China LLM

4 August 2026 | 7:27 AM China Standard Time

Jerry Shen
Goldman Sachs (Asia) L.L.C.
FICC & Equities

Did you know that Chinese AI models are now capturing the lion's share of token usage on global platforms like OpenRouter, fueled by pricing that's often just a third of their peers? Welcome to the new frontier of China's AI race. GIR Ronald and the China Internet team have just dropped a massive update, and the growth trajectory is staggering. We are witnessing a hyper-accelerated adoption curve where extreme value-for-money and breakthrough multi-modal capabilities are completely rewriting the rules of engagement.

Where GIR Stands Out While the street remains cautious on AI monetization timelines, GIR Ronald is pounding the table on revenue acceleration. **We have aggressively raised our 2026 aggregate Annualized Recurring Revenue (ARR) estimate for Chinese AI models to $13bn (up from our prior $10bn forecast)**. Factoring in this rapid API adoption, we've hiked our 2026 revenue estimates for Z.AI and MiniMax by 35% and 63%, respectively. The market is underestimating the shift from free-to-use consumer apps to highly monetizable, enterprise-grade agentic workflows.

Actionable Ideas

- **Buy MiniMax (TP HK$800):** The risk-reward here is heavily skewed to the upside. The recent launch of their H3 multi-modal model (seamlessly blending text, image, video, and audio) utilizes an open-weight approach priced at an incredibly disruptive 30%-50% of incumbents. This architectural efficiency is a massive catalyst for 2H26 ARR ramp-up and significant valuation repair.
- **Overweight Cloud & Data Centers: This remains our #1 preferred sub-sector play**. We love **GDS, VNET, Alibaba, and Kingsoft Cloud**. The explosion of AI token demand from coding and agentic AI, combined with hyperscalers aggressively ramping up capex into 2H26, provides clear margin upside and pricing power.
- **Domestic Compute Supply Easing:** Watch for Alibaba's new T-head super nodes and Tencent's aggressive domestic AI chip deployment. The infrastructure bottleneck is easing, paving the way for the next generation of 2T+ parameter models.

The contest has shifted from pure model intelligence to the best performance-to-price balance and agentic application ecosystems.

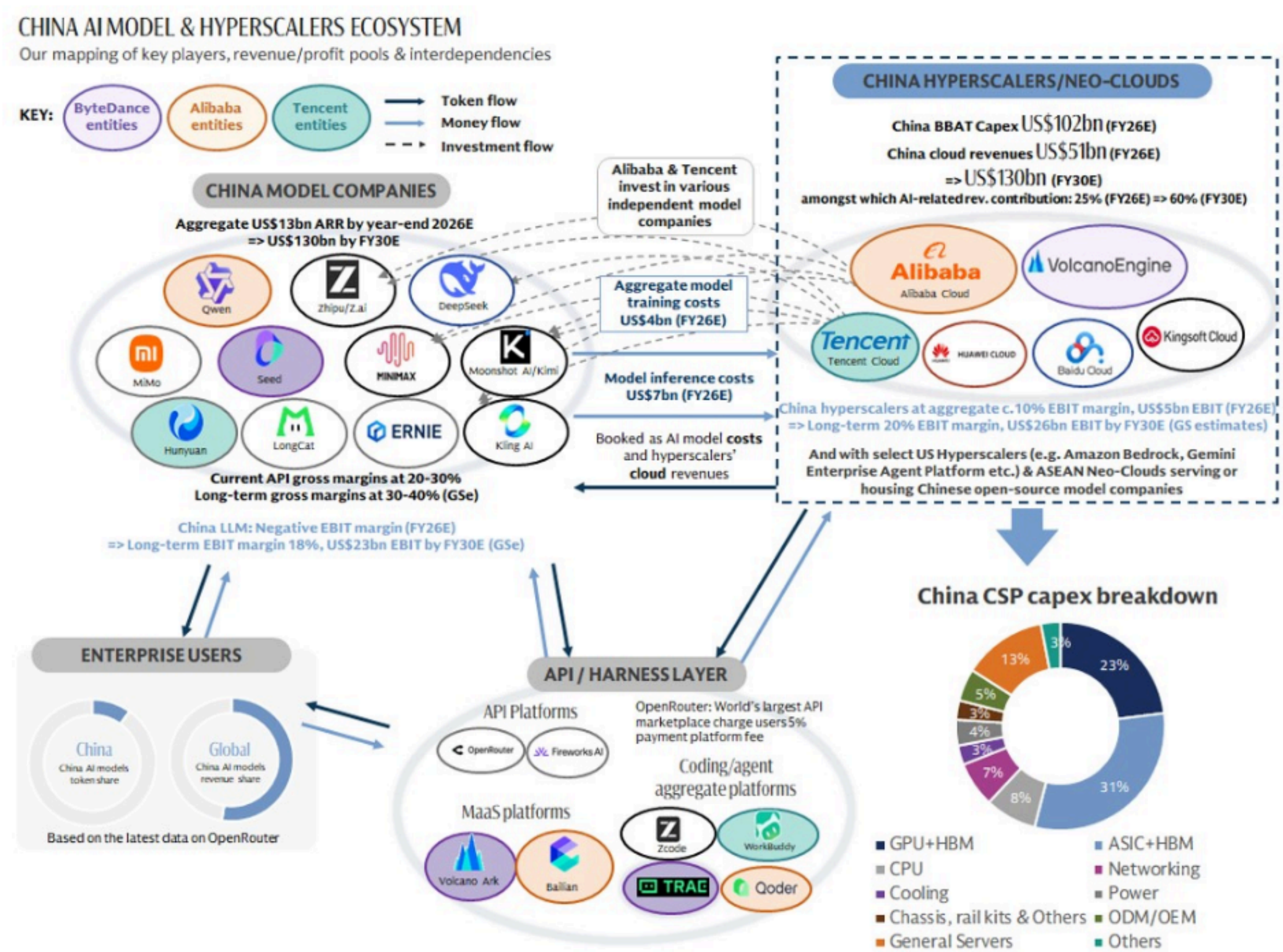
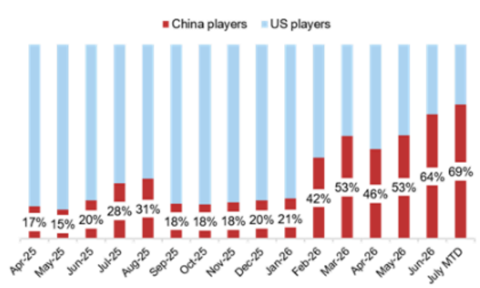
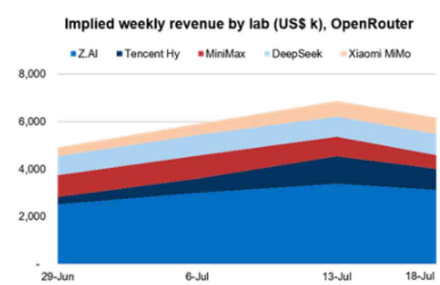


Exhibit 4: China model players continue to capture market share by token consumption on OpenRouter



Source: OpenRouter, Data compiled by Goldman Sachs Global Investment Research

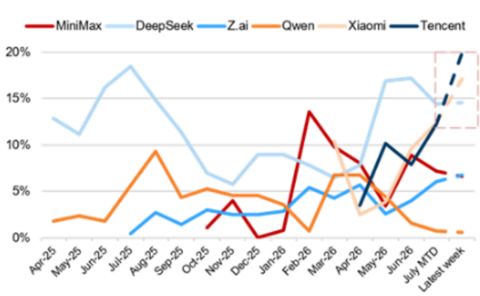
Exhibit 5: Implied weekly revenue rise according with token growth, especially for higher pricing coding models



Source: OpenRouter, Company data, Data compiled by Goldman Sachs Global Investment Research

Exhibit 6: We see leading share gain from value-for-money agentic models, esp. Xiaomi/Tencent over the latest week post DeepSeek rush hour price hike

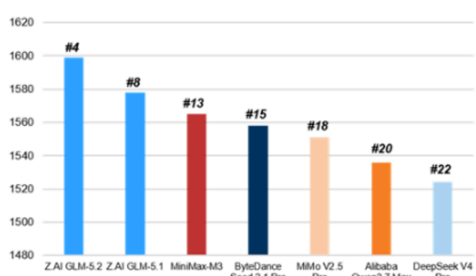
Market share of Chinese AI model players, of token consumption on OpenRouter



Source: OpenRouter, Data compiled by Goldman Sachs Global Investment Research

Exhibit 7: China models' ranks on Code Arena leaderboard for full-stack web development tasks

As of late July, 2026



Source: LMArena, Data compiled by Goldman Sachs Global Investment Research