

Why Enterprises Are Moving to Multi-Model AI Aggregation Platforms

300+ AI Models for OpenClaw & AI Agents

Save 20% on Costs

Free \$1 Tokens for New Members



Singapore, Singapore Jul 23, 2026 ([Issuewire.com](https://www.issuewire.com)) - Cost Reductions of 67 Percent and Threefold Deployment Speed Gains Drive the Shift

Enterprises are abandoning single-provider AI strategies and shifting to multi-model aggregation platforms as new data reveals cost savings exceeding 60 percent and deployment timelines cut by two-thirds, according to [Al.cc](https://www.al.cc), a unified AI API platform serving more than 10,000 active users worldwide.

The transition marks a fundamental change in how organizations approach artificial intelligence infrastructure. Global AI API market revenue reached \$64.41 billion in 2025 and is forecast to surpass \$900 billion by 2035, according to Precedence Research. As the market scales, enterprises are learning that routing all workloads to a single premium model — standard practice 12 months ago — is no longer cost-effective.

Analysis of 2.4 billion API calls shows enterprise token costs fell 67 percent year-over-year, with blended cost per million tokens declining from \$18.40 to \$6.07. The primary driver is intelligent task routing: in Q1 2025, 73 percent of enterprise token volume flowed to the two most expensive model tiers. By Q1 2026, that share dropped to 31 percent, with 69 percent distributed across mid-tier and cost-efficient models matched to task complexity.

"Enterprise teams were sending simple classification tasks and structured data extraction through frontier models because that was what they had integrated," said the [Al.cc](https://www.al.cc) research team. "Intelligent

routing alone accounts for 34 percentage points of the total 67 percent cost reduction."

Open-source models have accelerated the shift. These models captured 38 percent of enterprise token volume in Q1 2026, up from 11 percent in Q1 2025 — a 245 percent share increase driven by aggressive pricing from providers such as DeepSeek. Average models per enterprise account rose to 4.7, up from 2.1, indicating multi-model architecture has become the default.

Beyond cost, enterprises report significant speed improvements. Teams using multi-model infrastructure deployed production AI agents in a median of 3.6 weeks, compared with 11.2 weeks for single-provider integrations — a threefold improvement in time to market.

[AI.cc](#) operates as a unified aggregation layer across providers including OpenAI, Google, Anthropic, xAI, DeepSeek, Alibaba, ByteDance, and MiniMax. The platform delivers below-retail pricing, with effective discounts averaging 23 percent versus direct provider rates and reaching 35 to 40 percent for high-volume enterprise accounts. Enterprises using multi-model routing on the platform reported median cost reductions of 71 percent, with the top quartile exceeding 80 percent.

"Multi-model strategy is no longer optional," the report states. "Businesses still routing all AI requests to a single premium provider are overpaying by a significant margin."

The findings align with broader industry analysis. Research and Markets projects the global AI API market will grow by \$121.73 billion between 2025 and 2030 at a compound annual growth rate of 26.3 percent. As multi-model routing, prompt caching, and aggregated pricing redefine enterprise AI economics, organizations that fail to adapt risk falling behind competitors who have already captured these efficiencies.

About AI.cc

[AI.cc](#) is a unified AI API aggregation platform providing access to more than 300 artificial intelligence models through a single OpenAI-compatible interface. Founded in Singapore, the platform serves over 10,000 active users processing 90 million daily API requests across chat, video, image, voice, code, and embedding model categories. For more information, visit www.ai.cc.

Media Contact

AI.cc Media Relations
press@ai.cc



Media Contact

AICC

*****@ai.cc

<https://www.ai.cc>

Source : AICC PTE. LTD.

[See on IssueWire](#)