARR \$208M for 2026 (based on H1 results and 2025 H1/FY seasonality), combining API and enterprise solutions.
- Enterprise/government solutions: [Yes, including dedicated inference resources, customization, and customer deployment](#). Estimate: see above, we only know the revenue combined with API sales.
- Licensing fees: [Yes](#), but mainly commercial-use and attribution conditions.
- Other businesses complemented by AI: Not publicly identified.

Z.ai

Main AI revenue source: Selling model access. Note that to categorize these revenues accurately, we use the numbers from the interim report (H1 2026). As a result, the numbers shown here differ from the \$1.8 billion ARR reported by Z.ai management in September, and shown in the first figure of this report.

- Consumer-facing apps: [Subscription](#). Estimate: unknown, but low.
- Selling model access: [Yes](#), now its main business. Estimated 2026 revenue: \$805M (based on H1 results and 2025 H1/FY seasonality).
- Enterprise/government solutions: Used to be its main [business](#), but has [shifted away](#). Estimated 2026 revenue: \$60M for 2026 (based on H1 results and 2025 H1/FY seasonality).
- Licensing fees: [Yes](#). GLM-5.3 has a custom [license](#), and has started revenue sharing with Amazon Bedrock and multiple domestic cloud providers.

- Other businesses complemented by AI: Not publicly identified.

ByteDance

Main AI revenue source: other businesses complemented by AI.

- Consumer-facing apps: [Subscription](#) plus [commerce and channel fees](#), which funnel customers to ByteDance's Douyin shop. High future revenue potential.
- Selling model access: [Yes](#), through Volcano Engine. Estimate: [~\\$4B](#) MaaS ARR for 2026. This figure could also include enterprise solutions.
- Enterprise/government solutions: [Yes](#), through Volcano Engine. Estimate: see above. [~\\$4B](#) MaaS ARR for 2026. This figure also includes API sales.
- Licensing fees: No. Mostly closed-weight.
- Other businesses complemented by AI: Yes: [ads](#), [e-commerce](#), and [cloud](#).

Alibaba

Main AI revenue source: other businesses complemented by AI.

- Consumer-facing apps: [Subscription](#), rolled out August 2026, plus [Qwen funnels users into Alibaba commerce](#). Estimate: unknown.
- Selling model access: Yes, [QwenCloud](#). Estimate: unknown. Model and application services: roughly [\\$2.4B](#) annualized in August. This includes more than Qwen API sales, such as third-party model hosting.
- Enterprise/government solutions: [Yes](#), Bailian Dedicated Edition. Estimate: unknown.
- Licensing fees: Yes, e.g., [Qwen-3.8 Flash-Next](#).
- Other businesses complemented by AI: Yes: [e-commerce](#) and [cloud](#). Estimate: AI-related product revenue reached [\\$1.824B](#) in Q2 2026 (ARR \$7.3B).

Notes

1. Hailuo AI is MiniMax's consumer-facing video-generation product, available through its website and mobile apps and powered by its Hailuo video models; Talkie is an AI companion app.
2. According to MiniMax's IPO document, ad revenue is \$11,188,000 for the first 9 months of 2025, or on average \$41,000/day.
3. Note that this cross-period ratio is not a conventional monthly paid-conversion rate. MiniMax does not disclose the actual monthly conversion rate. In theory, the true average-month figure could lie anywhere from roughly 0.7% to 6.4%, depending on how frequently those 1.77 million users paid or remained active.
4. Note that WAU figures are not convertible to MAU when comparing the number to Chinese AI apps in Figure 1. Sensor Tower reports that ChatGPT reached approximately 1 billion mobile MAU in May 2026. We can also compare companies' app usage with a back-of-the-envelope calculation. According to QuestMobile, Doubao's MAU in June 2026 is 382.3 million with an average of 143.7 minutes per month per MAU. Multiplying the numbers, we get 915.6 million user-hours monthly. Sensor Tower reports approximately 215 minutes per ChatGPT user per month. Multiplying that by MAU, we get 3.58 billion user-hours monthly, approximately $4 \times$ Doubao's.
5. Metered API access charges per token used, with separate rates for input and output. A flat-rate plan charges a fixed monthly fee for a usage allowance; Z.ai's GLM Coding Plan, which works inside coding assistants such as Claude Code, is an example.
6. The other way to sell an unmodified model is a flat-rate plan: a monthly or per-seat subscription, usually bundling the model with a harness such as a coding agent.

7. Note that customers can also customize models themselves. For example, customers can [fine-tune](#) OpenAI models on their own training data. This is better classified as part of the API business than as a bespoke customization service.
8. Z.ai's gross margin on on-premises deployment went from [66.0%](#) (2024) to [48.8%](#) (2025) to [37.9%](#) (H1 2026, down from 59.1% in H1 2025).
9. In Q1 2026, approximately 122,000 of the 128,000 micro-dramas launched in China were AI-generated. However, note that for the Spring Festival season of 2026, "the number of live-action short dramas released was about 1/50 that of AI dramas, but the total number of views reached 25 times that of AI short dramas." It is possible that live-action short dramas earn more viewers than AI-generated short dramas.
10. Jiashu Technology is simultaneously [producing live-action short dramas and AI-generated short dramas](#), with the latter accounting for 60% of its domestic business.
11. We don't know the gross margin of advertising on Douyin, but in 2020, when advertising accounted for [~77%](#) of ByteDance's revenue, it had an overall gross margin of roughly [56%](#).
12. This is an illustrative estimate, not reported ByteDance revenue, and does not account for rebates, taxes, or other differences between gross billings and recognized revenue. We do think 25% of peak spending is within the reasonable range, however.
13. Alimama is a subsidiary of Alibaba.

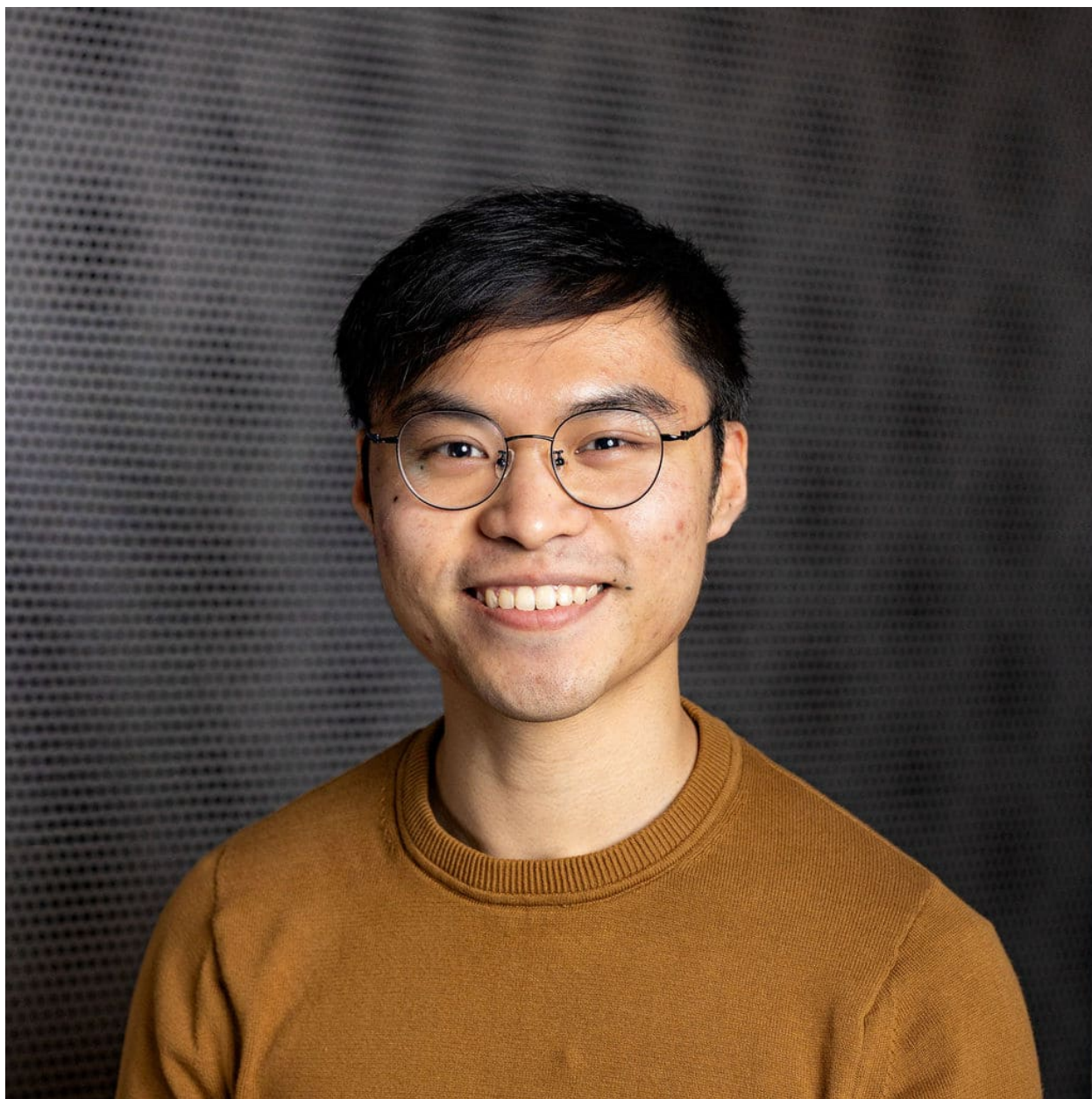
About the authors

Cheryl Wu

Cheryl Wu is a research scholar at GovAI and a Ph.D. candidate in Economics at Yale University. Her research focuses on the economics of AI and China's AI ecosystem.



Figure 9: Cheryl Wu' s avatar



Anson Ho

Anson Ho is a senior researcher at Epoch AI. He is interested in helping develop a more rigorous understanding of future developments in AI and its societal impacts.

Related topics

[Leading companies](#)[Finances](#)[Geopolitics](#)