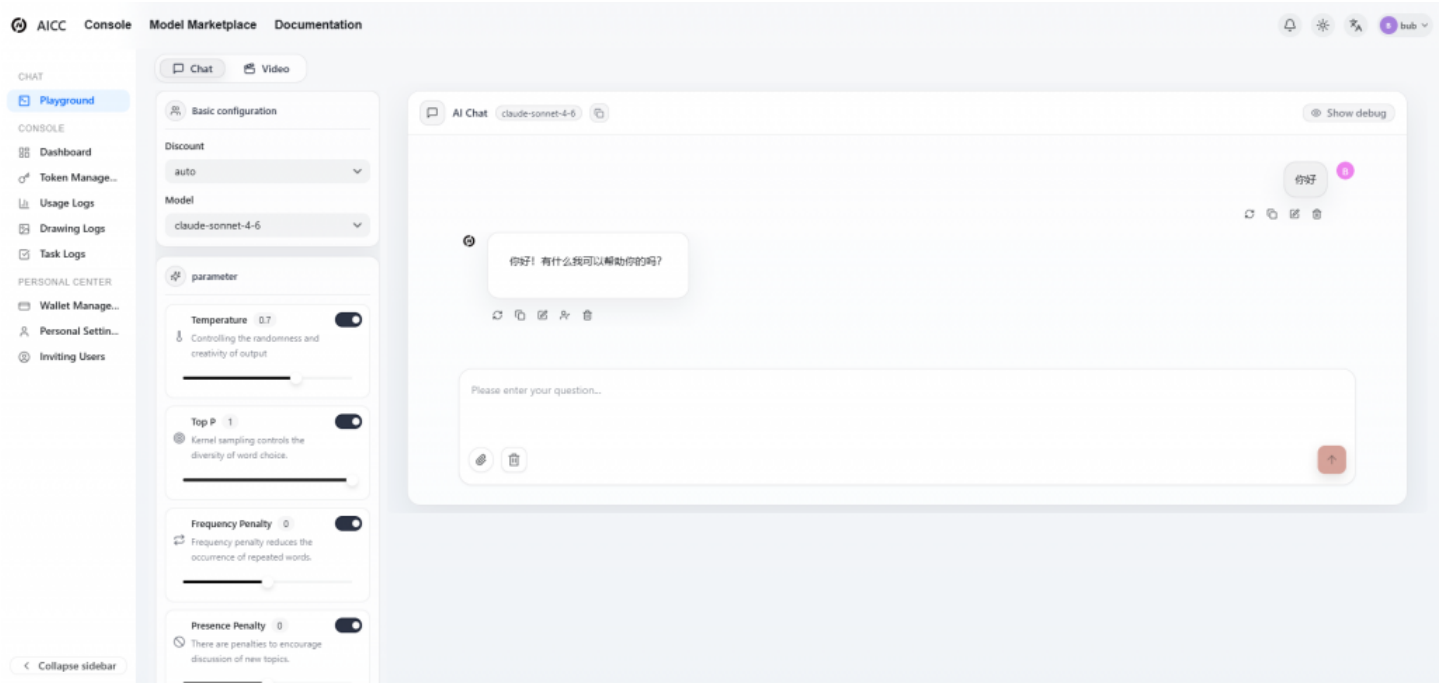
Enterprise AI Costs Drop 67 Percent as Companies Shift to Multi-Model Aggregation Platforms



Singapore, Singapore Jul 23, 2026 (Issuewire.com) - New data from 2.4 billion API calls reveals enterprises save up to 80 percent by routing workloads across multiple AI models

Enterprise spending on artificial intelligence APIs is falling rapidly as companies abandon single-provider strategies in favor of multi-model aggregation platforms, according to data released by AI.cc, a unified AI API platform processing more than 90 million daily requests across 300-plus models.

The shift reflects a broader industry transformation. Global AI API market revenue reached $64.41 billion in 2025 and is projected to surpass $900 billion by 2035, according to Precedence Research. As adoption scales, enterprises are discovering that routing every task to a single premium model — a common practice just 12 months ago — is no longer economically viable.

Data from AI.cc's 2026 AI API Infrastructure Report, drawn from anonymized analysis of 2.4 billion API calls processed between January and April 2026, shows enterprise token costs declined 67 percent year-over-year. The effective blended cost per million tokens dropped from $18.40 to $6.07 during the period.

The primary driver is intelligent task routing. In Q1 2025, 73 percent of enterprise token volume was directed to the two most expensive model tiers. By Q1 2026, that figure fell to 31 percent, with the remaining 69 percent distributed across mid-tier and cost-efficient models matched to task complexity.

"Enterprise teams were sending simple classification tasks, customer support queries, and structured data extraction through frontier models because that was what they had integrated," said the AI.cc research team. "The data shows that intelligent routing alone accounts for an estimated 34

percentage points of the total 67 percent cost reduction."

The economic case is significant. For an organization processing 2 billion tokens per month — equivalent to a mid-sized SaaS company running AI-powered customer support and document analysis — the cost difference between single-model and multi-model approaches amounts to approximately $295,920 in annual savings.

Open-source and open-weight models have accelerated the trend. These models captured 38 percent of enterprise token volume in Q1 2026, up from 11 percent in Q1 2025 — a 245 percent share increase in 12 months. The shift was triggered by aggressive pricing from models such as DeepSeek V4-Flash, which launched in April at $0.14 per million input tokens.

Average models per enterprise account rose to 4.7 in Q1 2026, up from 2.1 in Q1 2025. New adopters entering the platform averaged 5.3 models within their first 30 days, indicating that multi-model architecture has become the default rather than the exception.

Beyond cost, enterprises report faster deployment cycles. 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, headquartered in Singapore, operates as a unified aggregation layer across providers including OpenAI, Google, Anthropic, xAI, DeepSeek, Alibaba, ByteDance, and MiniMax. The platform's position as a high-volume aggregator enables below-retail pricing, with effective discounts averaging 23 percent versus direct provider rates and reaching 35 to 40 percent for highest-volume enterprise accounts.

"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."

Enterprises that adopted multi-model routing on the AI.cc platform reported median cost reductions of 71 percent versus equivalent single-provider deployments. The top quartile achieved reductions exceeding 80 percent while maintaining or improving output quality on customer-defined evaluation metrics.

The findings align with broader market 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. Grand View Research estimates North America held 38.8 percent of global market share in 2024, with cloud-based API deployments accounting for the largest revenue segment.

For business leaders evaluating AI deployment strategies, the data suggests that cost barriers which made large-scale adoption appear risky in 2024 and 2025 have shifted substantially. The combination of multi-model routing, prompt caching, and aggregated pricing has redefined the economics of enterprise AI infrastructure.

AI.cc offers a free tier with API access for new users. For more information, visit www.ai.cc.

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. AI.cc offers enterprise API plans, AI application development services, and GEO-optimized PR solutions. For more information, visit www.ai.cc.

Media Contact

AICC

*********@ai.cc

https://www.ai.cc

Source : AICC PTE. LTD.