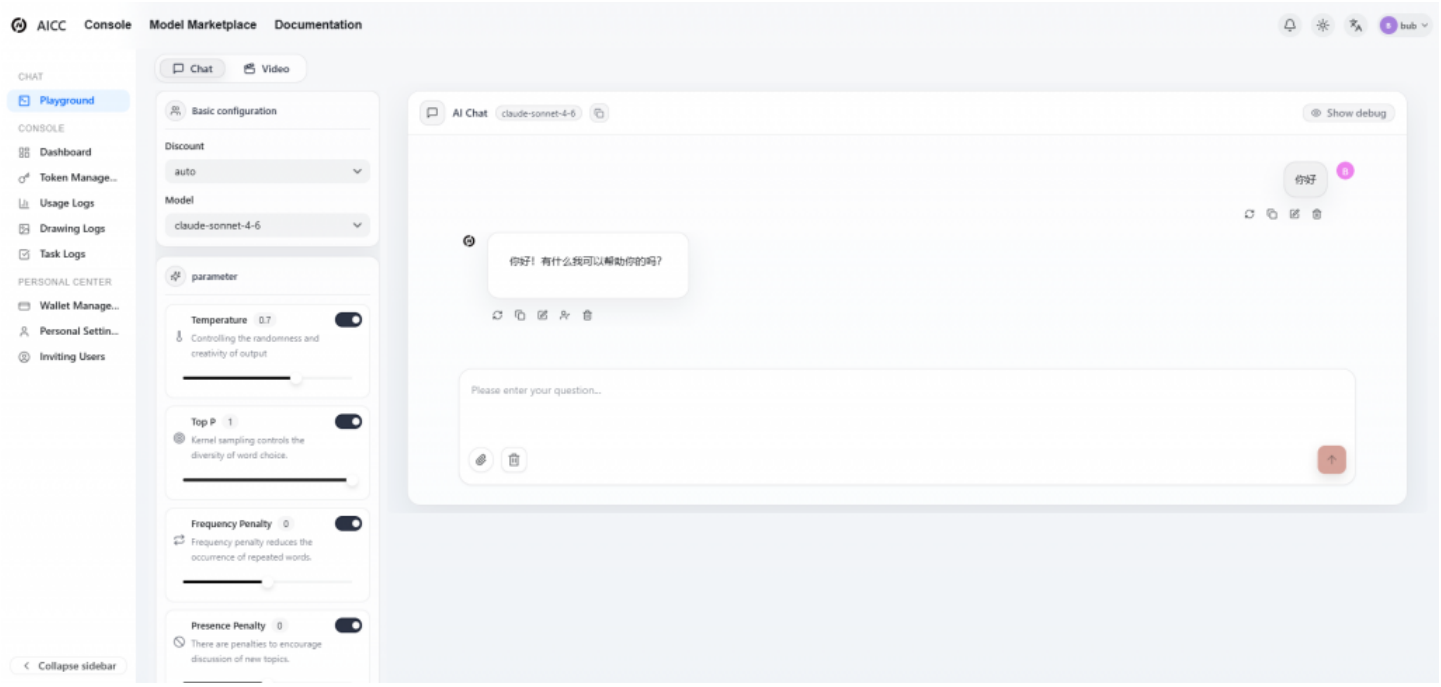
AICC Report: Enterprise Token Costs Drop 67% Year-Over-Year as Multi-Model AI Adoption Hits Record High



Singapore, Singapore May 8, 2026 (Issuewire.com) - Comprehensive analysis of 2.4 billion API calls across 8,000+ developers and enterprises reveals the definitive shift in how production AI is built, deployed, and priced in 2026 — with multi-model routing emerging as the dominant architecture and open-source models capturing 38% of enterprise token volume for the first time

AI.cc, the Singapore-based unified AI API aggregation platform, today released its 2026 AI API Infrastructure Report, the most comprehensive analysis of enterprise AI API usage patterns, cost benchmarks, and model adoption trends published to date. The report, drawn from anonymized data across more than 2.4 billion API calls processed through the AI.cc platform between January 1 and April 30, 2026, documents a structural transformation in enterprise AI deployment that is reshaping the economics, architecture, and competitive dynamics of AI-powered product development.

The headline finding is unambiguous: enterprise token costs fell **67% year-over-year** in the twelve months ending April 30, 2026, driven by three simultaneous forces — the collapse in open-source model pricing led by DeepSeek V4 and Qwen 3.6-Plus, the widespread adoption of intelligent multi-model routing strategies that direct traffic to the most cost-efficient qualified model for each task, and the compounding effect of AI.cc's aggregation-scale pricing advantages over direct retail API rates. For enterprise teams that adopted multi-model routing strategies on the AI.cc platform, the median cost reduction versus equivalent single-provider deployments reached **71%**, with the top quartile achieving reductions exceeding 80% while maintaining or improving output quality on customer-defined evaluation metrics.

The report encompasses data from more than 8,000 active developer and enterprise accounts across

47 countries, spanning use cases in software development, legal technology, financial services, e-commerce, healthcare administration, content production, customer experience, and internal knowledge management. It represents the first large-scale empirical analysis of multi-model API adoption patterns in production environments rather than benchmark conditions.

Section 1: The Cost Collapse — How Token Prices Fell 67% in Twelve Months

The most consequential finding in the 2026 AI API Infrastructure Report is the scale and pace of enterprise token cost reduction. Twelve months ago, the effective blended cost per million tokens for enterprise AI workloads — weighted by the actual mix of input and output token consumption across real production applications — averaged $18.40 across the AI.cc platform. As of April 30, 2026, that figure stands at **$6.07**, a 67% reduction representing one of the fastest cost deflations ever recorded in enterprise software infrastructure.

Three distinct mechanisms drove this reduction, each operating independently and compounding in combination.

Mechanism 1: Open-Source Model Price Disruption. The release of DeepSeek V4-Flash at $0.14 per million input tokens and $0.28 per million output tokens on April 24, 2026 — following DeepSeek V3.2's earlier disruption at $0.28 per million input tokens — established a new price floor for frontier-adjacent AI capability that forced a broad repricing across the model ecosystem. Qwen 3.5's 9B variant at $0.10 per million input tokens, Gemma 4's Apache 2.0 open-weight models at effectively zero self-hosted cost, and GLM-5.1's $3 per month subscription pricing for coding-heavy workflows collectively redefined what developers and enterprises should expect to pay for capable AI inference. AI.cc's platform data shows that open-source and open-weight models captured **38% of enterprise token volume** in Q1 2026, up from 11% in Q1 2025 — a 245% share increase in twelve months.

Mechanism 2: Multi-Model Routing Adoption. As developers moved from single-model to multi-model architectures, they stopped over-provisioning frontier model capacity for tasks that did not require it. AI.cc's platform data reveals that in Q1 2025, 73% of enterprise token volume was routed to the two most expensive model tiers — the equivalent of routing customer support queries and simple classification tasks through Claude Opus or GPT-4 simply because those were the models the team had integrated. By Q1 2026, that figure had fallen to 31%, with the remaining 69% distributed across mid-tier and cost-efficient models matched to task complexity. The aggregate cost impact of this routing optimization, holding model mix constant, accounts for an estimated 34 percentage points of the total 67% cost reduction.

Mechanism 3: Aggregation-Scale Pricing. AI.cc's position as a high-volume aggregator continued to compound cost advantages for customers. By routing consolidated volume across all supported models, AI.cc maintains below-retail pricing on the majority of its model catalog. The effective discount versus direct retail API pricing across the platform averaged **23%** in Q1 2026, representing a baseline cost advantage before routing optimization is applied. For the highest-volume enterprise accounts, effective discounts on specific model categories reached 35–40% versus direct provider retail rates.

Section 2: Multi-Model Adoption — From Experiment to Default Architecture

The 2026 AI API Infrastructure Report documents that multi-model deployment has crossed from experimental to default architecture across virtually all enterprise customer segments.

Average models per enterprise account reached 4.7 in Q1 2026, up from 2.1 in Q1 2025 — a 124%

increase in the diversity of models actively called per organization within a single year. Among accounts onboarded in Q1 2026, the figure reached 5.3 within the first 30 days, indicating that new adopters are entering the multi-model paradigm from day one rather than transitioning from single-model origins.

The distribution of model usage across enterprise accounts reveals a consistent architectural pattern that the report terms the **Tiered Intelligence Stack** — now the dominant deployment pattern across 64% of enterprise accounts by token volume:

A **cost-efficiency tier** handles the majority of request volume — typically 55–70% of all API calls — using models priced below $0.50 per million input tokens. DeepSeek V4-Flash, Qwen 3.5 9B, Gemma 4 12B, and Mistral Small 4 dominate this tier. Tasks routed here include intent classification, simple query resolution, content filtering, structured data extraction from well-formed inputs, and high-volume batch processing where latency is not a constraint.

A **mid-performance tier** handles the second-largest share of volume — typically 20–30% of API calls — using models priced between $0.50 and $5.00 per million input tokens. Claude Sonnet 4.6, Gemini 3.1 Flash, GPT-5.4, and DeepSeek V4-Pro are the most commonly called models in this tier. Tasks include standard response generation, moderate-complexity reasoning, document summarization, and customer-facing interactions that require quality above the cost-efficiency tier but do not justify frontier model pricing.

A **frontier tier** handles the smallest share of volume by request count — typically 5–15% of API calls — but the most complex and highest-value tasks. Claude Opus 4.7, GPT-5.5, and Gemini 3.1 Pro dominate this tier. Tasks include complex multi-step reasoning, long-context document analysis, sophisticated coding agent tasks, and any interaction where output quality directly and measurably impacts business outcomes. The defining characteristic of well-implemented Tiered Intelligence Stacks is that the frontier tier is reserved strictly for tasks that genuinely require frontier capability — not used as a default fallback for any query the routing logic cannot confidently classify.

Enterprises that fully implemented the Tiered Intelligence Stack pattern in 2026 — with deliberate routing logic across all three tiers — achieved a median blended cost per million tokens of **$2.31**, compared to $18.40 for equivalent workloads routed entirely through frontier models twelve months ago. This 87.4% reduction in effective cost per intelligence unit is the defining economic story of enterprise AI deployment in 2026.

Section 3: Model Rankings — What Enterprises Are Actually Using

Beyond aggregate statistics, the 2026 AI API Infrastructure Report provides the first empirical ranking of AI model usage in production enterprise environments — distinct from benchmark leaderboards and developer surveys, which measure stated preferences rather than actual deployment patterns.

Top 10 Models by Enterprise Token Volume, Q1 2026:

- **Claude Sonnet 4.6** (Anthropic) — Most-called model by token volume across the platform, reflecting its balance of performance, instruction-following quality, and mid-tier pricing. Dominant in customer-facing interaction and document processing workloads.
- **DeepSeek V3.2** (DeepSeek) — Second by volume, reflecting its position as the established cost-efficiency workhorse before V4's April launch. $0.28 per million input tokens with frontier-adjacent performance on most task categories.
- **GPT-5.4** (OpenAI) — Third by volume, with particular strength in tool-use-heavy agent

workflows and any integration where OpenAI's native computer use capabilities provide a measurable advantage.

- **Gemini 3.1 Flash** (Google) — Fourth, driven by multimodal workloads and its competitive $1.00 per million input token pricing for a mid-tier model with 1 million token context and vision capability.
- **Qwen 3.5 9B** (Alibaba) — Fifth, reflecting the extraordinary price-to-performance ratio at $0.10 per million input tokens and 81.7% GPQA Diamond score. Dominant in Asian-language workloads and high-volume classification tasks.
- **Claude Opus 4.6** (Anthropic) — Sixth, concentrated in high-value reasoning and coding agent workloads where output quality is the primary optimization target.
- **DeepSeek V4-Flash** (DeepSeek) — Seventh, having onboarded rapidly following its April 24 launch. Trajectory suggests it will reach top-three position by end of Q2 2026.
- **Llama 4 Maverick** (Meta) — Eighth, with strong adoption among teams prioritizing open-weight models for data sovereignty and self-hosting flexibility alongside API access for burst capacity.
- **Gemini 3.1 Pro** (Google) — Ninth by volume, with concentrated usage in scientific reasoning, multimodal analysis, and any workload where its leading GPQA Diamond score of 94.3% is the determining selection factor.
- **GLM-5.1** (Zhipu AI) — Tenth, reflecting significant adoption for long-horizon coding agent tasks and Chinese-language applications. MIT license and competitive pricing make it the dominant open-source choice for enterprise coding workloads.

Notable trends in the model rankings: Open-source and open-weight models occupy four of the top ten positions by token volume — a figure that would have been implausible twelve months ago when the equivalent list was dominated entirely by OpenAI and Anthropic proprietary models. The diversity of provider origin across the top ten — US, Chinese, and European models represented — confirms that enterprise AI procurement has become genuinely global in scope rather than concentrated among a handful of Silicon Valley providers.

Section 4: Agentic AI — The Fastest-Growing Workload Category

The 2026 AI API Infrastructure Report identifies agentic AI applications as the fastest-growing workload category on the platform by both request count growth rate and token volume growth rate, with agent-pattern API calls growing **680% year-over-year** in Q1 2026.

Agent-pattern API calls — defined as sequences of API requests exhibiting multi-turn reasoning, tool call invocation, result processing, and iterative refinement within a single coordinated workflow — represented 41% of new integration use cases registered on the platform in Q1 2026, up from 18% in Q1 2025.

The report identifies five dominant agent architectures deployed in production across the AI.cc platform in 2026:

The Research and Synthesis Agent uses a frontier reasoning model for source evaluation and synthesis, a fast mid-tier model for parallel document retrieval and summarization, and an embedding model for semantic search across large document collections. Median model count: 3.8 per workflow. Primary sectors: legal, financial services, academic, consulting.

The Software Development Agent orchestrates a frontier coding model for complex implementation and architecture decisions, a mid-tier model for code review and documentation, a specialized embedding model for codebase semantic search, and a fast model for test case generation and simple

refactoring. Median model count: 4.2 per workflow. Primary sectors: software development tooling, developer productivity, internal engineering automation.

The Customer Experience Agent routes customer interactions through a fast classification model for intent detection, a mid-tier model for standard response generation, a frontier model for escalation handling and complex problem resolution, and a multilingual model for non-English language interactions. Median model count: 3.1 per workflow. Primary sectors: e-commerce, fintech, SaaS support, telecommunications.

The Document Processing Agent uses a vision-capable model for document ingestion and OCR, a reasoning model for structured data extraction and validation, an embedding model for document classification and routing, and a fast model for high-volume batch processing. Median model count: 4.6 per workflow. Primary sectors: healthcare, legal, financial services, government.

The Content Production Agent orchestrates a research model for factual grounding, a generation model for draft production, a reasoning model for editorial quality evaluation, and specialized models for SEO optimization and multilingual localization. Median model count: 3.7 per workflow. Primary sectors: media, marketing technology, e-commerce content, publishing.

Across all agent architectures, the report finds that **OpenClaw-based implementations** — using AI.cc's purpose-built multi-model agent orchestration framework — achieved meaningfully lower rates of production incidents attributable to model failures, rate limit errors, and context management issues compared to custom-built equivalent implementations.

Section 5: Geographic Expansion — AI API Goes Global

The geographic distribution of AI.cc's customer base in Q1 2026 reflects the globalization of AI-powered product development, with meaningful growth across every major region and the Asia-Pacific market reinforcing its position as the largest single market for unified AI API infrastructure.

Asia-Pacific remains the largest region by customer count, representing 44% of active accounts. Singapore, India, Australia, Japan, South Korea, and Indonesia are the top six markets by account volume within the region. The concentration of Chinese-origin model usage is highest in Asia-Pacific, with DeepSeek, Qwen, GLM, and Doubao models collectively representing 52% of token volume in the region — reflecting both language requirements and cost optimization priorities of Asia-Pacific developers.

Europe was the fastest-growing region in Q1 2026, with new account activations growing 290% year-over-year. Germany, the Netherlands, France, and the United Kingdom led European growth. European enterprise customers exhibited the strongest preference for open-source and open-weight models — Llama 4, Gemma 4, Mistral Small 4, and GLM-5.1 collectively represented 61% of European enterprise token volume — driven by data sovereignty requirements, GDPR compliance considerations, and the operational advantages of self-hosted inference for sensitive workload categories.

North America grew 180% year-over-year in Q1 2026, with AI.cc's model breadth and pricing cited most frequently as primary selection drivers over US-based alternatives. Canadian and US enterprise customers showed the highest average model diversity at 5.9 distinct models per account, reflecting more mature multi-model strategy implementation.

Middle East and Africa grew 340% year-over-year from a smaller base, with the United Arab

Emirates, Saudi Arabia, Israel, Nigeria, and South Africa leading regional adoption. Multilingual capability across Arabic, Hebrew, Swahili, and other regional languages was the primary differentiation factor cited in this region.

Latin America grew 220% year-over-year, led by Brazil, Mexico, Colombia, and Argentina. Cost efficiency and open-source model access were the primary drivers, with DeepSeek V3.2 and Qwen 3.5 9B representing over 60% of Latin American token volume.

Section 6: Enterprise Benchmark Data — What Teams Are Paying in 2026

The report provides the first publicly available benchmark dataset for enterprise AI API costs in 2026, segmented by workload type, model tier, and optimization level. These figures represent observed outcomes across the AI.cc platform in Q1 2026 and are intended to provide development teams with a realistic baseline for evaluating their own AI infrastructure cost efficiency.

Effective cost per million blended tokens by optimization level:

Unoptimized single-frontier-model deployment (equivalent to routing all traffic through Claude Opus 4.7 or GPT-5.5 at retail pricing): **$17–28 per million blended tokens**

Single-provider deployment with basic model selection (using a mid-tier model as default, frontier only for flagged requests): **$4–8 per million blended tokens**

Multi-provider deployment with manual routing logic (developer-defined rules directing traffic across three to five models): **$2–5 per million blended tokens**

AI.cc platform with intelligent routing recommendations applied (platform-assisted routing across the full model catalog): **$1.20–3.40 per million blended tokens**

Section 7: Looking Ahead — H2 2026 Projections

Based on Q1 2026 platform data and the model release pipeline visible as of publication date, the 2026 AI API Infrastructure Report projects the following trends for the second half of 2026.

Token costs will continue to fall, but at a decelerating rate. The steepest cost reductions — driven by the shift from single-model to multi-model architectures and the initial disruption of open-source pricing — are largely captured. Continued price reduction in H2 2026 is projected at 20–30% from current levels, driven primarily by further open-source model releases and competitive pressure on mid-tier proprietary model pricing. The era of 60–70% annual cost reductions is likely transitioning to a more moderate 20–35% annual reduction trajectory.

Agentic workloads will surpass conversational workloads in token volume. At the current growth trajectory, agent-pattern API calls are projected to exceed conversational and single-turn inference workloads in total token volume by Q3 2026. This transition will further accelerate multi-model adoption, as agentic architectures are inherently multi-model by nature and cannot be efficiently served by single-provider strategies.

Open-source model share will reach 50% of enterprise token volume by year-end. From 38% in Q1 2026, open-source and open-weight model adoption is projected to reach 48–52% of enterprise token volume by Q4 2026, driven by the continued improvement of open-source model quality, the

anticipated release of DeepSeek V4 at full production scale, and the maturation of self-hosted inference infrastructure that reduces the operational overhead of running open-weight models.

Geographic diversification of model origin will accelerate. The Q1 2026 data showing Chinese-origin models at 38% of enterprise token volume and open-source European models (Mistral family) at 8% points to a long-term trajectory in which no single country's AI labs dominate enterprise deployment. H2 2026 is expected to see continued share gains from Chinese, European, and emerging-market AI labs as their model quality continues to converge with US-based frontier providers.

About AI.cc

AI.cc is a unified AI API aggregation platform headquartered in Singapore, providing developers and enterprises with seamless access to 300+ AI models — including GPT-5.5, Claude Opus 4.7, Gemini 3.1 Pro, DeepSeek V4 and more — through a single OpenAI-compatible API. The platform supports text, image, video, voice, code, embedding, and OCR model categories. Additional offerings include the OpenClaw AI agent framework, enterprise plans with SLA guarantees, AI application development services and AI Translator API

Register for a free API key: www.ai.cc Platform documentation: docs.ai.cc

Media Contact

AICC

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

https://www.ai.cc

Source : AICC PTE. LTD.

See on IssueWire