

Xiaomi gives away the second tier of AI

On 22 September Xiaomi published the weights of MiMo-V2.6-Pro, a trillion-parameter agent model, under an MIT licence, with API output at \$0.87 per million tokens. It ties Grok 4.7 on the Artificial Analysis Intelligence Index and leads every open rival, yet sits 12 points behind the closed frontier. Xiaomi has not caught the frontier. It has made the tier just below it free to download, and that tier is where most paid agent work sits.

AUTHOR

AT Alicante Tech Vibes *Research Labs*

~8 MIN READ · ALL FIGURES SOURCED & DATED

46 PRO ON INTELLIGENCE INDEX, TOP OPEN WEIGHTS · ARTIFICIAL ANALYSIS, SEP 2026	\$0.435/\$0.87 PRO API, PER M INPUT/ OUTPUT TOKENS · XIAOMI, 22 SEP 2026	58 FRONTIER LEADER CLAUDE OPUS 5.5, SAME INDEX · TNW, 2 OCT 2026
\$0.13 PRO COST PER INDEX TASK; GPT-6 ASTRA \$3.26 · AA, SEP 2026	\$2.62M PRO'S FINAL RL TRAINING RUN, UNDER SIX DAYS · XIAOMI, 22 SEP 2026	88s vs 18s MEDIAN WALL TIME, PRO VS GPT-6 SOL · COMPUTINGFORGEEKS, 25 SEP 2026

I. A phone maker shipped a trillion-parameter agent for free

On 22 September, Xiaomi released the MiMo-V2.6 series: a Pro model, a Flash model and a 9-billion-parameter distilled version, with weights on Hugging Face. VentureBeat reported the specifications: Pro is a sparse mixture-of-experts model with 1.02 trillion total and 42 billion active parameters; Flash has 310 billion and 15 billion. Both take a one-million-token context and carry the MIT licence, which permits commercial use and modification.

The API prices are \$0.435 and \$0.87 per million input and output tokens for Pro, and \$0.14 and \$0.28 for Flash. Xiaomi says the series keeps V2.5's pricing and costs "1/20 to 1/60" of overseas models at the same intelligence level.

Xiaomi also published what the final training stage cost: about \$2.62 million of reinforcement learning for Pro and \$850,000 for Flash, each run covering roughly 750,000 trajectories in under six days. That figure covers the last RL stage only, not pre-training, and it should not be read as the cost of building the model.

The headline claim is that the world's best open model now matches a closed one. That is true, and the closed model it matches is not the frontier.

II. The tie is with the second tier, not the top

Artificial Analysis scores MiMo-V2.6-Pro at 46 on its Intelligence Index. Unite.AI's roundup of the leaderboard puts that level with xAI's Grok 4.7 and ahead of the strongest other open models, GLM-5.3 at 45 and Kimi K3 at 44.

The closed frontier sits higher. On 2 October The Next Web reported the same index with Claude Opus 5.5 at 58, Gemini 4 Argon and GPT-6 Astra at 53, and GPT-6.1 Sol at 52. Xiaomi's own announcement concedes a gap to the strongest closed models. The open-weight lead is real, but it is a lead of one or two points over other open models, and it trails the frontier by twelve.

The open leaderboard also turns over fast. In May, Xiaomi was already saying its previous model, V2.5-Pro, ranked first among open models on Artificial

Analysis. Being first among open weights is a position that changes hands every few months between Chinese labs, not a lasting advantage.

FIG. 1: ARTIFICIAL ANALYSIS INTELLIGENCE INDEX, SELECTED MODELS



BARS SCALED TO THE HIGHEST SCORE (58). * MARKS OPEN-WEIGHT MODELS. INDEX SCORES ARE ARTIFICIAL ANALYSIS MEASUREMENTS, NOT VENDOR CLAIMS. SOURCES: THE NEXT WEB, 2 OCTOBER 2026; UNITE.AI, 23 SEPTEMBER 2026.

III. The honest technical read

The agentic-coding numbers are strong and mostly vendor-run, and the sources disagree on them. [Xiaomi's announcement](#) says Pro rose on DeepSWE v1.1 from 58.4 to 72.6 during the final RL run. [VentureBeat reports 71.9](#). [Unite.AI reports 72.57](#) and compares it with 19.0 for the older V2.5-Pro. These describe two different baselines: one is the same model before its last RL run, the other is the previous generation. On Unite.AI's table, Claude Opus 5 and GPT-6 Astra both score 74.0, and DeepSeek V4.1 Flash, another open model, scores 74.2. So Pro is not the open leader on this test.

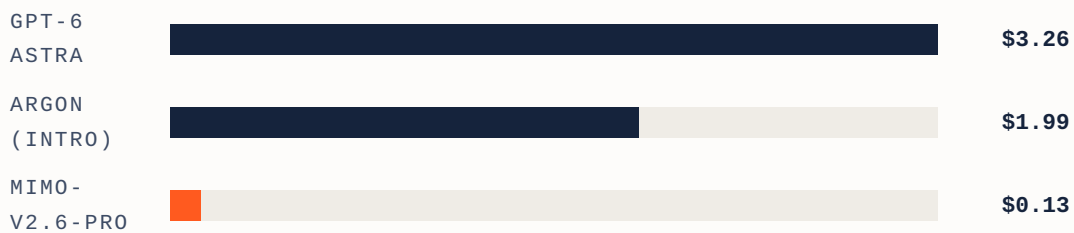
Independent hands-on tests found weaker results. [MindStudio put Pro at 34.9 on Terminal Bench 4.0](#) against 49.0 for Claude Opus 5, and 47.9 on ExploitBench against 70.0. One debugging task took more than 25 minutes. [ComputingForGeeks ran infrastructure tasks on 25 September](#): Pro produced a working Kubernetes deployment in one of three runs, and its median wall time was 88 seconds against 18 for GPT-6 Sol. One request hung for more than ten minutes.

Latency is a real constraint. [Tabbit measured about 18 seconds to first token](#) on a 10,000-token input, most of it spent thinking. That is acceptable for

overnight agent jobs and poor for interactive chat. Reported output speed also varies by source, from about 40 tokens per second on Artificial Analysis to about 130 on Unite.AI's figures.

Self-hosting is not cheap either. ComputingForGeeks [lists Pro's weights at 566 GB](#), served across eight GPUs with tensor parallelism; Flash needs 173 GB and at least two 80 GB GPUs. “Free weights” means no licence fee. It does not mean free to run.

FIG. 2: COST TO RUN ONE ARTIFICIAL ANALYSIS INDEX TASK



BARs SCALED TO THE HIGHEST COST (\$3.26). INDEX SCORES DIFFER (ASTRA AND ARGON 53, MIMO 46), SO THIS IS NOT A LIKE-FOR-LIKE COMPARISON. SOURCES: ARTIFICIAL ANALYSIS VIA UNITE.AI, SEPTEMBER 2026; THE NEXT WEB, 2 OCTOBER 2026.

IV. The squeeze lands on the \$2/\$10 middle

Prices. Closed mid-tier models now cluster at [\\$2/\\$10 per million tokens](#) for Claude Sonnet 5.5 and GPT-6.1 Sol, with Opus 5.5 at \$4/\$20. MiMo-V2.6-Pro's output tokens cost about one-eleventh of the mid-tier rate and one twenty-third of Opus. Per indexed task the gap is about 25 times against GPT-6 Astra, though for a lower score. The model competes with Sonnet-class pricing. It does not compete with Opus-class capability.

Market structure. In our reading, most enterprise agent work runs on mid-tier models: classification, extraction, routine code changes and batch pipelines. A free model that matches Grok 4.7 lowers the ceiling on what closed labs can charge for that tier, while the frontier keeps its premium. Expect closed labs to defend the middle with cheaper SKUs and longer introductory pricing, not to lose it overnight.

Build versus buy. Our estimate, not a sourced figure: assume eight 80 GB GPUs rented at \$2.50 per GPU-hour. Self-hosting Pro then costs about \$14,600 a month. At \$0.87 per million, the same money buys roughly 16.8 billion output tokens from Xiaomi's API. Below that volume, the API is cheaper. Self-hosting makes sense for very heavy users, or for firms that need data to stay on their own hardware.

Capital. Xiaomi is paying for this out of hardware margins. Lei Jun committed more than 60 billion yuan to AI over three years, including more than 16 billion yuan in 2026, and Xiaomi's shares fell more than 6% on the day. In the first quarter, revenue fell 10.9% to 99.1 billion yuan and adjusted net profit fell 43.1%. A company with shrinking profit is giving away a model to win developers for its phones, cars and its own API, not to earn money from the model itself.

Labour. The tasks a cheap 46-level model handles well are high-volume, verifiable and tolerant of delay. That puts pressure on outsourced routine development and data work first. Senior engineering is less exposed, because that work still depends on frontier models and fast feedback.

Xiaomi did not catch the frontier. It made the tier just below it free to download, and that tier is where most of the paid tokens sit.

MIMO-V2.6-PRO SCORES 46 ON THE ARTIFICIAL ANALYSIS INDEX, LEVEL WITH GROK 4.7 AND 12 POINTS BEHIND CLAUDE OPUS 5.5, AT ABOUT \$0.13 PER TASK. SEPTEMBER-OCTOBER 2026.

V. Second-order: open tokens, closed revenue

Open models already carry most of the traffic without earning most of the money. Open-weight models account for about 60% of OpenRouter's US token consumption, most of them Chinese, according to The New Stack on 14 September. But a Linux Foundation analysis of OpenRouter traffic, cited by Tech Times, found closed providers took about 96% of model-layer revenue from May to September 2025. That data is a year old. It is still the strongest evidence against the idea that cheaper open models drain closed-lab revenue.

Procurement is the bottleneck. The same New Stack report cites Deloitte finding that 77% of companies weigh a vendor's country of origin. OpenRouter now offers US-only routing for Chinese models hosted by US providers. A Chinese-built model only reaches regulated buyers after a Western host runs it, which hands part of the margin to the infrastructure layer.

The release also lowers costs for other labs. Xiaomi says it will open-source more than 7,000 RL task environments and its training framework. If that happens, the cost of post-training an agent model falls for every competitor, open or closed. And Xiaomi's 9B distilled model is a cheaper path to fine-tuned agents at the edge, though ComputingForGeeks found it failed most of its infrastructure tasks.

VI. Precedent: the DeepSeek shock, with less noise

In January 2025, DeepSeek's R1 release coincided with Nvidia losing about \$589 billion in market value in one day. The fear was that cheap training would cut demand for chips.

We found no report of a comparable market reaction to MiMo-V2.6, and that is the point. Investors now expect a cheap Chinese open model a few points behind the frontier every few months. The effect shows up slowly, in mid-tier price lists and enterprise model routing, rather than in a single day's share price.

VII. How it plays out: three scenarios

Probabilities are our analytical judgment, not forecasts from a model, and they sum to 100%.

1 · Baseline: the middle tier commoditises ~55%

12-24 MONTHS ·
STEADY

Open models stay about 8–12 index points behind the frontier while matching the closed second tier. Closed mid-tier list prices drift down, and frontier prices hold. Enterprises send bulk agent work to open models through Western hosts and keep frontier models for hard tasks. Watch for: Sonnet- or Sol-class prices cut below \$2/\$10, US hosts adding MiMo-V2.6-Pro to compliant routing, and the open model share of OpenRouter tokens holding near 60%.

2 · Acceleration: open weights close the gap ~20%

9-18 MONTHS · FAST

Open RL environments and cheap post-training let the next Chinese open release come within about five points of the frontier. Closed labs then have to cut prices across the whole lineup. Watch for: an open model scoring above 52 on the index, independent Terminal Bench results within five points of Opus, and a large enterprise publicly moving production work to open weights.

3 · Stall: tokens without money ~25%

6-12 MONTHS · QUIET

Latency, reliability and country-of-origin rules keep open models in hobby and batch use. Closed labs keep almost all the revenue, and Xiaomi slows its releases or tightens the licence as AI spending weighs on profit. Watch for: no Western enterprise host listing MiMo, Xiaomi reducing its AI spending guidance, and the next V2.7 or V3 release shipping under a more restrictive licence.

WHAT WOULD FALSIFY THIS THESIS

The claim is that MiMo-V2.6 commoditises the closed second tier without threatening the frontier. It fails in one direction if independent agentic tests put Pro within a few points of Opus 5.5. Then the frontier itself is exposed, and this reading is too cautious. It fails in the other direction if closed mid-tier prices and revenue shares stay flat through mid-2027 despite free 46-level weights. That would mean the binding constraints are latency, reliability and procurement, not price.

WHAT TO WATCH, IN ORDER

1. **By 31 October 2026** — whether US hosts such as Fireworks or Baseten serve MiMo-V2.6-Pro in OpenRouter's US-only routing. This is the procurement test.
2. **By 30 November 2026** — whether Xiaomi actually releases the 7,000-plus RL environments and training framework it promised, which would lower post-training costs for every lab.
3. **Xiaomi Q3 2026 results, expected late November 2026** — AI spending against the 16-billion-yuan guidance, and whether management puts a figure on MiMo API revenue.
4. **Closed mid-tier pricing, Q4 2026** — whether Claude Sonnet 5.5 and GPT-6.1 Sol keep \$2/\$10 after their introductory windows end, or go lower.
5. **Next major Chinese open release, by Q1 2027** — whether DeepSeek, Moonshot or Zhipu scores above 50 on the Artificial Analysis index, narrowing the gap to the frontier.

SOURCES

1. Xiaomi MiMo, "MiMo-V2.6 series release," 22 September 2026 · mimo.mi.com
2. VentureBeat, "'Better than DeepSeek': Xiaomi's MiMo-V2.6-Pro debuts as the top open weights model in the world alongside cheaper V2.6-Flash," 21 September 2026 · venturebeat.com
3. Artificial Analysis, "MiMo-V2.6-Pro: Intelligence, Performance & Price Analysis," accessed 4 October 2026 · artificialanalysis.ai
4. Unite.AI, "Xiaomi's New Flagship Model Leads Open-Weight Rankings With a Score of 46," 21 September 2026 (updated 23 September) · unite.ai
5. WinBuzzer, "Xiaomi MiMo V2.6: downloadable AI with low-cost APIs," 24 September 2026 · winbuzzer.com
6. Latent Space, "[AINews] Xiaomi MiMo-V2.6-Pro 1T-A42B: the new top Open Weights model, trained for \$3M," 22 September 2026 · latent.space
7. The Next Web, "Gemini 4 Argon matches GPT-6 Astra and hallucinates far less in early tests," 2 October 2026 · thenextweb.com
8. FinanceFeeds, "Google Prices Gemini 4 Argon at \$2/\$10, the Same Rate as Claude Sonnet 5.5 and GPT-6.1 Sol," 1 October 2026 · financefeeds.com
9. ComputingForGeeks, "MiMo-V2.6 Pro and Flash: Benchmarks, Price, Tested," 25 September 2026 · computingforgeeks.com

10. MindStudio, "Xiaomi MiMo-V2.6-Pro Tested: Does the 1T-Param Open Model Hold Up?," 22 September 2026 · mindstudio.ai
11. Tabbit, "MiMo-V2.6-Pro Review: The Smartest Open Model Makes You Wait," 22 September 2026 · go.tabbit.ai
12. 36Kr, "Lei Jun's Plan to Invest 60 Billion Yuan in AI in Three Years Spooks the Market," 20 March 2026 · 36kr.com
13. 36Kr, "Xiaomi Q1 Performance: 99.1 Billion Yuan Revenue ... 60 Billion Yuan AI Investment in Next Three Years," 27 May 2026 · 36kr.com
14. The New Stack, "Chinese AI models dominate OpenRouter's US token consumption. It can now guarantee that traffic stays entirely in the US," 14 September 2026 · thenewstack.io
15. Tech Times, "Chinese Open-Weight AI Handles Most Developer Tokens: Closed Models Capture 96% of Revenue," 17 September 2026 · techtimes.com
16. Forbes, "Biggest Market Loss In History: Nvidia Stock Sheds Nearly \$600 Billion As DeepSeek Shakes AI Darling," 27 January 2025 · forbes.com

METHOD NOTE: FIGURES ARE ATTRIBUTED AND DATED INLINE; WHERE SOURCES DISAGREE (DEEPSWE SCORES AND BASELINES, OUTPUT SPEED), THE DISAGREEMENT IS REPORTED RATHER THAN AVERAGED. INTELLIGENCE INDEX SCORES ARE ARTIFICIAL ANALYSIS MEASUREMENTS; DEEPSWE FIGURES ARE VENDOR-RUN AND LABELLED AS SUCH. THE SELF-HOSTING COST COMPARISON IS OUR ESTIMATE UNDER A STATED GPU-RENTAL ASSUMPTION. SCENARIO PROBABILITIES ARE ANALYTICAL JUDGMENT.

Built by the [Alicante Tech Vibes](#) community, a tope.

2026-10-04