WHITE PAPER · OPEN MODELS · JULY 2026

BRIEF

Open models: *the co-frontier*

In roughly eighteen months, open-weight models went from a promising second tier to a genuine co-frontier: months behind the best closed models, but ahead in certain areas like maths, agentic tool use and, importantly, price. We mapped who builds them, how they run, and the economics that decide build-versus-buy. What follows is the overview, alongside the detailed model explorer.

3–4 months

how far the best open-weight models trail the closed state of the art (Epoch AI)

~3.3%

how much the best closed model outscores the best open one in head-to-head human preference votes (Stanford HAI AI Index, 2026)

10–50×

cheaper per token: open-weight serverless APIs vs closed frontier APIs (mid-2026 price sheets)

~2.9M

models on Hugging Face by mid-2026: roughly doubling year on year

DEFINITION

Definition and categories

Open weights

The trained parameters are downloadable and runnable; the training data and code usually are not. The overwhelming majority of “open” models live here.

Llama (Meta) · Qwen (Alibaba) · DeepSeek · Mistral · gpt-oss (OpenAI)

Fully open

Weights *plus* training data, code and intermediate checkpoints: the model is reproducible end to end. Rare, and mostly research-led.

OLMo (Ai2) · Pythia (EleutherAI) · Nemotron (NVIDIA, close behind)

The licence axis

A separate question: from genuinely permissive terms to user caps, output restrictions and non-commercial clauses attached to otherwise-open weights.

Apache-2.0 · MIT · Llama Community · Gemma Terms · CC-BY-NC

What *open* means

"Open" is a combination of independent factors, not a single property. It's important to decide how reproducible and auditable the model is, what its commercial and legal terms allow, and whether you self-host or consume an API. A model can carry a permissive licence yet be impossible to reproduce, or ship with full data and code yet forbid commercial use.

The licences sort into four tiers. **Permissive** (Apache-2.0 / MIT) is now where the frontier lives: Qwen, DeepSeek, gpt-oss, GLM, Granite and ERNIE all ship under one or the other. **Restrictive open weights** attach conditions: Meta's Llama Community Licence caps usage above 700 million monthly active users; xAI's Grok-2 terms forbid using outputs to train other models. **Non-commercial** (Cohere's Command and Aya, CC-BY-NC) is open to study but not to sell. **Fully open** (Ai2's OLMo, with NVIDIA's Nemotron close behind) publishes the data and recipes too. The centre of gravity has moved decisively toward the first tier: Meta's licence, once the reference point for "open", now reads as comparatively restrictive.

THE ARGUMENT

The case *for* — and the case *against*

For

Control. Data residency, privacy, compliance and air-gapped deployment: weights on your own hardware, in your own jurisdiction. A 2025 survey of ~1,500 IT leaders ranked data privacy the top AI-adoption barrier (53%).

Cost. No per-token markup at scale and the commodity tier is deflating. Equivalent-capability inference falls roughly **10× a year** (a16z); Epoch AI measures a median ~50× across benchmarks.

Customisation. Full fine-tuning and weight surgery. For a specific task: code completion, transcription, embeddings, a single language. A fine-tuned open model is frequently the best cost-adjusted choice, closed models included.

Latency and offline. Sub-20ms local inference against 50–200ms cloud round-trips; edge and on-device operation with no connectivity requirement at all.

Vendor independence. No lock-in, no unilateral deprecation: a model you host cannot be switched off, repriced or retired from outside.

Against

The residual frontier gap. The hardest generalist reasoning and agentic-assistant work still favours the closed anchors. GPT-5, Claude Opus 4.5 (~81% SWE-bench Verified), Gemini 3 Pro, which benefit from massive reinforcement learning on real user feedback and proprietary agent infrastructure.

Operational burden. Self-hosting means GPUs, serving stacks, quantisation choices and on-call rotas. A GPU costs the same idle or busy: at low utilisation, effective cost can run ~10× the theoretical minimum.

Provenance and geopolitics. The strongest open models are disproportionately Chinese; government-device bans on DeepSeek span US federal agencies, Australia, Italy and Taiwan, and the regulatory environment around model origin is fluid.

The enterprise counter-signal. An a16z survey of enterprise CIOs (mid-2025) put open source at ~13% of workloads: **down** from ~19% six months earlier with OpenAI dominant in production.

Benchmark noise. Invalid-question rates of 2–42% across popular benchmarks, near-saturated maths tests and vendor best-case configurations.

THE SHIFT

Who builds the *open frontier*

The centre of gravity moved to China. The strongest open-weight models now come predominantly from Chinese labs: DeepSeek, Alibaba's Qwen, Moonshot's Kimi, Zhipu/Z.ai's GLM and MiniMax, with Baidu (ERNIE 4.5: a notable open pivot) and Tencent (Hunyuan). The single clearest structural change of 2025 was **Qwen overtaking Llama** as the most-used open family: 100,000+ fine-tuned derivatives on Hugging Face, more than Google and Meta combined, and roughly **700 million cumulative downloads** by January 2026 (Xinhua).

Production traffic tells the same story. On OpenRouter, the combined token share of US providers fell from roughly **70% in mid-2025 to about 30%** by mid-2026, with DeepSeek at times the single largest provider of token volume.

Meta pioneered the movement but appears to be retreating. Llama 4 (April 2025) fell behind the frontier it created, the flagship "Behemoth" never shipped, and in April 2026 Meta's Superintelligence Labs released a proprietary model with no weights. *Whether Llama 4 is Meta's last open release is rumoured, not confirmed.* Filling the gap, **OpenAI re-entered open weights** in August 2025 with gpt-oss (Apache-2.0). Its first open release since GPT-2. Google ships Gemma, Microsoft ships Phi, NVIDIA's near-fully-open Nemotron line has become the strongest US open model, IBM targets the enterprise with Granite, and Ai2's OLMo remains fully-open.

Outside the two poles: **Mistral** (FR) is Europe's flagship, with permissive *Small* models plus code (Codestral, Devstral), reasoning (Magistral) and audio (Voxtral) lines; the UAE's TII ships Falcon and the strongest Arabic models. Underneath it all sits scale: Hugging Face passed **2 million public models in 2025** and ~2.4–2.9M by mid-2026 with adoption extremely concentrated. The top 0.01% of models take roughly half of all downloads, and ~92% of downloads go to sub-1B-parameter models: embeddings, edge and pipeline use, not frontier chat.

CAPABILITY

The gap

Three independent trackers converge on the same conclusion: **the gap is small, and no longer closing.** Stanford HAI's AI Index puts the top closed model **3.3% ahead** on Chatbot Arena (up from 1.7% a year earlier, and 0.5% in mid-2024). Epoch AI measures the open-weights lag at **3–4 months**. Artificial Analysis shows its composite intelligence gap collapsing from ~13 points in early 2025 to **3–6 points** in 2026, with every one of its top-ranked open models coming from China-based labs.

Mathematics: effectively at parity. On competition maths (AIME 2025), the top open models — gpt-oss-120b at ~97.9% with tools, DeepSeek at ~96% — sit alongside or above GPT-5-class closed models. **Agentic tool use: open has led.** Kimi K2 Thinking posted state-of-the-art results on Humanity's Last Exam (with tools) and BrowseComp, exceeding reported GPT-5 numbers, and stayed stable across 200–300 sequential tool calls. **Coding: within roughly ten points.** The best open models reach ~64–71% on SWE-bench Verified against the closed leader's ~81% — behind, but still production-grade.

The single biggest driver of the 2025 leap was the **open-sourcing of the reasoning paradigm**. DeepSeek-R1 published the reinforcement-learning recipe for eliciting long chain-of-thought in January 2025; within a year, toggleable "thinking" modes were standard across Qwen, gpt-oss, Phi and Mistral's Magistral. And because reasoning models distil into small dense ones (R1 shipped distillations from 1.5B to 70B), frontier-style reasoning now runs on consumer hardware.

A caution on the numbers. The AI Index found invalid-question rates of 2–42% across popular benchmarks; AIME is near-saturated, with memorisation concerns; contamination-resistant tests show drops of up to ~18 points against static ones; and vendors report best-case configurations (tools on, parallel sampling). Treat single-number leaderboard positions as indicative, not definitive.

Running it *yourself*

The tooling grew up. At the base sits llama.cpp and its GGUF format, a way to pack a whole model into a single file that runs on ordinary CPUs and consumer GPUs. Ollama and LM Studio are the desktop apps most people use to run a model on their own machine; vLLM and SGLang are what production servers run on. Hugging Face retired its own serving software, Text Generation Inference (TGI), in March 2026. Every layer speaks the OpenAI-compatible API, so the same application runs on a laptop or in a datacentre with a one-line change.

Mixture-of-experts made big models fit small hardware. The trick is that a model can carry an enormous total size but only fire a fraction of it per token, so memory stays modest while keeping quality high: a 671B-total model computes like a 37B one. That is why trillion-parameter open models now run on a single unified-memory workstation or high-memory server. The bottleneck has shifted off the models and onto the hardware.

Per token, open is dramatically cheap. Serving a Llama-70B-class model runs ~$0.10–1.04 per million tokens depending on provider, a ~10× spread that rewards shopping around, and DeepSeek-V3-class ~$0.20–0.56, against closed frontier APIs **10–50× higher**. Prices for equivalent capability fall ~10× a year. In a mid-2026 price war, DeepSeek made a 75% cut to its flagship pricing.

Build-versus-buy depends on how utilised the hardware is, which has largely converged for everyday enterprise work. Rent per-token APIs when traffic is bursty or hard to predict. Host your own when volume is high and steady, which for a 70B-class model starts paying off somewhere in the hundreds of millions to low billions of tokens a month. Most organisations land on **hybrid**: a self-hosted baseline for the predictable load, with serverless to soak up the spikes. Teams already running their own inference juggle ~7 models on average, matching each job to the cheapest model that clears the bar.

SYNTHESIS

Five findings

01 China leads the open frontier.

Every top-ranked open model on Artificial Analysis comes from a China-based lab; Qwen displaced Llama as the ecosystem's base layer; DeepSeek has at times been the largest single source of token volume on OpenRouter. The counterweights are regulatory and strategic. Government-device bans and procurement limits; as of July 2026 the US was *considering*, but had not enacted, restrictions on using Chinese open models. The July 2025 America's AI Action Plan frames "leading open models founded on American values" as an explicit goal.

02 The licence centre of gravity is now Apache-2.0 and MIT.

Qwen, DeepSeek, GLM, gpt-oss, Granite and ERNIE all ship genuinely permissive. Meta's Llama Community Licence, the original reference for "open", now reads as the restrictive outlier, and Google's Gemma is reportedly (not confirmed) moving to Apache-2.0 with Gemma 4. For adopters this collapses most of the legal ambiguity that dogged early open deployment.

03 Mixture-of-Experts made frontier-scale models locally viable.

Total parameters drive memory; active parameters drive speed. DeepSeek-V3's 671B total computes through 37B per token; gpt-oss-120b activates 5.1B and runs on a single 80GB GPU.

04 Local inference went mainstream.

Ollama at ~52 million monthly downloads; llama.cpp past 100,000 GitHub stars; ~135,000 GGUF files on Hugging Face; NPU phones decoding at >100 tokens per second. Production serving consolidated to vLLM and SGLang, with hardened air-gap builds from Red Hat and NVIDIA. The privacy, compliance and offline arguments now have a mature stack behind them. Reported enterprise switches include Coinbase, Airbnb, Uber and Shopify.

05 Economics, not capability, is the battleground

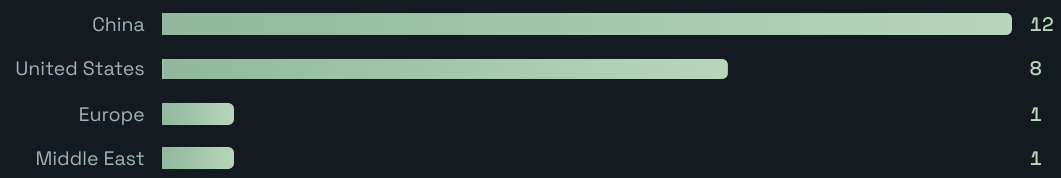
Inference prices fall ~10× a year while a memory and power supply crunch inflates the hardware underneath. Capability has largely converged for common enterprise tasks; what remains is governance, provenance and total cost of ownership. The rational posture for nearly every organisation is neither "all open" nor "all closed": closed frontier models for the hardest user-facing work, fine-tuned open weights for everything high-volume, private, cost-sensitive or offline — chosen workload by workload.

LANDSCAPE

The key models

The ~22 open models that define the mid-2026 landscape: who builds them, what they activate, what they ship under, and what each one signals. Figures as of 8 July 2026; prices, rankings and specifications are point-in-time snapshots of a fast-moving field.

Profiled models by builder region — July 2026.



The architectural extremes tell the story. **DeepSeek-V4-Pro** spans 1.6 trillion total parameters yet activates 49B per token; **gpt-oss-120b** activates just 5.1B and fits a single 80GB GPU; **Llama 4 Scout** claims a 10-million-token context window; **Kimi K2 Thinking** ships a trillion parameters quantisation-aware-trained to INT4. Sparse MoE is now the default above ~100B — and it is why the biggest open models run outside the biggest datacentres.

CHINA

DeepSeek-V4-Pro ↗

WHAT IT IS & CAPABILITIES

DeepSeek's latest open frontier model (April 2026): 1.6 trillion total parameters with only 49B active per token, sparse attention, and a 1M-token context — released under MIT.

ARCHITECTURE & LICENCE

1.6T total / 49B active · MoE + sparse attention · 1M context · MIT

STANDOUT & READ

The current high-water mark of the open frontier. Verified against primary sources; post-dates many third-party trackers.

DeepSeek-R1 ↗

WHAT IT IS & CAPABILITIES

The January 2025 shock: an open reasoning model matching OpenAI's o1 via pure reinforcement learning — with the recipe published. Wiped a record ~$589B off NVIDIA's market cap in a day.

ARCHITECTURE & LICENCE

671B total / 37B active · MoE · 128K context · MIT

STANDOUT & READ

The release that open-sourced the reasoning paradigm — and the sovereignty catalyst for half the world's AI programs.

DeepSeek-V3.2 ↗

WHAT IT IS & CAPABILITIES

Sparse-attention efficiency evolution of the V3 base; a maths-specialised variant reportedly reached IMO-2025 gold-level performance.

ARCHITECTURE & LICENCE

671B total / 37B active · MoE + sparse attention · 128K context · MIT

STANDOUT & READ

Frontier maths at commodity token prices — a mid-2026 price war made DeepSeek's 75% flagship cut permanent.

Qwen3-235B-A22B ↗

WHAT IT IS & CAPABILITIES

Alibaba's all-round flagship and the most-derived base model on Hugging Face: 100,000+ fine-tuned derivatives — more than Google and Meta combined — and ~700M cumulative family downloads by January 2026.

ARCHITECTURE & LICENCE

235B total / 22B active · MoE · 256K–1M context · Apache-2.0

STANDOUT & READ

The family that overtook Llama — the base layer of the open ecosystem, and of many nations' "sovereign" models.

Qwen3-Coder-480B ↗

WHAT IT IS & CAPABILITIES

The leading open coding model: agentic, repository-scale, long-context.

ARCHITECTURE & LICENCE

480B total / 35B active · MoE · 256K–1M context · Apache-2.0

STANDOUT & READ

State of the art for open code.

Qwen3.5-397B-A17B ↗

WHAT IT IS & CAPABILITIES

The February 2026 generation: multimodal, spanning 201 languages, on a leaner 17B-active architecture.

ARCHITECTURE & LICENCE

397B total / 17B active · MoE · large context · Apache-2.0

STANDOUT & READ

Verified against primary sources; post-dates most third-party trackers — specifications may be refined by the vendor.

Kimi K2 Thinking ↗

WHAT IT IS & CAPABILITIES

Moonshot's trillion-parameter reasoner, quantisation-aware-trained to INT4. Beat reported GPT-5 numbers on Humanity's Last Exam (with tools) and BrowseComp, and stayed stable across 200–300 sequential tool calls.

ARCHITECTURE & LICENCE

1T total / 32B active · MoE, INT4 QAT · 256K context · Modified-MIT

STANDOUT & READ

The landmark "open beats closed" agentic data point.

GLM-4.6 ↗

WHAT IT IS & CAPABILITIES

Zhipu/Z.ai's agentic and coding specialist; approaches Claude Sonnet-class results on LiveCodeBench.

ARCHITECTURE & LICENCE

357B total / 32B active · MoE · 200K context · MIT

STANDOUT & READ

Top-tier open for agents and code.

GLM-5.2 ↗

WHAT IT IS & CAPABILITIES

Z.ai's June 2026 flagship: #1 open model on the Artificial Analysis intelligence index at release.

ARCHITECTURE & LICENCE

744B total / 40B active · MoE · 1M context · MIT

STANDOUT & READ

China's newest number one. Verified against primary sources; post-dates many third-party trackers.

MiniMax-M2 ↗

WHAT IT IS & CAPABILITIES

Compact-activation MoE tuned for cheap agentic coding — 10B active parameters doing agent work.

ARCHITECTURE & LICENCE

230B total / 10B active · MoE · 192K context · MIT

STANDOUT & READ

Agentic capability at commodity prices.

ERNIE 4.5 (424B) ↗

WHAT IT IS & CAPABILITIES

Baidu's shift from closed to open: a multimodal MoE family released under Apache-2.0.

ARCHITECTURE & LICENCE

424B total / 47B active · multimodal MoE · Apache-2.0

STANDOUT & READ

The incumbent that switched sides — evidence the permissive tier is now table stakes in China.

Hunyuan-A13B ↗

WHAT IT IS & CAPABILITIES

Tencent's efficient fine-grained MoE with dual fast/slow thinking modes — strong capability per active parameter.

ARCHITECTURE & LICENCE

80B total / 13B active · fine-grained MoE · 256K context · Tencent licence

STANDOUT & READ

Efficiency-first engineering — but on custom terms, not Apache/MIT.

UNITED STATES

gpt-oss-120b ↗

WHAT IT IS & CAPABILITIES

OpenAI's first open weights since GPT-2 (August 2025): configurable reasoning effort, ships in MXFP4, ~97.9% on AIME with tools — and runs on a single 80GB GPU.

ARCHITECTURE & LICENCE

117B total / 5.1B active · MoE · 128K context · Apache-2.0

STANDOUT & READ

The best Western open model — and a strategic re-entry that reset expectations for US openness.

Llama 4 Maverick ↗

WHAT IT IS & CAPABILITIES

Meta's natively multimodal flagship (April 2025) — now trailing the open frontier it created; the larger "Behemoth" never shipped. The family crossed 1 billion downloads in March 2025.

ARCHITECTURE & LICENCE

~400B total / 17B active · MoE, native multimodal · 1M context · Llama Community

STANDOUT & READ

The pioneer, mid-retreat. Whether Llama 4 is Meta's last open release is rumoured, not confirmed.

Llama 4 Scout ↗

WHAT IT IS & CAPABILITIES

The long-context sibling: a claimed 10-million-token window on a single-host MoE.

ARCHITECTURE & LICENCE

109B total / 17B active · MoE, native multimodal · 10M context · Llama Community

STANDOUT & READ

Extreme context as the differentiator.

Nemotron 3 Ultra ↗

WHAT IT IS & CAPABILITIES

NVIDIA's hybrid Mamba-Transformer MoE with published recipes and much of the training data — the strongest US open model of 2026.

ARCHITECTURE & LICENCE

550B total / 55B active · hybrid Mamba-MoE · 1M context · OpenMDW

STANDOUT & READ

The chipmaker as open-model lab — near-fully open, and the likeliest vehicle for a US open-frontier comeback. Verified against primary sources.

Gemma 3 27B ↗

WHAT IT IS & CAPABILITIES

Google's small multimodal workhorse: dense with SigLIP vision, quantisation-aware builds for the edge, siblings down to 270M parameters.

ARCHITECTURE & LICENCE

27B dense · dense + SigLIP vision · 128K context · Gemma Terms

STANDOUT & READ

Best small multimodal. Custom terms — reportedly (not confirmed) moving to Apache-2.0 with Gemma 4.

Phi-4-reasoning ↗

WHAT IT IS & CAPABILITIES

Microsoft's 14B reasoner: distilled frontier-style reasoning in a laptop-class dense model.

ARCHITECTURE & LICENCE

14B dense · 32K–128K context · MIT

STANDOUT & READ

The reasoning-per-parameter leader.

Granite 4.0 H-Small ↗

WHAT IT IS & CAPABILITIES

IBM's hybrid Mamba-2/Transformer for the enterprise: low-RAM long context, certified and cryptographically signed, with Nano siblings that run in a browser.

ARCHITECTURE & LICENCE

32B total / 9B active · hybrid Mamba-2/Transformer · 512K context · Apache-2.0

STANDOUT & READ

Open weights dressed for procurement.

OLMo 3 32B ↗

WHAT IT IS & CAPABILITIES

Ai2's fully open release: weights, training data, code and checkpoints — reproducible end to end.

ARCHITECTURE & LICENCE

32B dense · 65K context · Apache-2.0

STANDOUT & READ

The reference for what "open source" strictly means.

EUROPE

Mistral Small 3.2 ↗

WHAT IT IS & CAPABILITIES

Europe's flagship open lab: the strongest Western small model, flanked by code (Codestral, Devstral), reasoning (Magistral) and audio (Voxtral) lines — Voxtral now beats Whisper on ASR accuracy.

ARCHITECTURE & LICENCE

24B dense · dense + vision · 128K context · Apache-2.0

STANDOUT & READ

The European answer — permissive where Llama is not, and the anchor of France's sovereign-AI agenda.

MIDDLE EAST

Falcon-H1 34B ↗

WHAT IT IS & CAPABILITIES

TII's hybrid Transformer-SSM line from Abu Dhabi, including the strongest open Arabic models.

ARCHITECTURE & LICENCE

34B · hybrid Transformer + Mamba · long context · TII Falcon (Apache-based)

STANDOUT & READ

The leading open family outside the US, China and Europe.

About Mirai Collective

Mirai Collective is an AI development studio. We build AI products end-to-end, train whole teams from zero to one, and guide organisations through AI adoption — practitioners, not theorists. The full research corpus behind this briefing — the model comparison and hardware/cloud cost sheets, the local-setup quickstart and the source log — is available on request: hello@miraicollective.io

Sources. Stanford HAI AI Index · Epoch AI · Artificial Analysis · LMArena · MLPerf · OpenRouter · Hugging Face · a16z · vendor blogs, model cards and arXiv papers · press of record. Figures as of 8 July 2026; prices and rankings are volatile snapshots and should be re-checked before use in a quote or business case.