

REFERENCE

Model Atlas

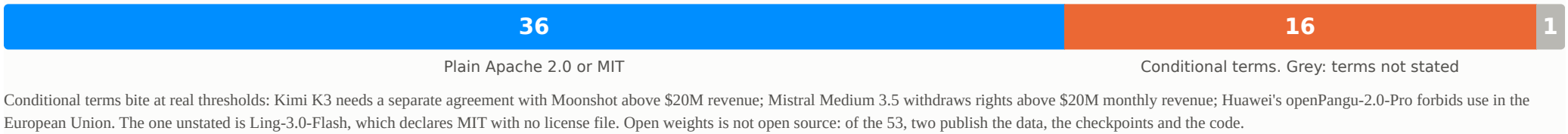
Seventy model families from the United States, China and everywhere else, plus twenty-six embedding, reranker and speech families. Who built each one, whether you can download it, what it costs, what it takes in, what it runs on, and what it is honestly good and bad at.

Prices are USD per million tokens, input then output, as each vendor quotes them. Scores are the Artificial Analysis Intelligence Index v4.3.2. Blank price cells mean the vendor publishes no list price.

FAMILIES LISTED	OPEN WEIGHTS	BEST OPEN MODEL	BEST US OPEN MODEL
70	53	46	26
across 14 countries of origin	downloadable and self-hostable	MiMo-V2.6-Pro, Xiaomi	Inkling-Small, Thinking Machines

Start with the license

All 53 downloadable models, by how clearly you are allowed to use them.

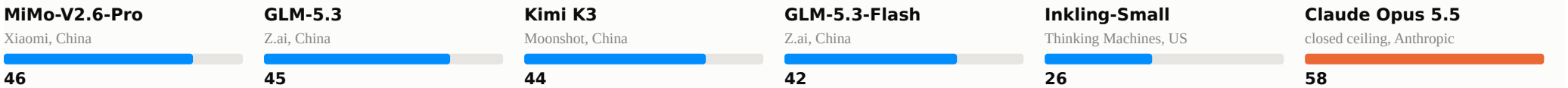


Then pick by the constraint you are actually under

Lowest cost per finished task	License certainty for a client contract	Runs on hardware you already own	You must show an auditor what went in
GLM-5.3-Flash or DeepSeek V4.1-Flash. Both MIT, both \$0.15 per million input tokens. Mistral Small 4 matches the price under Apache 2.0 if model origin is the constraint.	IBM Granite 4.2, Apache 2.0, cryptographically signed, with IP indemnification. Cohere Command A+, Apache 2.0, air-gapped on two H100s.	gpt-oss-120b on one 80GB GPU, gpt-oss-20b in 16GB, Gemma 4 E4B or Qwen3.5 4B on an 8GB board with image input, MiniCPM5-2B on a CPU. All Apache 2.0.	Olmo 3 or K2 Horizon. The two lines publishing weights, every checkpoint, the training data and the training code. Then the list ends.

For reference, the top of the board

Artificial Analysis Intelligence Index v4.3.2, highest published reasoning-effort setting. A tier, not a ranking.



The atlas

Grouped by where it was built and whether the weights are public. The Takes in and Fits columns are defined in the method section at the back.

United States: closed, API only 13 listed									
MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
GPT-6 Astra OpenAI	US	CLOSED API only	undisclosed	1M	\$10 / \$50	53	text, image	API	<p>STRONG The hardest end-to-end agentic work. Saturated ARC-AGI-3 at 99.9%. During unsafeguarded testing it found two previously unknown zero-days.</p> <p>WEAK Nothing is disclosed about size or architecture. Meets OpenAI's own Critical threshold for cyber capability, so parts of it are gated, and prompts above 272K tokens cost double on input and 1.5 times on output.</p> <p>BUILT Not disclosed. Announced 3 September 2026. Knowledge cutoff 30 April 2026.</p>
GPT-6.1 Sol OpenAI	US	CLOSED API only	undisclosed	1M	\$2 / \$10	52	text, image	API	<p>STRONG Agentic coding at a fifth of Astra's price. OpenAI says it matches GPT-6 Astra on DeepSWE v1.1 and cuts the factual error rate from 11.4% to 7.7%.</p> <p>WEAK The third Sol in ten weeks: GPT-5.6 Sol in July, GPT-6 Sol on 22 September, 6.1 Sol on 29 September. There is no GPT-6 Terra, so this is the middle tier as well as the top of the Sol line.</p> <p>BUILT Not disclosed. OpenAI says it uses the same types of data and training as GPT-6 Astra. Critical on cyber under the Preparedness Framework, so the same gating applies.</p>
GPT-6 Luna OpenAI	US	CLOSED API only	undisclosed	1M	\$0.1 / \$0.5	38	text, image	API	<p>STRONG Ten cents per million input tokens from a frontier lab, half the GPT-5.6 Luna price, with a 1M context window.</p> <p>WEAK GPT-5.6 Luna still carried more OpenRouter traffic in the week to 1 October, so the installed base has not moved yet. Index 38 is well below the Sol tier.</p> <p>BUILT Not disclosed. Released 22 September 2026. Knowledge cutoff 18 May 2026.</p>
Claude Fable 5.1 Anthropic	US	CLOSED API only	undisclosed	1M	\$10 / \$50	53	text, image	API	<p>STRONG The long-horizon reasoning flagship, and the only safeguarded route to Mythos-class capability for organisations outside Anthropic's verification programmes.</p> <p>WEAK No longer top of the board: Opus 5.5 and Sonnet 5.5 both score higher at a fraction of the price. Requests flagged for cyber route to Opus 4.8 and for biology to Opus 5, and Anthropic says so on the product page.</p> <p>BUILT Not disclosed. Fable is the safeguarded twin of Mythos 5.1, which is gated to approved US organisations. Announced 1 September 2026.</p>

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Claude Opus 5.5 Anthropic	US	CLOSED API only	undisclosed	1M	\$4 / \$20	58	text, image	API	STRONG Top of the intelligence index at 58, five points clear of Fable 5.1 and GPT-6 Astra, at 40% of Fable's price. Terminal-Bench 4.0 at 66.4% against Fable's 55.8%. WEAK Carries the same cyber and biology fallbacks as Fable, so flagged requests are answered by an older model. The benchmark figures are Anthropic's own and two weeks old. BUILT Not disclosed. Released 22 September 2026. Anthropic's recommended default for most workloads.
Claude Sonnet 5.5 Anthropic	US	CLOSED API only	undisclosed	1M	\$2 / \$10	56	text, image	API	STRONG Second on the index at a fifth of Fable's price. Terminal-Bench 4.0 at 70.6%, higher than Opus 5.5, from a line that scored 10.3% one version earlier. WEAK Higher-risk cybersecurity tasks visibly fall back to Sonnet 5, in Anthropic's words. Two points behind Opus 5.5 on GDPval-AA. BUILT Not disclosed. Released 28 September 2026. Anthropic says Haiku 5.5 follows in the coming weeks.
Claude Haiku 4.5 Anthropic	US	CLOSED API only	undisclosed	200K	\$1 / \$5	17	text, image	API	STRONG The cheap Claude. Still fine for classification and routing. WEAK A full generation behind the rest of the line, with a February 2025 knowledge cutoff. The retirement floor is 15 October 2026 and Haiku 5.5 is announced for the coming weeks. BUILT Released October 2025 and not refreshed since.
Gemini 3.1 Pro Google DeepMind	US	CLOSED API only, preview	undisclosed	1M	\$2 / \$12	30	text, image, audio, video	API	STRONG ARC-AGI-2 at 77.1%, more than double the previous Pro. Abstract reasoning is where this one earns its keep. WEAK Still labelled preview seven months after rollout began, and now eleven index points below the cheaper 3.8 Flash. Gemini 4 Argon, announced 30 September at \$2 / \$10, is gated to the Fairwind cyber programme and not yet on the API. BUILT Not disclosed. Price steps up to \$4 / \$18 above 200K tokens of context.
Gemini 3.8 Flash Google DeepMind	US	CLOSED API only	undisclosed	1M	\$0.75 / \$3.75	41	text, image, audio, video	API	STRONG Long-horizon software engineering and autonomous agents at a fraction of frontier pricing. Google's fastest-moving line. WEAK The price is promotional and doubles to \$1.50 / \$7.50 on 1 January 2027. BUILT Not disclosed. Generally available 2 September 2026. The Flash line now runs ahead of Pro on the index, which is the opposite of how the naming reads.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Gemini 3.5 Flash-Lite Google DeepMind	US	CLOSED API only	undisclosed	1M	\$0.3 / \$2.5	22	text, image, audio, video	API	STRONG High-volume classification and extraction. Takes text, image, video and audio at one rate on the cheap tier. WEAK Replaces 3.1 Flash-Lite, which Google has scheduled for shutdown on 7 May 2027. Twenty percent dearer on input and two thirds dearer on output than the model it replaces. BUILT Not disclosed. Google's changelog of 18 September 2026 points new projects here or to 3.8 Flash.
Grok 4.7 SpaceXAI	US	CLOSED API only	undisclosed	500K	\$2 / \$6	46	text, image	API	STRONG Same \$2 / \$6 price as Grok 4.6 with a higher index. SpaceXAI's own comparison puts it ahead of Fable 5.1 on electrical engineering and legal agent benchmarks. WEAK Prices double above 200K tokens of prompt. The fast variant is restricted to Cursor and Grok Build, not the public API. BUILT Not disclosed. Knowledge cutoff May 2026. Released 21 September 2026, forty days after Grok 4.6.
Muse Spark 1.3 Meta	US	CLOSED API only	undisclosed	1M	\$1.25 / \$4.25	48	text, image, audio, video	API	STRONG Index 48, the highest closed score outside Anthropic and OpenAI, generally available through Meta's own API since 24 September. Within a point of Claude Opus 5 on Meta's own coding composite. WEAK The Contributor tier at \$0.10 / \$0.20 is paid for with your prompts, which Meta may use for training. Meta hopes to open source future versions and has committed to no date. BUILT Meta says it rebuilt its stack from the ground up after abandoning Llama as a frontier line. The max reasoning level is sold on the Standard tier only.
MAI-Thinking-1 Microsoft	US	CLOSED Public preview	~1T / 35B active	256K	\$2 / \$8	–	text	API	STRONG Markets the absence of distillation as a feature: Microsoft says it was trained without distillation from third-party models, on what it calls clean, traceable, enterprise-grade data. WEAK Benchmarks are vendor-reported and it is not on the independent index. The \$2 / \$8 price comes from Microsoft's Foundry blog rather than a pricing page. BUILT Trained from the ground up on Microsoft's own data pipeline. The claim is the product. Public preview on Foundry since August 2026.

United States: open weights 14 listed									
MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Nemotron 3 Ultra NVIDIA	US	OPEN OpenMDW-1.1	550B / 55B active	1M	not published	23	text	Large 8x B200 minimum	STRONG Throughput. NVIDIA claims 5.9 times the inference rate of GLM-5.1 at similar accuracy, and the free tier is the only American open model in OpenRouter's top ten by tokens. WEAK Needs eight B200 GPUs to run. Raw intelligence trails the leading Chinese open models by half, and Thinking Machines' Inkling now scores above it. BUILT Trained from scratch on about 20T tokens in four-bit NVFP4, then SFT, multi-environment RL, and multi-teacher on-policy distillation.
Muse Glimmer 30B Meta	US	OPEN Apache 2.0	30B	128K	not published	17	text, image	Small 24GB VRAM quantized	STRONG End-to-end agentic task completion on a single consumer GPU. Multimodal, 100+ languages, 4-bit quantized into about 20GB. WEAK Meta's launch blog names no limitations; the model card does, including errors in multi-step reasoning. BUILT The clearest public example of a lab distilling its own flagship: pre-trained by logit distillation from Muse Spark, then post-trained with SFT, on-policy distillation and RL.
Gemma 4 31B Google	US	OPEN Apache 2.0	30.7B dense	256K	not published	15	text, image, video	Small	STRONG 140+ languages, text, image and video in, built on the same research as Gemini 3. Google says the Gemma family passed a billion downloads in August 2026. WEAK Data cutoff January 2025. Google's own card flags trouble with ambiguity, sarcasm and figurative language. BUILT Distilled during pretraining from a larger sibling.
Gemma 4 26B-A4B Google	US	OPEN Apache 2.0	25.2B / 3.8B active	256K	not published	17	text, image, video	Small	STRONG The sparse option in the Gemma line, and now the more downloaded of the two at 13 million a month. Scores two points above the dense 31B on the index. WEAK A 25B download for 4B of active compute, so the memory bill is the dense model's. Takes video as frames, like the 31B. BUILT Same pretraining distillation approach as the rest of Gemma 4. First Gemma release under Apache 2.0.

Model	Built in	Weights	Size	Context	Price in/out	Index	Takes in	Fits	What it is good and bad at, and how it was made
Gemma 4 E2B, E4B, 12B Google	US	OPEN Apache 2.0	5.1B, 8B, 12B dense	128K; 256K for 12B	<i>not published</i>	–	text, image, audio, video	Tiny E4B 4.5GB at Q4; 12B in 16GB	<p>STRONG The most downloaded family in the band with audio input: 4.3 million, 3.1 million and 1.6 million downloads a month for E4B, E2B and 12B, each taking text, image, video and audio under Apache 2.0. Google names Raspberry Pi and Jetson Orin Nano as targets for the two E models and says the 12B runs in 16GB.</p> <p>WEAK Every index score in this tier is an Artificial Analysis estimate rather than a measured result (14 for the 12B, 9 for E4B, 8 for E2B), so the column is blank. Data cutoff January 2025, and the E names understate the download: E4B is 8B of weights once embeddings are counted.</p> <p>BUILT Same family as the 31B and 26B-A4B rows. The E models shipped 2 April 2026 with per-layer embeddings that give a 2B or 4B effective footprint; the 12B followed 3 June 2026 as an encoder-free model with raw audio projected into the token space.</p>
Nemotron 3 Nano 4B NVIDIA	US	OPEN NVIDIA Nemotron Open Model License	4B hybrid	262K	<i>not published</i>	5	text	Tiny	<p>STRONG 3.5 million downloads a month for a 4B hybrid Mamba-Transformer with a 262K context, which NVIDIA aims at Jetson, GeForce RTX and DGX Spark for local assistants and IoT automation. The pretraining and post-training datasets are published.</p> <p>WEAK Scores 5 on the index and is text only. The licence is the Nemotron Open Model License rather than OpenMDW: commercial use is allowed, but derivatives must carry a Licensed by NVIDIA notice and the grant ends if a guardrail is removed.</p> <p>BUILT Mamba2-Transformer hybrid trained from scratch on more than 10 trillion tokens and compressed from Nemotron Nano 9B v2 with NVIDIA's Elastic framework. Released 16 March 2026.</p>
LFM2.5 Liquid AI	US	OPEN LFM Open License v1.0	230M to 2.6B; 8B / 1B active	128K	<i>not published</i>	–	text	Tiny 2.6B in under 2.5GB	<p>STRONG Built for CPUs and phones rather than shrunk from a server model: the 2.6B runs in under 2.5GB at 30 tokens a second on a phone by Liquid's figures, and the family adds vision and speech-in, speech-out siblings. The GGUF build had 968,000 downloads in a month.</p> <p>WEAK The LFM Open License withdraws commercial rights from any entity with \$10 million or more in annual revenue, the one conditional licence in this band. The index score of 8 is an estimate, so the column is blank, and Liquid's own card says the 2.6B is not for agentic coding or knowledge-heavy work.</p> <p>BUILT Hybrid of gated short-convolution blocks and grouped-query attention, pretrained on about 34 trillion tokens, then SFT, per-domain teacher specialisation, on-policy distillation and agentic RL. The 2.6B shipped 4 August 2026.</p>

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Nemotron 3.5 Lightning NVIDIA	US	OPEN OpenMDW-1.1	30B / 3B active	1M	not published	13	text	Small	STRONG Agentic speed. Terminal-Bench v2.1 at 24.6 against Nemotron 3 Nano's 8.3, and 10,000 tasks completed 30% faster than Qwen3.6 35B at comparable accuracy. WEAK Low on the raw intelligence index, and NVIDIA's own table has Qwen3.6 35B at 44 on the same Terminal-Bench. This is a specialist, not a generalist. BUILT Hybrid Mamba-Transformer MoE. NVIDIA also publishes the training datasets, part of them gated behind an approval.
gpt-oss-120b OpenAI	US	OPEN Apache 2.0	117B / 5.1B active	128K	not published	12	text	Medium one 80GB GPU	STRONG Fits on a single 80GB GPU thanks to 4-bit quantized expert weights. Genuinely permissive licence from a lab that ships nothing else open. WEAK Not refreshed since August 2025 while the closed line moved four generations. Text only, no image input. BUILT Token-choice MoE with sliding-window attention and learned attention sinks.
gpt-oss-20b OpenAI	US	OPEN Apache 2.0	21B / 3.6B active	128K	not published	9	text	Small 16GB	STRONG Runs in 16GB. The most deployable thing OpenAI has released, at 6.7 million downloads a month. WEAK Same August 2025 vintage as its larger sibling. BUILT Same architecture, smaller. A safety-classifier variant also exists.
Granite 4.2 IBM	US	OPEN Apache 2.0	3B, 8B, 30B dense	128K / 512K ext.	not published	15	text	Tiny to Small	STRONG Cryptographically signed releases and uncapped IP indemnification, which matters more to a procurement team than a benchmark does. The 30B scores 15 on the index, the 8B 11. WEAK Text only, no multimodal. Coding trails: SWE-Bench Verified 57 for the 30B despite a trillion synthetic code tokens. BUILT 15T tokens across five phases. IBM went back to plain dense attention after the hybrid Mamba variants of Granite 4.0 and 4.1.
Olmo 3 32B Allen Institute (Ai2)	US	OPEN Apache 2.0	32B	64K	not published	7	text	Tiny to Small	STRONG The only genuinely open line from a US lab: weights, every intermediate checkpoint, the 9.3T-token Dolma 3 corpus with the 6T mix actually used, and all the training code. WEAK Competitive within its transparency class, not at the frontier. The 32B context is 64K. BUILT Fully documented, which is the entire point of the project. Olmo Hybrid 7B followed in March 2026 on the same terms, and the Olmo-core 3 MoE training stack shipped on 1 October 2026 with no new model yet.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Inkling Thinking Machines	US	OPEN Apache 2.0	975B / 41B active; Small 276B / 12B	1M	not published	26	text, image, audio	Large 600GB VRAM at NVFP4	STRONG A customisation substrate rather than a chat model, pretrained from scratch on 45T tokens across text, image, audio and video. Inkling-Small, released 30 July, scores 26, the highest of any American open-weights model; the full model scores 25. WEAK The company says outright that it is not the strongest model available today, open or closed. Apache 2.0 is declared on the model card and repository metadata, with no separate LICENSE file in the repository. Reflection AI announced Beam, a 501B / 23B active model under Apache 2.0, on 5 October with weights promised later in the month; it enters when they ship. BUILT Trained entirely on NVIDIA GB300 systems, and used other labs' open models, including Moonshot's Kimi K2.5, to generate early post-training data. Small is distilled on-policy from the full model.
Laguna S 2.1 Poolside	US	OPEN OpenMDW-1.1	118B / 8B active	1M	\$0.1 / \$0.2	–	text	Medium about 236GB BF16, multi-GPU	STRONG An agentic coding specialist from a San Francisco startup that pretrains in-house. Terminal-Bench 2.1 at 70.2% and SWE-bench Multilingual at 78.5% in the vendor's own table, with interleaved thinking between tool calls. WEAK Poolside lists its own failure modes: harness overfitting with third-party tools, malformed JSON in nested tool calls, and overthinking. Coding only, and no general index score was found. The price is the one Poolside quotes for its own OpenRouter endpoint. BUILT 118B sparse MoE pretrained from 22 May 2026 on 4,096 H200 GPUs for about 60 days. Ships a draft model for speculative decoding.
China: open weights 18 listed									
MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
GLM-5.3 Z.ai (Zhipu)	China	OPEN custom GLM-5.3	753B / 40B active	1M	\$1.4 / \$4.4	45	text	Large	STRONG Second-highest open-weights score in the world at 45, one point behind MiMo-V2.6-Pro. Terminal Bench 3.0 at 28.3, six times GLM-5.2, and long-horizon agent work. WEAK Z.ai advertises more than double its predecessor's performance on exploitation benchmarks and emergent vulnerability discovery, in a model anyone can download. BUILT Post-training only. The vendor states plainly that every gain came from post-training on the existing GLM-5.2 base. The licence adds a security review for model-as-a-service operators above \$10B revenue.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Kimi K3 Moonshot AI	China	OPEN custom, revenue-gated	2.8T / 104B active	1M	\$3 / \$15	44	text, image, video	Datacenter	<p>STRONG BrowseComp at 91.2, the highest agentic search number found anywhere. Long-horizon coding and multi-agent research.</p> <p>WEAK Priced at parity with US frontier models, so the cheap-China story does not apply here. Training tokens and RL method are undisclosed.</p> <p>BUILT Not disclosed. The licence requires a separate agreement above \$20M revenue in any twelve months and attribution above 100M monthly users or \$20M monthly revenue.</p>
GLM-5.3-Flash Z.ai (Zhipu)	China	OPEN MIT	320B / 18B active	1M	\$0.15 / \$0.5	42	text, image	Large	<p>STRONG An MIT-licensed, natively multimodal 320B model at fifteen cents per million input tokens, and the most-used model of any kind on OpenRouter over the thirty days to 1 October at 56.5 trillion tokens.</p> <p>WEAK Evaluated at 300K context rather than the full million.</p> <p>BUILT Newly trained base on a 30T-token multimodal corpus, hybrid sparse and linear attention.</p>
Qwen3.8-Max Alibaba	China	OPEN custom qwen3.8-max	2.4T / 95B active	1M	\$2 / \$6	40	text	Datacenter	<p>STRONG GPQA Diamond at 92.6 and PaperBench at 93.0. Alibaba's largest flagship by parameter count.</p> <p>WEAK The downloadable weights are the Qwen3.8-2.4T-A95B base, which is text-only and cannot turn thinking off. The Max API product adds vision, 1M default context and tools, and scores 45 on the index against 40 for the base. The price is Alibaba's international list; other regions are lower.</p> <p>BUILT Not disclosed. Base weights are on Hugging Face under a custom licence with a \$50M revenue threshold for model-as-a-service businesses; the branded Max product is not the same artifact.</p>
Qwen3.8-Flash-Next Alibaba	China	OPEN qwen-community-1.0	125B / 6B active	1M	<i>not published</i>	40	text, image, video	Medium	<p>STRONG The most extreme sparsity in the field: six billion active parameters out of a 125B backbone, and it still posts SWE-bench Pro at 62.5.</p> <p>WEAK A research vehicle as much as a product. The community licence requires a separate licence for any model-as-a-service or AI work assistant business, with no revenue floor. Alibaba's qwen3.8-flash API is priced at \$0.15 / \$0.47 but Alibaba does not say it is this artifact.</p> <p>BUILT Hybrid Gated DeltaNet and Qwen Sparse Attention, n-gram embeddings to add parameters without routing overhead, mixed Muon and AdamW optimisers.</p>

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Qwen3.8-27B Alibaba	China	OPEN Apache 2.0	27B dense	1M	\$0.5 / \$3	34	text, image, video	Small	STRONG Natively multimodal, Apache 2.0, small enough to deploy, and nearly seven million downloads a month. OSWorld-Verified at 84.3. WEAK Dense, so it costs more per token to serve than the sparse models around it. Context is 256K native, extensible to 1M. BUILT Built on the Qwen3.5 architectural foundation. Qwen is the most-derived base on Hugging Face by a wide margin: 151,448 derivative models in Hugging Face's August 2026 count.
Qwen3.5 0.8B to 9B Alibaba	China	OPEN Apache 2.0	0.8B, 2B, 4B, 9B dense	262K	not published	11	text, image, video	Tiny	STRONG The most downloaded open weights under 12B of any kind: 8.3 million downloads a month for the 9B and 7.9 million for the 4B, and every size down to 0.8B carries a vision encoder for image and video input. The 9B posts GPQA Diamond 81.7 on its card. WEAK Released 2 March 2026 and not followed: Qwen3.6 and Qwen3.8 start at 27B, so this tier is two generations behind the flagship. Scores 11 on the index, below MiniCPM5-2B at a quarter of the size. No audio input. BUILT Causal language model with a vision encoder, early-fusion multimodal pretraining and reinforcement learning post-training; token counts are not published. Base checkpoints are published for each size.
MiniCPM5 1B, 2B OpenBMB	China	OPEN Apache 2.0	1B, 2B dense	128K	not published	12	text	Tiny	STRONG The highest measured index score under 3B at 12 for the 2B, above Qwen3.5 9B and above every estimated Gemma 4 small score, with 1.3 million downloads a month. AIME 2026 at 86.5 and LiveCodeBench v6 at 69.1 on the card. WEAK Text only; the vision sibling MiniCPM-V 4.6 scores an estimated 6. Two of the card's headline figures are copied from Artificial Analysis rather than reproduced. Released 7 September 2026, so independent results beyond the index are thin. BUILT Standard dense transformer trained in stable and decay phases, then SFT on 400B tokens of deep-thinking data, RL, and on-policy distillation that merges 16 expert models. OpenBMB says the 2B was adapted to nine AI chips, including Ascend, through FlagOS.
DeepSeek V4-Pro DeepSeek	China	OPEN MIT	1.6T / 49B active	1M	\$0.66 / \$1.98	36	text	Datacenter 4x GB300 node, DeepSeek docs	STRONG Terminal Bench 2.1 at 87.9 and CyberGym at 83.3, under MIT, at a fifth of Kimi's price. Off-peak cache hits cost about two cents per million tokens. WEAK DeepSeek publishes no training cost, GPU hours or hardware for V4. Any figure you see is somebody's estimate. Peak-hour prices are double, \$1.32 / \$3.96. BUILT More than 32T tokens. Domain expert models trained with GRPO, then consolidated into one model by on-policy distillation. Compressed sparse attention, Muon optimiser.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
DeepSeek V4.1-Flash <small>DeepSeek</small>	China	OPEN MIT	552B / 8B prefill, 16B decode	1M	\$0.15 / \$0.6	39	text, image	Large	<p>STRONG Terminal-Bench 2.1 at 90.6 from eight billion active parameters, at fifteen cents per million input tokens off-peak. Third on OpenRouter by tokens over thirty days. Global KV cache of 890 bytes per token.</p> <p>WEAK Shipped 10 September 2026 and inherited the V4-Flash endpoint and price the same day. Peak-hour rates are double.</p> <p>BUILT A causal encoder-decoder: 20 encoder and 20 decoder layers, one shared and 384 routed experts, six routed per token. 45T pretraining tokens, native vision.</p>
DeepSeek V4-Flash <small>DeepSeek</small>	China	OPEN MIT	284B / 13B active	1M	<i>not published</i>	–	text	Large <small>4x GB300 node, DeepSeek docs</small>	<p>STRONG The 0731 checkpoint beat the V4-Pro preview on DeepSeek's own card despite far fewer active parameters, and it was still the fifth most-used model on OpenRouter in the thirty days to 1 October.</p> <p>WEAK DeepSeek retired the endpoint on 10 September 2026 and routes the name to V4.1-Flash, so the vendor no longer sells or prices it. Third-party hosts pushed 41.7 trillion tokens through it in a month, which is why it stays.</p> <p>BUILT 32T tokens. Same architecture family as V4-Pro at a quarter the size. Weights remain on Hugging Face under MIT.</p>
Hy4-preview <small>Tencent</small>	China	OPEN Apache 2.0	770B / 49B active	1M	\$0.834 / \$2.501	–	text	Large	<p>STRONG A 770B model under Apache 2.0 with no commercial restrictions, and the second most-used open model on OpenRouter over thirty days, behind GLM-5.3-Flash.</p> <p>WEAK Still labelled preview five weeks after release; Tencent says the next batch is coming soon. Not on the independent index. Token count, compute and cost are all undisclosed.</p> <p>BUILT Ships something Tencent calls Gated DeepSeek Sparse Attention, naming a competitor's published mechanism in its own model card. Post-training built around real workplace tasks.</p>
MiMo-V2.6-Pro <small>Xiaomi</small>	China	OPEN MIT	1.02T / 42B active	1M	\$0.435 / \$0.87	46	text, image, audio, video	Large	<p>STRONG The highest-scoring open-weights model on the index at 46, under MIT, at under half a dollar per million input tokens. Terminal Bench 2.1 at 89.9 with native image and audio input.</p> <p>WEAK Released 22 September 2026, so the long-horizon results are Xiaomi's own. Two variants ship, RL and MOPD, and the vendor does not say which the API serves. The V2.5 API it replaces is deprecated on 21 October.</p> <p>BUILT Six days of what Xiaomi calls live reinforcement learning on the V2.5 base, then multi-teacher on-policy distillation. Xiaomi puts the RL bill at about \$2.62M. A 309B Flash sibling at \$0.14 / \$0.28 scores 38.</p>

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
MiniMax M3 MiniMax	China	OPEN minimax-community	428B / 23B active	1M	\$0.3 / \$1.2	29	text, image, video	Large	STRONG SWE-Bench Verified at 80.5 and trained mixed-modality from the first step rather than bolted together. WEAK Long-Horizon Terminal Bench at 38.5 against single-task scores of 85 to 90 across the field. That gap is the clearest picture of where these models break. The price is a permanent 50% discount on a \$0.6 / \$2.4 list, for prompts up to 512K. BUILT MiniMax Sparse Attention. Per-token compute cut to a twentieth of previous versions. Attribution required above \$20M yearly revenue.
Ling-3.0-Flash Ant Group	China	OPEN stated as MIT	124B / 5.1B active	256K	not published	–	text	Medium 4x 141GB GPUs	STRONG Activates one expert in sixty-four and, on Ant's own numbers, matches its previous 1T flagship at about 12% of the parameters. WEAK MIT is declared in the repository metadata, but the repository carries no LICENSE file, so the terms are a declaration rather than a document. BUILT Alternates Kimi Delta Attention and multi-head latent attention at five to one. Ant is shipping Moonshot's published attention operator, by name, in its own model.
Step-3.7-Flash StepFun	China	OPEN Apache 2.0	198B / 11B active	256K	\$0.2 / \$1.15	–	text, image	Medium 120GB unified, quantized	STRONG A vision-language MoE aimed squarely at coding agents and search workflows, with quantized variants down to 104B. WEAK Positioned on multimodal efficiency rather than frontier capability. Not on the independent index. StepFun's newer Step 5 Preview scores 44, but its weights are not public. BUILT Not disclosed in detail. StepFun, the lab, is named in the September 2026 NSA/CISA/FBI advisory, which cites its Step 4 model rather than this one.
openPangu-2.0-Pro Huawei	China	OPEN custom openPangu 2.0	505B / 18B active	512K	not published	–	text	Large	STRONG Pretrained on about 34T tokens entirely on Huawei Ascend NPUs with no NVIDIA hardware, which is the only claim of its kind at this scale. WEAK No independent benchmarks exist. The licence forbids use within the European Union and requires Powered by openPangu attribution. Huawei's model team was accused in 2025 of shipping something suspiciously similar to Qwen, which it denied. BUILT Weights and inference code on Huawei's GitCode and, since July 2026, a Hugging Face organisation.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Intern-S1-Pro Shanghai AI Lab	China	OPEN Apache 2.0	1T / 22B active	—	<i>not published</i>	—	text, image	Large	STRONG Built for science rather than chat: chemistry, materials, life sciences, earth science and physics, benchmarked on olympiad physics and scientific reasoning. WEAK Narrow by design. Context length is not stated on the card. BUILT Eight of 512 experts active per token. Its September 2026 successor, Intern-S2-397B, is also Apache 2.0 and declares a Qwen3.5 MoE architecture in its config.

Elsewhere: open weights 21 listed

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Mistral Large 3 Mistral AI	France	OPEN Apache 2.0	675B / 41B active	256K	\$0.5 / \$1.5	9	text, image	Large one B200 or H200 node, FP8	STRONG A 675B sparse MoE under Apache 2.0 from a lab that genuinely pretrains its own bases. One of the few sovereignty claims outside the US and China that survives inspection at the weights level. WEAK Outscored by the Chinese open models it competes with on price and openness: 9 on the index against 45 for GLM-5.3. BUILT Own pretraining on 3,000 H200s. Mistral moved from a restrictive research licence to Apache 2.0 across its flagship tier during 2025 and 2026.
Mistral Medium 3.5 Mistral AI	France	OPEN Modified MIT	128B dense	256K	\$1.5 / \$7.5	14	text, image	Medium	STRONG SWE-Bench Verified at 77.6, multimodal, 24 languages, and dense rather than sparse, which makes it predictable to serve. WEAK The Modified MIT licence withdraws all rights from companies above \$20M monthly revenue, so read it before you deploy. BUILT Absorbed both the Magistral reasoning line and the Devstral coding line. Those no longer exist separately.
Mistral Small 4 Mistral AI	France	OPEN Apache 2.0	119B / 6.5B active	256K	\$0.15 / \$0.6	11	text, image	Medium	STRONG Instruct, reasoning and coding unified into one Apache-licensed model with a toggleable reasoning mode. Fifteen cents per million input tokens from a European lab, for buyers who cannot use a Chinese model at that price. WEAK Small is a relative term: 119B total is not a laptop model. BUILT 128 experts, four active. Own pretraining.

Model	Built in	Weights	Size	Context	Price in/out	Index	Takes in	Fits	What it is good and bad at, and how it was made
Ministral 3 Mistral AI	France	OPEN Apache 2.0	3B, 8B dense	256K	\$0.15 / \$0.15	5	text, image	Tiny 3B in 8GB, 8B in 12GB at FP8	STRONG The European option for the band: Apache 2.0, image input at every size, 256K context, and the only vendor here that states a VRAM floor, 8GB for the 3B and 12GB for the 8B in FP8. Base, instruct and reasoning variants of each. WEAK Both sizes score 5 on the index, the lowest of the on-device families here. Downloads are modest at about 114,000 a month per repo, and the price sits in the launch post rather than on Mistral's pricing page, which omits Ministral. A 14B sits above the band. BUILT Dense transformers with a 0.4B vision encoder on 3.4B and 8.4B language models, released 2 December 2025 as the edge tier of Mistral 3. Own pretraining; training data not disclosed. Voxtral Mini 4B Realtime in the speech section is built on the 3B base.
Command A+ Cohere	Canada	OPEN Apache 2.0	218B / 25B active	128K	not published	13	text, image	Large 2x H100 at W4A4	STRONG Built explicitly for air-gapped and on-premise sovereign deployment, with full visibility into the architecture. Runs on two H100s. 48 languages including every official EU language. WEAK Scores 13 on the index. Cohere sells private deployment, not benchmark leadership. BUILT Cohere's first mixture-of-experts model and, by its own docs, the last in the Command A family. Own pretraining. Cohere and Aleph Alpha signed a definitive combination agreement in September 2026, pending regulatory approval.
Jais 2 70B Inception / G42	UAE	OPEN Apache 2.0	70B	8K	not published	–	text	Medium	STRONG Trained from scratch on 2.6T tokens, roughly a seventh of what a comparable Llama used, with a custom Arabic vocabulary. Highest Arabic score on AraGen. WEAK An 8,192-token context window in 2026 rules it out for most agentic work. BUILT Trained on Cerebras wafer-scale systems rather than GPU clusters, which is the interesting part: national-scale capability without a GPU megacuster.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
K2 Horizon 375B IFM / MBZUAI	UAE	OPEN Apache 2.0	375B / 23B active	512K	not published	31	text	Large validated on 8x H200	STRONG What IFM calls the largest fully open model launch to date: final weights, intermediate checkpoints, training code and the TxT360-v2 corpus are all published under Apache 2.0. Scores 31 on the index, above every American open model. WEAK Released 6 September 2026, so independent results are thin, and 23B active still needs eight H200s to run. BUILT Sparse MoE pretrained in two phases of 7.08T and 8.05T tokens, then mid-training at native 512K context and SFT. One of two lines in this atlas you can rebuild from published parts.
Falcon-H1 34B TII	UAE	OPEN Falcon LLM License	34B	256K	not published	–	text	Small	STRONG Hybrid Transformer and Mamba, 18 languages, with a 7B reasoning variant that posts AIME-24 at 88.1 and beats models several times its size. WEAK The Falcon licence is not Apache and not MIT: derivatives must carry a built-with-Falcon attribution. Check the terms. BUILT Own pretraining. TII also ships a 1.58-bit ternary-weight edge line pretrained at that precision from scratch, not quantized afterwards.
Sarvam 105B Sarvam AI	India	OPEN Apache 2.0	105B / 10.3B active	128K	not published	9	text	Medium	STRONG MMLU at 90.6 and Math500 at 98.6, trained from scratch on 12T tokens entirely on Indian national compute, covering 22 scheduled Indian languages. WEAK Evaluation infrastructure for Indic performance is thin, which makes the claims hard to audit from outside. Scores 9 on the general index. BUILT The cleanest upgrade story in sovereign AI: Sarvam's previous model was a Mistral Small fine-tune. Ten months later it was training its own base.
A.X K2 SK Telecom	South Korea	OPEN Apache 2.0	688B / 33B active	256K	not published	23	text	Large 800GiB GPU memory, FP8	STRONG Trained from scratch on 8.2T tokens entirely in FP8, under a rule that bans foreign weights outright. AIME26 at 97.1%. WEAK BrowseComp at 9.3 against 29.1 for leading alternatives. Open-ended web research is a hole. BUILT Three curriculum stages, all FP8. Part of Korea's sovereign programme, which eliminated Naver Cloud in January 2026 for using frozen Qwen encoder weights and advanced SKT, LG and Upstage in August.
K-EXAONE 2.0 LG AI Research	South Korea	OPEN Apache 2.0	750B / 37B active	262K	not published	20	text	Large two 8x H200 nodes, quickstart	STRONG Best long-context retrieval of the Korean cohort at 94.4 on OpenAI-MRCR, plus SWE-Bench Verified at 68.2 and AIME 2026 at 92.3%. WEAK Upcycled from LG's own 236B model rather than trained fresh, which is a different claim from what the programme's rhetoric implies. BUILT Continual pretraining, difficulty-focused mid-training, then post-training. More than three times the size of its predecessor.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Solar Open 2 Upstage	South Korea	OPEN Upstage Solar License	250B / 15B active	1M	<i>not published</i>	25	text	Large 4x H200 minimum	STRONG A Korean-optimised tokenizer processes Korean text in 50 to 80% of the tokens global models need, which is a direct cost saving, not a benchmark. Runs on four H200s. Highest index score of the Korean open cohort at 25. WEAK Its own licence rather than Apache: derivatives must carry the Solar name and a Built with Solar notice, though commercial use is permitted. BUILT Selective weight transfer from Upstage's own earlier 100B model, reaching target loss in 59% of the tokens random initialisation needed.
Motif-3 Motif Technologies	South Korea	OPEN MIT	314B / 13.2B active	262K	<i>not published</i>	34	text	Large	STRONG MIT-licensed, built from scratch on about 12.5T tokens, and scores 34 on the index, the highest of Korea's sovereign cohort. SWE-Bench Verified 76.2 and Terminal-Bench 2.1 74.9 on the vendor's table. WEAK Cut from Korea's national programme on 18 August 2026 on usability and ecosystem grounds despite topping its benchmarks, which leaves the lab without the programme's compute. BUILT 384 routed experts, eight active plus one shared, 256K native context. Motif calls the design fully in-house; the weights are MIT.
Rakuten AI 3.0 Rakuten	Japan	OPEN Apache 2.0	671B / 37B active	128K	<i>not published</i>	–	text	Large	STRONG Free under Apache 2.0, and by parameter count the largest model released by a Japanese company. WEAK Its config.json declares model_type deepseek_v3 with 256 routed experts and a 129,280 vocabulary, matching DeepSeek-V3 exactly. Rakuten's own announcement never claimed from-scratch pretraining. BUILT Rakuten says it was developed by leveraging the best from the open-source community. Part-funded by a Japanese government programme.
GigaChat 3.5 Ultra Sber	Russia	OPEN MIT	432B / 28B active	262K	<i>not published</i>	–	text	Large	STRONG A 432B model under MIT with a reasoning variant shipped in September 2026. Sber says it is 40% smaller than the 702B it replaces and stronger on code, maths and agents. WEAK The from-scratch claim was made for the 702B preview. The 3.5 cards do not repeat it, and the architecture now adds GatedDeltaNet linear attention to the DeepSeek-style latent attention and multi-token prediction already present. BUILT Multi-head latent attention interleaved with GatedDeltaNet layers, gated normalisation and multi-token prediction heads. About 5.5T synthetic tokens in the corpus, per the 3.1 card.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
AliceAI Foundation 80B Yandex	Russia	OPEN Apache 2.0	80B / 3B active	262K	<i>not published</i>	–	text	Medium	STRONG Yandex's first base model trained from scratch and released openly: 18T tokens, 512 experts, and a hybrid of Kimi Delta Attention and gated attention. Yandex says it beats its own 235B Alice AI LLM on knowledge, maths, code and long context. WEAK Base model only, no instruct or chat checkpoint, and Russian and English only. The benchmarks are Yandex's own. BUILT Four training stages totalling about 18T tokens. Released 21 September 2026. The closed Alice AI LLM it sits beside was initialised from Qwen3-235B.
EuroLLM-22B Unbabel / IST	EU	OPEN Apache 2.0	22B	32K	<i>not published</i>	–	text	Small	STRONG 35 languages including all 24 official EU languages, from scratch, Apache 2.0. Strongest on translation. WEAK 22B is a research-scale model in a year when its competitors ship 700B. Not to be confused with OpenEuroLLM, a separate consortium that has shipped no models at all. BUILT 4T tokens on 400 NVIDIA H100s on Spain's MareNostrum5 supercomputer, funded through EU research programmes.
Apertus 1.5 70B Swiss AI Initiative	Switzerland	OPEN Apache 2.0	70B dense	262K	<i>not published</i>	–	text, image, audio	Medium	STRONG Switzerland's national model, now multimodal: text, image and audio in, under Apache 2.0, with open data and full training details promised on the same terms as Apertus 1.0. Context grew four-fold to 262K. WEAK The 1.5 technical report and benchmark tables were still pending when checked, so the fully open claim rests on the 1.0 precedent until it lands. 70B dense is mid-sized against the sparse 300B to 700B models around it. BUILT Continued pretraining of Apertus 1.0 with a further 2T tokens for the 70B, including multimodal data. Released July 2026 by EPFL, ETH Zurich and CSCS.
Jamba Large 1.7 AI21 Labs	Israel	OPEN Jamba Open Model License	398B / 94B active	256K	\$2 / \$8	6	text	Large 8x 80GB, Int8	STRONG Genuinely original architecture: a hybrid Mamba and Transformer design that meaningfully cuts KV-cache cost at 256K context. WEAK Knowledge cutoff of August 2024. In May 2026 AI21 cut about 60% of staff and, according to press reports, said it would stop selling models to focus on its agent product. The endpoint is still listed and priced, and the licence ends commercial rights above \$50M annual revenue. BUILT State space model layers interleaved with attention. The most pointed commercial outcome in the dataset: technically distinctive, commercially unviable.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
SEA-LION v4.5 AI Singapore	Singapore	OPEN MIT	5.1B and 27B	128K / 262K	<i>not published</i>	–	text, image	Tiny to Small	STRONG Real Southeast Asian language coverage that nobody else provides, in sizes small enough to deploy locally. WEAK Two series built on Google Gemma 4 and Alibaba Qwen 3.6. AI Singapore declares MIT on both repositories, though the Gemma-based one links Google's own licence page, so check which terms bind you. BUILT Post-training on foreign bases, using logit distillation and model merging, by Singapore's own documentation.
Latam-GPT CENIA	Chile	OPEN Llama 3.1 Community License	70B	–	<i>not published</i>	–	text	Medium about 140GB BF16	STRONG A 297B-token corpus from more than 75 institutions across 20 Latin American countries, published in September 2026, addressing a real gap: Spanish and Portuguese are 2 to 3% of existing training data. WEAK Built on Meta's Llama 3.1 70B, so Meta's licence and its 700M-user threshold apply. Context length is not stated. BUILT Continued pretraining and fine-tuning on a foreign base. The corpus is now public; the base is not Latin American.
Elsewhere: closed, API only 4 listed									
MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Doubao Seed 2.1 Pro ByteDance	China	CLOSED API only	undisclosed	–	<i>not published</i>	–	text, image, video	API	STRONG Powers China's most-used consumer AI app, reportedly processing 120 trillion tokens a day. WEAK Nothing is open. Volcengine prices it in yuan with no USD list, no parameter count has ever been published, and ByteDance's only open LLM is the 36B Seed-OSS from August 2025. BUILT Not disclosed. ByteDance had its OpenAI API access suspended in December 2023 for using GPT output in development.
ERNIE 5.1 Baidu	China	CLOSED API only	undisclosed	–	<i>not published</i>	–	not stated	API	STRONG Fourth globally and first among Chinese models on LMArena's search leaderboard when the 5.1 preview landed. WEAK Baidu claims leading performance at 6% of the pre-training cost of comparable models, names no baseline, publishes no absolute figure, and keeps the model closed so nobody can check. Its international price page lists ERNIE 5.0 only. BUILT Not disclosed. Its open line stalled at ERNIE 4.5.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	INDEX	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
PLaMo 3.0 Prime Preferred Networks	Japan	CLOSED API only	undisclosed	256K	<i>not published</i>	–	not stated	API	STRONG Genuinely from scratch, co-developed with Japan's national research institute, with a proprietary tokenizer that cuts Japanese token counts and therefore cost. WEAK Parameter count, architecture and absolute benchmark scores are all undisclosed. The yen price circulating in developer blogs is not on a PFN page that could be read. BUILT Priced in yen against GPT-5.4-mini and Claude Haiku. The open NICT base models are on Hugging Face.
Alice AI LLM Yandex	Russia	CLOSED API only	235B MoE	128K	\$4.1 / \$9.84	–	text	API	STRONG Russia's most-used AI assistant, heavily engineered for serving: 5.8x faster than its own baseline through quantization and speculative decoding. WEAK Yandex's own engineering report says the flagship was initialised from the weights of an open-source model, specifically Alibaba's Qwen3-235B. The price is the USD list for customers outside Russia. BUILT Trained on 4,000 GPUs with a custom stack after foreign initialisation. Yandex's from-scratch open 80B base, released September 2026, is in the open table.

Beyond text generation

Twenty-six embedding, reranker and speech families, listed on the same terms. The score column is the public board for that section: MTEB Multilingual v2 for embeddings and rerankers, and Open ASR Leaderboard average word error rate for speech, where lower is better. These rows are not in the cover counts or the licence ladder.

- Embeddings.** An embedding model writes nothing. It turns a piece of text, or an image or a clip, into a list of numbers whose distance from other such lists measures how alike the two are. That is how search, retrieval-augmented generation and clustering find the right passages before a generative model ever sees them. It runs once per document and once per query, is small and cheap, and is judged on retrieval quality, not reasoning.
- Rerankers.** A reranker reads a query and a shortlist of candidates together and scores how well each one answers it. It is slower per item than an embedding and more accurate, so it sits after the embedding search and before the generative model, deciding which twenty passages out of two hundred deserve the context window.
- Speech.** Speech models convert between sound and text in either direction: recognition turns audio into a transcript, synthesis turns text into a voice. They are priced per hour of audio or per character rather than per token, and judged by word error rate, where lower is better, rather than by an intelligence score. Some generative rows above accept audio in; these are the specialists that do it best and run on the smallest hardware.
- How they differ from the rows above.** None of them reasons, converses or writes. They are the components around a generative model: what finds the material, what sorts it, and what carries it in and out as speech. Licence, price and hardware fit matter exactly as much, which is why they are listed on the same terms.

Embeddings 9 listed										
MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	MTEB	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE	
<div>Harrier-OSS-v1 27B</div> <div>Microsoft (Bing)</div>	US	OPEN MIT	27B, 0.6B, 270M	32K	not published	74.3	text	Tiny to Small	<div>STRONG Top of MTEB Multilingual v2 at 74.27, two points clear of the next model, under MIT, with 32K context and more than 100 languages. The 0.6B sibling scores 69.0, above every closed API on the board.</div> <div>WEAK Only 78% of the benchmark tasks count as zero-shot, so part of the lead is training-set overlap. No base model, no release date on the card, and no hosted price.</div> <div>BUILT Contrastive pretraining on more than 2 billion weakly supervised pairs, then fine-tuning on over 10 million examples, with multilingual synthetic pairs generated by GPT-5. Decoder-only, 5,376-dimensional output. Released April 2026.</div>	
<div>Qwen3-Embedding</div> <div>Alibaba</div>	China	OPEN Apache 2.0	0.6B, 4B, 8B	32K	not published	70.6	text	Tiny	<div>STRONG Third on the board at 70.58 for the 8B, 99% zero-shot, Apache 2.0, output size selectable from 32 to 4,096 dimensions, more than 100 languages. The 0.6B is downloaded over nine million times a month.</div> <div>WEAK Sixteen months old with no Qwen3.5 or Qwen3.8 successor shipped. Text only. No vendor price for the open artifact.</div> <div>BUILT Contrastive pretraining on pairs synthesised by Qwen3, supervised training, then checkpoint merging. LoRA on the Qwen3 bases. Released June 2025.</div>	

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	MTEB	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
EmbeddingGemma 2 Google DeepMind	US	OPEN Apache 2.0	740M	8K	<i>not published</i>	61.4	text, image, audio, video	Tiny 567MB RAM on a Pixel	STRONG Five modalities into one 768-dimensional space from 740M parameters, modular so the text-only path is 270M. Google's card gives MTEB multilingual v2 61.36. Context grew four-fold to 8,192 tokens. WEAK Released 6 October 2026 with no board entry yet, and 61.36 is where the 300M EmbeddingGemma 1 already sat, so the gain is modality coverage rather than text quality. The repository is gated. BUILT An adapted Gemma 4 decoder for text with separate vision and audio encoders, trained with Matryoshka representation learning across more than 140 languages. Runs in about 567MB on a phone by Google's figure.
Gemini Embedding 2 Google DeepMind	US	CLOSED API only	undisclosed	8K	\$0.15 / –	–	text, image, audio, video	API	STRONG Google's first natively multimodal embedding: text, up to six images, two minutes of video, audio and PDFs into one 3,072-dimensional space with truncation to 128. Generally available since April 2026. WEAK No complete MTEB submission, so no board score; the 001 model it replaces scores 68.37 and its space is incompatible, so upgrading means re-embedding every corpus. The price is the Vertex rate card; the Gemini API pricing page carried no embedding row when read. BUILT Not disclosed beyond being built on the Gemini architecture. Image, video and audio input are priced separately from text.
text-embedding-3-large OpenAI	US	CLOSED API only	undisclosed	8K	\$0.13 / –	–	text	API	STRONG The default most teams already run. 3,072 dimensions with a parameter to shrink them. text-embedding-3-small costs \$0.02. WEAK Released January 2024 and never refreshed, though OpenAI's docs still call it newest. Where the multilingual board can compare it, its retrieval score trails Qwen3-Embedding-8B by eleven points. Text only. BUILT Not disclosed.
Embed 5 Cohere	Canada	CLOSED API only	undisclosed	128K	\$0.12 / –	–	text, image	API	STRONG Released 30 September 2026 as Pro and Fast (\$0.08) with 128K context, 2,048 dimensions down to 256, and one shared space so a corpus indexed with Fast can be queried with Pro. Cohere's own visual document retrieval table puts Pro at 85.8 against Voyage 4 Large at 83.7. WEAK No MTEB submission and no MTEB figure in the launch post. The pay-as-you-go price is in the launch blog, not on the pricing page, which lists only Model Vault instances by the hour. BUILT Not disclosed. Available on Cohere's API, Model Vault, Microsoft Foundry and Amazon SageMaker.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	MTEB	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
voyage-4-large Voyage AI (MongoDB)	US	CLOSED API only	undisclosed	32K	\$0.12 / –	–	text	API	STRONG Voyage calls it the first production embedding model built as a mixture of experts and, on its own 29-dataset table, puts it 14% above OpenAI's text-embedding-3-large. Four sizes share one space, down to a free 340M nano under Apache 2.0. WEAK No MTEB Multilingual v2 submission for any Voyage 4 size; the only complete Voyage entry on the board is voyage-3.5 at 58.46. Voyage is a MongoDB company since February 2025, not Anthropic's, whatever the coverage says. BUILT Mixture of experts for the large model, training data undisclosed. voyage-context-4 and voyage-code-4 extend the line. Released January 2026.
jina-embeddings-v5 Jina AI	Germany	OPEN CC BY-NC 4.0	239M to 1.7B	32K	not published	67.0	text, image, audio, video	Tiny	STRONG The strongest sub-1B text embedder on the board, and the omni-small adds image, video and audio towers at 1.7B, which Jina calls the best open omni model under 2B parameters. WEAK Every Jina model is CC BY-NC 4.0: non-commercial unless you buy a licence or use the API, whose price is not printed anywhere readable. BUILT text-small is Qwen3-0.6B-Base trained by distillation from Qwen3-Embedding-4B; omni adds locked modality towers. Released February and May 2026.
BGE-M3 BAAI	China	OPEN MIT	568M	8K	not published	–	text	Tiny	STRONG The most downloaded multilingual embedder in existence at 34 million a month, MIT, more than 100 languages, and dense, sparse and multi-vector output from one 568M model. WEAK Released February 2024 and not refreshed; BAAI lists no 2026 embedding or reranker. The board leaves its mean blank over incomplete subsets, and its retrieval score is sixteen points behind Qwen3-Embedding-8B. BUILT XLM-RoBERTa extended to 8,192 tokens by RetroMAE pretraining, then self-distillation across the three retrieval modes.

Rerankers 8 listed									
MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	MTEB	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Qwen3-Reranker Alibaba	China	OPEN Apache 2.0	0.6B, 4B, 8B	32K	not published	–	text	Tiny	STRONG The open reranker everyone else benchmarks against: MTEB-R 72.94 and MTEB-Code 81.22 for the 8B on the vendor's table, more than 100 languages, and the 0.6B is downloaded 1.6 million times a month. WEAK June 2025 with no successor; the April 2026 Qwen3-VL-Reranker covers images, not a text refresh. Instruction following is weak on the vendor's own FollowIR column. BUILT Single-stage supervised training on the Qwen3 bases. Scores a yes or no token given query and document.
Rerank 4 Cohere	Canada	CLOSED API only	undisclosed	32K	not published	–	text	API	STRONG Pro and Fast variants released December 2025 lift the window from 4K to 32K, rerank JSON documents, cover more than 100 languages and take up to 1,000 documents per call. WEAK Cohere publishes no benchmark figure for Rerank 4 anywhere readable and no per-search price; the pricing page lists only Model Vault instances by the hour. BUILT Not disclosed.
rerank-3 Voyage AI (MongoDB)	US	CLOSED API only	undisclosed	32K	\$0.05 / –	–	text	API	STRONG Released 30 September 2026. On Voyage's own 95-dataset evaluation it beats Cohere rerank-v4.0-pro by 2.72% and Qwen3-Reranker-8B by 3.02%; the lite version costs \$0.02. WEAK A week old and every number is Voyage's. No parameter count, no architecture, no MTEB submission. BUILT An updated backbone and an improved mixture of training data is the whole disclosure.
jina-reranker-v3.5 Jina AI	Germany	OPEN CC BY-NC 4.0	0.6B	131K	not published	67.8	text	Tiny	STRONG A listwise reranker that scores up to 64 documents in one 131K pass. Jina's card gives BEIR 63.20, matching a 4B model at roughly a seventh of the parameters. WEAK CC BY-NC 4.0, so commercial use needs a paid licence or the API, whose price is not published. The board score of 67.84 belongs to the September 2025 v3; v3.5 has no board results. BUILT Qwen3-0.6B-Base with sliding-window layers interleaved with global ones, trained by self-distillation. Released July 2026.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	MTEB	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
bge-reranker-v2-m3 BAAI	China	OPEN Apache 2.0	568M	8K	<i>not published</i>	–	text	Tiny	STRONG The most downloaded reranker on Hugging Face at 17 million a month, Apache 2.0, multilingual, and cheap to run. WEAK February 2024 and unrefreshed. No leaderboard presence and no modern benchmark on the card. BUILT A cross-encoder fine-tuned from bge-m3.
zerank-2 ZeroEntropy (Notion)	US	OPEN Apache 2.0	4B, 2B, nano	32K	<i>not published</i>	–	text	Tiny	STRONG The clearest openness story in the section: after Notion acquired ZeroEntropy in July 2026 every model was relicensed to Apache 2.0, so a reranker that was API-gated is now free to self-host. Downloaded 680,000 times a month. WEAK Dead as a service: the API was sunset on 4 September 2026 and no further releases should be expected from a team now inside Notion. The card lists English only while the launch post calls it multilingual. BUILT Qwen3-4B trained by ZeroEntropy's zELO method, which models relevance as Elo ratings.
mxbai-rerank-large-v2 Mixedbread	Germany	OPEN Apache 2.0	1.5B, 0.5B	8K	<i>not published</i>	–	text	Tiny	STRONG Apache 2.0, 109 languages, and the first reranker trained with GRPO reinforcement learning before its contrastive and preference stages. WEAK March 2025 and no v3; Mixedbread's current product is a bundled search API priced per thousand queries, not a standalone rerank rate. Its BEIR 57.49 is below jina-reranker-v3.5. BUILT Qwen2.5 bases, three stages: GRPO, contrastive learning, preference learning.
llama-nemotron-rerank-1b-v2 NVIDIA	US	OPEN OpenMDW-1.1	1B	8K	<i>not published</i>	–	text	Tiny L4 or A10G class GPU	STRONG A 1B cross-encoder under OpenMDW with 737,000 downloads a month, evaluated on 26 languages: Recall@5 73.6% on BEIR plus TechQA and 86.8% cross-lingual on MLQA. WEAK 8K context against 32K and 131K elsewhere. Trained on 800,000 samples, small by current standards, and NVIDIA's newer Nemotron 3 embedding line has no matching reranker yet. BUILT Fine-tuned from Llama 3.2 1B as a cross-encoder on public datasets licensed for commercial use.

Speech: recognition and synthesis 9 listed									
MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	WER	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Whisper large-v3 OpenAI	US	OPEN Apache 2.0	1.55B	30 s windows	<i>not published</i>	5.78	audio	Tiny	STRONG The default open speech recogniser: four million downloads a month, 99 languages, supported by every inference stack. Still the reference point every other row benchmarks against. WEAK Three years old and mid-table at 5.78 against 4.31 for Qwen3-ASR at a similar size. The card itself lists hallucination, repetition and weaker accuracy on accents. OpenAI shuts the hosted whisper-1 on 26 February 2027. BUILT Encoder-decoder Transformer trained on one million hours of weakly labelled audio and four million hours pseudo-labelled by large-v2. Released November 2023 and not updated since.
GPT-Transcribe, GPT-Realtime 2.1 OpenAI	US	CLOSED API only	undisclosed	25MB file	STT \$0.27 / hr		– text, audio	API	STRONG One account covers batch and live transcription, translation and a speech-to-speech model with tool use. gpt-transcribe at \$0.27 an hour is the cheapest hosted rate here after Alibaba. WEAK OpenAI publishes no WER and is absent from the public board. whisper-1 and the gpt-4o transcribe models shut down 26 February 2027, the tts-1 line on 6 January 2027, and the replacement TTS is priced per audio token at \$10 / \$20 per million, so cost no longer maps to character counts. BUILT Not disclosed. gpt-realtime-2.1-mini is described as a distilled reasoning model with a 128K context.
Parakeet TDT 0.6B v3, Canary-Qwen 2.5B NVIDIA	US	OPEN CC BY 4.0	0.6B / 2.5B	24 min	<i>not published</i>	4.86	audio	Tiny 2GB RAM to load	STRONG Throughput: Parakeet runs at 6,076 times real time on the board, the fastest model near the top, within a point of the best closed APIs. Canary-Qwen scores 4.43, the best open English result with a disclosed training set. Attribution-only licence. WEAK Parakeet v3 covers 25 European languages and nothing else; Canary-Qwen is English only. The Granary training data is pseudo-labelled. BUILT Parakeet: FastConformer encoder with a token-and-duration transducer, trained on 660,000 hours of pseudo-labelled Granary data plus 10,000 human-transcribed hours. Canary-Qwen joins a FastConformer encoder to Qwen3-1.7B.

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	WER	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Qwen3-ASR, Qwen3-TTS <small>Alibaba</small>	China	OPEN Apache 2.0	0.6B / 1.7B	—	STT \$0.126 / hr	4.31	text, audio	Tiny	<p>STRONG Best open recogniser on the board at 4.31, Apache 2.0, 30 languages and 22 Chinese dialects. Qwen3-TTS is the most downloaded speech model in the table at 3.5 million a month and clones a voice from three seconds. Hosted ASR at \$0.126 an hour is the lowest list price here.</p> <p>WEAK Cards disclose no training hours or corpus provenance. Hosted prices differ by region, hosted TTS pricing could not be read in English, and the open audio repos were last updated July 2026.</p> <p>BUILT Qwen3-ASR is built on Qwen3-Omni with a custom audio encoder feeding a Qwen3 decoder. Qwen3-TTS is a 12 Hz speech-token language model on a 1.7B backbone.</p>
Voxtral <small>Mistral AI</small>	France	OPEN Apache 2.0; TTS CC BY-NC 4.0	3B / 4B / 24B	3 h per request	STT \$0.18 / hr	4.99	text, audio	Tiny to Small 4B Realtime on a 16GB GPU	<p>STRONG The only European family with an open streaming recogniser, an open audio LLM and a TTS model under one name. Voxtral Realtime had 1.9 million downloads a month, the most of any STT model here.</p> <p>WEAK The hosted Transcribe 2 model is not open and the Voxtral TTS weights are CC BY-NC 4.0, which bars commercial use of the download. Thirteen transcription languages. The realtime 4B scores 6.46 on the board, worse than Whisper. Mistral's pricing page shows no audio prices; the figures are from the launch post.</p> <p>BUILT Voxtral Small extends Mistral Small 3 with a Whisper-style audio encoder; the 4B Realtime is built on Ministral 3. Training data undisclosed.</p>
Scribe v2, Eleven v4 <small>ElevenLabs</small>	US	CLOSED API only	undisclosed	3GB file	STT \$0.22 / hr; TTS \$80 / M chars	3.97	text, audio	API	<p>STRONG Best board score of any vendor selling a general API, in more than 90 languages, with diarisation and entity detection. Eleven v4, released 28 September 2026, claims about 75% listener preference in blind tests and ten-second cloning.</p> <p>WEAK The dearest TTS in the table at \$80 per million characters against Deepgram's \$30, with a launch discount that ends 12 October 2026. Accuracy claims are the vendor's own charts without numbers. Nothing on architecture, size or training data.</p> <p>BUILT Not disclosed. Scribe v2 shipped January 2026 and Eleven v4 replaced v3 as flagship in September.</p>
Kyutai STT, TTS, Pocket TTS <small>Kyutai</small>	France	OPEN CC BY 4.0	2.6B / 1.8B / 100M	2 h	<i>not published</i>	5.57	text, audio	Tiny Pocket TTS on a laptop CPU	<p>STRONG The only non-profit lab here, releasing the whole stack with training code: streaming STT and TTS, the Moshi dialogue model, and Pocket TTS, a 100M model that runs faster than real time on a laptop CPU.</p> <p>WEAK English only for the 2.6B STT and six languages for Pocket TTS, with a fixed 2.5-second delay. Mid-table at 5.57. Voice cloning is locked to precomputed embeddings and the TTS does no watermarking.</p> <p>BUILT All on the Moshi multistream architecture over Mimi audio tokens. STT pretrained on 2.5 million hours of public audio, fine-tuned on 24,000 hours.</p>

MODEL	BUILT IN	WEIGHTS	SIZE	CONTEXT	PRICE IN/OUT	WER	TAKES IN	FITS	WHAT IT IS GOOD AND BAD AT, AND HOW IT WAS MADE
Nova-3, Flux, Aura-2 Deepgram	US	CLOSED API only	undisclosed	–	STT \$0.258 / hr; TTS \$30 / M chars	–	text, audio	API	STRONG The voice-agent incumbent. Flux is built around model-native end-of-turn detection, Flux TTS answers in as little as 80 ms, and the bundled Voice Agent API is \$0.075 a minute. WEAK No public WER and no leaderboard entry; the accuracy claims are against unnamed competitors on Deepgram's own set. Nova-3 dates from early 2025 and the pricing page shows no Nova-4. Nothing on architecture or training data. BUILT Not disclosed.
Saaras V4, Bulbul v3 Sarvam AI	India	CLOSED API only	3B decoder	2 h batch	STT ₹30 / hr; TTS ₹3,000 / M chars	–	text, audio	API	STRONG The sovereign Indian speech stack. Saaras V4, released 25 September 2026, auto-detects and transcribes 22 Indian languages plus English, with language identification error of 5.22% across all 22 by Sarvam's figures. The cheapest STT list price here once converted. WEAK Every accuracy figure is Sarvam's own; nothing is on a public board. Bulbul v3 covers 11 of the promised 22 languages. REST calls are capped at 30 seconds and prices are in rupees only. BUILT Saaras V4 is an audio encoder plus a 3B hybrid state space decoder trained from scratch in-house.

How to read this, and what it does not contain

Method, limits and known gaps. Published because a reference without them is worth less.

What verified means here

Every figure came from the vendor's own page, its model card, or a primary file. Where a vendor publishes no figure, the cell says so rather than carrying a reseller's number. Two claims were checked against primary files rather than press coverage: Rakuten AI 3.0's config.json, and the leaderboard scores. This edition added the LICENSE files of six repositories.

There are no unsourced parameter counts for closed models here, and no prices from aggregators. Modality comes from the model card or config file; speech and embedding scores from the public boards named in their section.

What Fits means

A band derived from total parameters, not a vendor claim. It assumes 4-bit weights, about 0.55 bytes per parameter plus headroom for the KV cache, and says what you can load, not what runs fast.

Tiny. Under about 12B. A desktop GPU with 8GB to 12GB, a Jetson Orin Nano, a phone, or a Raspberry Pi 5 at the low end, where only the 4B-and-under sizes are practical.

Small. 12B to about 35B. One consumer GPU of 16GB to 24GB, a 32GB M4 Mac mini, or a Jetson Orin.

Medium. 35B to about 200B. One 48GB to 80GB datacenter GPU, or a 64GB to 128GB unified-memory machine: an M4 Pro Mac mini at 64GB, a Mac Studio, a Jetson Thor or a DGX Spark at 128GB.

Large. 200B to about 1.2T. A node of two to eight datacenter GPUs, 2x to 8x H100, H200 or B200.

Datacenter. Above about 1.2T. A node of 288GB-class GPUs or more than one node.

API. Closed weights. Nothing to buy; the vendor runs it.

Total parameters set memory even when few are active: a 125B model with 6B active still needs about 70GB. A family that ships several sizes shows a range, Tiny to Small. Where a vendor states a minimum or documents its smallest deployment, it is printed under the band; those figures are usually at BF16 or FP8 and so heavier than the band.

What Takes in means

Input modality as the model card states it. Output is text unless a row says otherwise; embeddings return a vector and speech rows list both directions.

text. Text only. The card or config declares no image, audio or video encoder.

text, image. Images in: screenshots, documents, photos. Video, if accepted at all, as separate frames.

text, image, video. Video as a sequence of frames, by the card's own description.

text, image, audio. Speech or sound in as audio, not transcribed first.

text, image, audio, video. All four. Only Gemini, Muse Spark, MiMo-V2.6 and the omni embedders.

not stated. The vendor's page is silent. Two closed rows.

A modality the API accepts is not the same as one the open weights accept: Qwen3.8-Max's base download is text only while the Max API takes images.

1. **Everything has a shelf life of weeks.** Re-check the vendor page before quoting a number. Nine of these models shipped in the three weeks before this was compiled, and OpenAI replaced one of them after seven days.
2. **Version numbers stopped meaning anything.** OpenAI shipped GPT-6 Sol and GPT-6.1 Sol a week apart and dropped the Terra tier. Google's Flash line is at 3.8 while Pro sits at 3.1 and is still labelled preview. You cannot compare across vendors by reading a version string.
3. **Index rows are per reasoning-effort setting.** A model appears several times on the source leaderboard at different effort levels. Scores here are the highest published setting. Treat them as a tier, not a ranking, and do not build an argument on a two-point gap.
4. **These are v4.3.2 scores.** The index was rescored in September and re-anchored again on 19 September. Higher figures still circulating from launch coverage, GLM-5.3 at 60 and Nemotron 3 Ultra at 47.7, are

1. **Gemini 3.8 Flash pricing is promotional.** The \$0.75 / \$3.75 rate holds through 31 December 2026 and doubles to \$1.50 / \$7.50 on 1 January 2027.
2. **Parameter counts for closed models are mostly unpublished.** Figures circulating for Claude, GPT-6 Astra and Muse Spark are not from the vendors and are not in this file. Microsoft and Yandex are the exceptions that publish one.
3. **DeepSeek prices vary by time of day.** Off-peak is half of peak. Peak is 01:00 to 04:00 and 06:00 to 10:00 UTC on weekdays outside Chinese holidays. Off-peak rates are shown throughout. DeepSeek retired V4-Flash's endpoint on 10 September.
4. **Blank price cells are deliberate.** NVIDIA, Google's Gemma, OpenAI's gpt-oss, IBM, Cohere and most sovereign programmes sell no API, so they have no list price. ByteDance prices in yuan only. A blank cell is not a low number.

from the previous version and do not compare. Three scores here moved by a point or more between 4.3 and 4.3.2.

5. **Open weights is not open source.** It is the most common error in the coverage. Open weights means you can run the parameters. Open source means you can rebuild the thing.
6. **Fits is derived, not reported.** The band comes from total parameters at 4-bit. It is a loading test, not a speed test, and 4-bit quality loss is unmeasured for most of these. Vendor minimums, where printed, are the vendor's own words.

What changed since 16 September 2026

Added

- Claude Opus 5.5 and Sonnet 5.5 replace Opus 5 and Sonnet 5. Opus 5.5 scores 58, the new closed ceiling.
- GPT-6.1 Sol replaces GPT-5.6 Sol and Terra; GPT-6 Luna replaces GPT-5.6 Luna at half the price.
- Grok 4.7 replaces Grok 4.6. Gemini 3.5 Flash-Lite replaces 3.1 Flash-Lite, which Google retires on 7 May 2027.
- MiMo-V2.6-Pro replaces MiMo-V2.5-Pro and is the new best open-weights model at 46.
- GigaChat 3.5 Ultra replaces the 702B GigaChat 3 Ultra preview, two releases behind.
- Laguna S 2.1 (Poolside, US) and the Inkling-Small checkpoint, both missed in September.
- K2 Horizon 375B (UAE, fully open), Motif-3 (Korea, MIT), AliceAI Foundation 80B (Yandex) and Apertus 1.5 70B (Switzerland).

Removed

- GPT-5.6 Terra: superseded 29 September. No GPT-6 Terra exists and GPT-6.1 Sol is cheaper and stronger. Still served.
- Grok-2: a 2024 model, 8,500 downloads a month, no derivatives of note, and a custom no-training licence rather than an open one.
- Phi-4-reasoning-vision: 3,000 downloads a month, no fine-tunes, no index score. Microsoft's later open models are mostly Qwen fine-tunes.

Changed

- Index relabelled v4.3.2. Muse Glimmer 18 to 17, Nemotron 3.5 Lightning 14 to 13, Granite 4.2 9 to 15 (September used the 3B's score).
- Twenty-two rows gain an index score that was blank: every Mistral, DeepSeek V4-Pro and V4.1-Flash, both Gemini, the three returning Korean rows.
- Thirteen prices move from not published to a vendor price, including Qwen3.8-Max, Hy4-preview, MiniMax M3, all three Mistral rows and Muse Spark.
- MiMo price corrected to \$0.435 / \$0.87 from Xiaomi's own page; the September \$1 / \$3 was wrong.
- Licences resolved from repositories: Step-3.7-Flash, Intern-S1-Pro and Inkling are Apache 2.0; openPangu-2.0-Pro is custom and forbids EU use.
- Ling-3.0-Flash declares MIT in metadata only. Latam-GPT is Llama 3.1 Community License. SEA-LION v4.5 is MIT as declared by AI Singapore.
- DeepSeek retired the V4-Flash endpoint on 10 September and routes it to V4.1-Flash. The row stays because third parties still run it at scale.
- Qwen3.8-27B has shipped. GLM-5.3-Flash, not GPT-5.6 Luna, is the most-used model on OpenRouter; Hy4 is second among open models.
- Cohere has not acquired Aleph Alpha; the agreement signed 16 September is pending approval. Solar Open 2 and Mistral Medium 3.5 figures corrected.
- Two new columns on every row, input modality from the model card and a derived hardware-fit band, and three new sections: nine embedding, eight reranker and nine speech families.
- A Tiny band for 8GB to 12GB devices, with six on-device families added: Gemma 4 E2B/E4B/12B, Nemotron 3 Nano 4B, LFM2.5, Qwen3.5 small, MiniCPM5 and Ministral 3.

Known gaps, carried rather than hidden

Qwen3.8-Max. The downloadable weights are the Qwen3.8-2.4T-A95B base, which is feature-reduced against the Max API product. Treat the row as the base model, not the branded product.

Ling-3.0-Flash. MIT is declared in the repository metadata and there is no license file. Step-3.7-Flash, Intern-S1-Pro and openPangu-2.0-Pro were resolved from their repositories this edition.

SEA-LION v4.5. AI Singapore declares MIT on both series, but the Gemma-based one links Google's Gemma license page. Which terms bind you is for AI Singapore to clarify.

Prices read in one region. Alibaba, DeepSeek and Xiaomi quote different prices by region or by hour. International list prices are shown; Alibaba's are lower in Beijing and Frankfurt.

Africa. No significant 2025 or 2026 African foundation-model release was found. That is a research gap rather than evidence of absence.

Announced, not shipped. Reflection AI's Beam (501B, Apache 2.0 stated, announced 5 October) and StepFun's Step 5 Preview have weights promised but not public. Neither is a row until the weights are downloadable.

Sources: vendor pricing pages and model cards, Hugging Face model repositories and its “State of Open Models, Summer 2026” analysis, and the Artificial Analysis leaderboard read live, with the MTEB and Open ASR leaderboards for the tool sections. Refreshed monthly.

aienablement.io/model-atlas