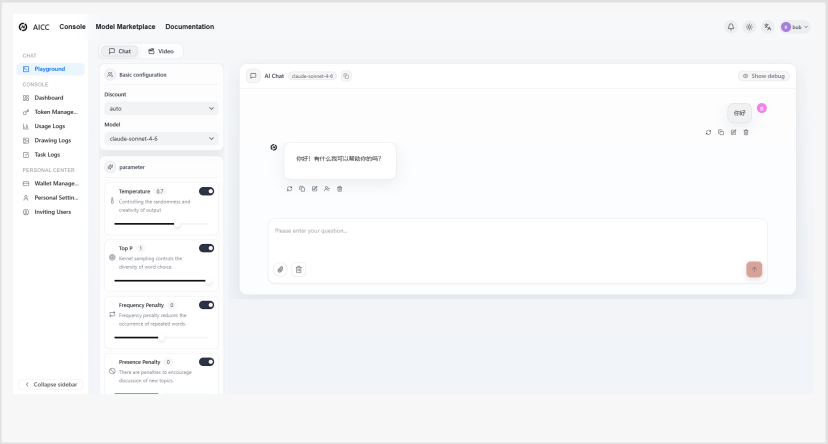
AI.cc Now Supports 500+ Hugging Face Open-Source Models via Unified API

SINGAPORE, SINGAPORE, SINGAPORE, July 3, 2026 /EINPresswire.com/ -- PRESS RELEASE
FOR IMMEDIATE RELEASE



Date: May 30, 2026

Contact: press@ai.cc | www.ai.cc

AI.cc Now Supports 500+ Hugging Face Open-Source Models via Unified API, Eliminating Self-Hosting Barriers for Enterprise Teams

Singapore-based platform adds full open-source model catalog access — including Llama 4, Mistral, Falcon, GLM-5.1 and 500+ community models — through existing OpenAI-compatible endpoint, with no self-hosting infrastructure required

SINGAPORE, May 30, 2026 — AI.cc, the Singapore-based unified AI API aggregation platform, today announced that enterprise customers can now access 500+ open-source models from the Hugging Face Hub through AI.cc's unified API — eliminating the GPU infrastructure, DevOps overhead, and model management complexity that has historically prevented enterprise teams from deploying open-source models at production scale.

The expanded model catalog, now totaling 800+ models across proprietary and open-source categories, is available immediately through AI.cc's existing OpenAI-compatible endpoint. No new SDK integration, no separate Hugging Face Inference API account, and no self-hosting infrastructure is required. Enterprise teams access Llama 4, Mistral Large 3, GLM-5.1, DeepSeek V4, Gemma 4, and hundreds of additional open-source models using the same API key and the same call structure they already use for Claude Opus 4.7, GPT-5.5, and Gemini 3.1 Pro.

"Open-source models have crossed the capability threshold where they belong in enterprise production deployments — not just research environments," said an AI.cc spokesperson. "The remaining barrier has been operational, not technical. Managing GPU infrastructure, quantization, container orchestration, and model updates is a full-time engineering job that most product teams cannot afford to maintain alongside their core product work. We have removed that barrier entirely."

Why Open-Source Models Now Belong in Enterprise Production

The case for open-source model access in enterprise AI deployments has strengthened substantially in the first half of 2026, driven by three developments that together close the remaining gap between open-source capability and enterprise requirements.

Benchmark convergence with proprietary frontier models. Six months ago, proprietary models held a commanding lead over open-source alternatives on enterprise-relevant benchmarks. That lead has narrowed to single-digit percentage points on most evaluation categories. GLM-5.1 reaches 94.6% of Claude Opus 4.6's coding performance on SWE-bench. MiniMax M2.5 scores 80.2% on SWE-bench Verified — within 0.6 points of Claude Opus 4.6. Mistral Small 4 outperforms GPT-OSS 120B on LiveCodeBench. Llama 4 Maverick beats previous-generation frontier models across major benchmarks while running on a single H100.

For the majority of enterprise workload categories — document processing, classification, summarization, standard response generation, code assistance — open-source models now deliver output quality that enterprise users cannot distinguish from proprietary frontier model output. Reserving proprietary frontier models for the minority of tasks where their marginal capability advantage is genuinely consequential, and routing the remainder to open-source models, is the rational enterprise strategy in 2026.

Licensing that supports commercial deployment. Early open-source AI models carried licensing restrictions that made commercial deployment legally ambiguous. The 2026 model generation has largely resolved this. Llama 4 ships under Meta's commercial license. Gemma 4 is Apache 2.0. GLM-5.1 is MIT. Mistral models carry Apache 2.0 licensing. DeepSeek V4 is MIT. The open-source models now available through AI.cc's expanded catalog are commercially deployable without licensing restrictions that would complicate enterprise procurement.

Cost differential that justifies architectural investment. AI.cc's platform data shows that enterprises routing appropriate workloads to open-source models reduce blended token costs by 40–65% compared to equivalent proprietary-only deployments. At enterprise processing volumes, this differential reaches hundreds of thousands of dollars annually — sufficient to justify the routing architecture investment required to implement open-source model access, even before accounting for the elimination of self-hosting infrastructure costs.

The 500+ Model Catalog: What Is Now Available

The expanded AI.cc catalog includes 500+ curated open-source models selected based on download volume, benchmark performance, license permissiveness, and enterprise deployment suitability. The catalog spans every major open-source model family available as of May 2026.

Foundation and reasoning models: The complete Llama 4 family including Scout (10M token context, commercial license) and Maverick (multimodal, single-H100 deployable). Mistral Large 3, Mistral Small 4, and Devstral 2 (123B coding specialist). The full Qwen 3.x series from 1.5B to 480B, including Qwen 3 Coder 480B and Qwen 3.5 at $0.10 per million input tokens. Google's Gemma 4 family across all four variants. Zhipu AI's GLM-5.1 (744B MoE, MIT license) and GLM-5V-Turbo. The complete DeepSeek open-weight catalog including V4-Pro, V4-Flash, V3.2, and R1. Falcon 3 and Falcon 2 from the Technology Innovation Institute. Arcee Trinity (400B, Apache 2.0).

Specialized and fine-tuned models: 300+ community and organization fine-tuned models across

biomedical, legal, financial, multilingual, and code-specific categories. Models fine-tuned for specific regulatory compliance domains, clinical documentation, financial report generation, and Southeast Asian language coverage that proprietary frontier APIs do not support natively.
Embedding and retrieval models: High-performance open-source embedding models for RAG pipelines, semantic search, and document classification — including models specifically optimized for multilingual embedding across Asian language pairs where proprietary embedding models show degraded performance.

Technical Implementation: One Endpoint, All Models
Accessing open-source models through AI.cc requires no changes to existing integration code beyond the model parameter. Open-source models from the expanded catalog use a standardized naming format that distinguishes them from proprietary models while maintaining identical call structure:

```
python# Existing proprietary model call — unchanged
response = client.chat.completions.create(
model="claude-opus-4-7",
messages=[{"role": "user", "content": complex_prompt}]
)

# Open-source model — identical structure, one parameter change
response = client.chat.completions.create(
model="hf/meta-llama/llama-4-scout",
messages=[{"role": "user", "content": standard_prompt}]
)

# Cost-efficient open-source for high-volume classification
response = client.chat.completions.create(
model="hf/qwen/qwen3.5-9b",
messages=[{"role": "user", "content": classification_prompt}]
)
```

All three calls use the same client instance, the same API key, and return responses in identical format. Token consumption across proprietary and open-source models is consolidated in AI.cc's unified billing dashboard, providing a single cost view across the full model catalog.
AI.cc's OpenClaw agent framework supports open-source models identically to proprietary models, enabling multi-step agent workflows that route dynamically between open-source and proprietary models at the task level. A single agent workflow can use GLM-5.1 for coding subtasks, Llama 4 Scout for long-context document retrieval, Mistral Small 4 for classification steps, and Claude Opus 4.7 for high-stakes reasoning — all coordinated through OpenClaw without any framework-level distinction between model categories.

Multi-Model Routing Across Open-Source and Proprietary Models
The practical value of unified access to open-source and proprietary models through a single API is most apparent in the Tiered Intelligence Stack architectures that represent the dominant

enterprise deployment pattern in 2026.
A representative enterprise document processing deployment using the expanded AI.cc catalog might route as follows: document ingestion and OCR to Gemma 4 12B (Apache 2.0, low cost), content classification and extraction to Qwen 3.5 9B ($0.10/M input), standard summarization and drafting to Mistral Large 3 (Apache 2.0, strong European language performance), complex reasoning and risk analysis to Claude Opus 4.7 (frontier quality for high-stakes steps), and final output formatting to GLM-5.1 for coding-adjacent structured outputs.
This architecture routes approximately 70% of token volume through open-source models priced below $0.50 per million input tokens, reserving proprietary frontier model capacity for the 30% of workflow steps where frontier capability is genuinely necessary. The blended cost across the full workflow reaches $0.35–0.65 per million tokens — compared to $5.00–18.00 per million tokens for equivalent workflows routed entirely through frontier proprietary models.

Pricing and Availability
Open-source models from the expanded catalog are priced based on underlying inference cost plus AI.cc's standard margin. The majority of the catalog is priced below $0.50 per million input tokens, with higher-capability large models priced between $0.80 and $2.00 per million input tokens.
Free-tier access includes evaluation quota for all 500+ open-source models. Enterprise customers requiring dedicated inference capacity with SLA guarantees for specific models can access dedicated endpoints through AI.cc's enterprise plan.
The complete open-source model catalog, pricing table, benchmark data, and integration documentation are available at docs.ai.cc/open-source-models. Enterprise inquiries: www.ai.cc/enterprise-plans.

About AI.cc
AI.cc is a unified AI API aggregation platform headquartered in Singapore, providing developers and enterprises with access to 800+ AI models — including GPT-5.5, Claude Opus 4.7, Gemini 3.1 Pro, DeepSeek V4, and 500+ Hugging Face open-source models — through a single OpenAI-compatible API. Additional offerings include the OpenClaw AI agent framework, enterprise SLA plans, AI Translator API, and AI Web Scraping API.
Open-source model catalog: docs.ai.cc/open-source-models

Free API access: www.ai.cc

Enterprise plans: www.ai.cc/enterprise-plans

AICC
AICC
+44 7716 940759
support@ai.cc