

Frontier AI Models 2026

Landscape, Projections & Industry Implications

• POINT-IN-TIME SNAPSHOT · MODEL RANKINGS & PRICING AS OF JULY 2026 · THE FRONTIER MOVES MONTHLY — SEE THE EDITOR'S NOTE IN SECTION 03

A senior-researcher assessment of where the frontier model market stands, where it is heading over the next three years, and what it means for industries and allocators — with a focus on **Reka AI's** position in physical-world intelligence.

2.8T

Parameters in Kimi K3 — the first open-weight model at frontier scale

~50x

Median annual fall in the price of a fixed level of intelligence

\$407B

Global enterprise AI spend, 2026 — up 35% on 2025

72%

Enterprises running AI in production, 2026

IN THIS ISSUE

- | | |
|---|---|
| 01 Executive Summary — the race has split; price is the weapon | 02 The Market Map — four camps of frontier AI |
| 03 State of the Frontier — who leads today | 04 Market Signal & Demand — intelligence commoditised |
| 05 From Pilots to Production — where the money goes | 06 Token Economics — unit of production, not value |
| 07 Projections 2026–2029 — where the frontier goes next | 08 Industry Implications — sector by sector |
| 09 Company Focus — Reka AI, the physical-world specialist | 10 Risks & Watch Items — what could change the picture |

01 — EXECUTIVE SUMMARY

SNAPSHOT · JUL 2026

The race has split, and price is the weapon

The frontier AI market is no longer a single race for the smartest model. It has split into four camps, defined by two questions: is the model **open or closed**, and is it built for **everything or for something specific**. July 2026 made this split impossible to ignore — Moonshot AI's Kimi K3, a 2.8-trillion-parameter open-weight model, now sits within touching distance of the best closed US systems on independent leaderboards, while pricing far below them. When capability converges and prices collapse, general-purpose intelligence becomes a commodity.

01 · The closed frontier still leads — narrowly

Claude Fable 5 and GPT-5.6 Sol top independent rankings, but the margin over the best open model is now single benchmark points, not generations.

02 · Open weights reached frontier scale

Kimi K3 is first-in-class at ~3-trillion parameters, with a 1-million-token context window and full weights promised publicly.

03 · Price deflation is brutal

Frontier-level output now costs roughly a third to a half of comparable US closed pricing; inference prices are falling at double-digit multiples per year.

04 · Enterprises crossed the tipping point

~72% now run AI in production — but most value is still concentrated in a small minority of mature deployments; adoption has outpaced measurable P&L impact.

05 · The next frontier is physical

World models, robotics, sensing and infrastructure intelligence are where scale alone does not win. NVIDIA and labs including Reka are building here — and sovereign buyers, including the GCC, have structural demand for exactly this category.

CORE THESIS

Durable value in AI is shifting from raw scale to **ownership** — of a hard domain, proprietary data, a physical deployment surface, and trusted customer relationships. That is precisely the ground **Reka AI** is consolidating in physical-world intelligence.

02 — THE MARKET MAP

Four camps of frontier AI

Two axes organise the field. Horizontal: closed, proprietary models versus open-weight models anyone can download. Vertical: general-purpose intelligence versus specialised, domain-focused intelligence.

GENERAL-PURPOSE ← · → SPECIALISED / DOMAIN

Closed Frontier Leaders**Claude Fable 5 · GPT-5.6 Sol · Gemini · Grok**

Most capable, most expensive — a capital war measured in billions of dollars of compute.

Proprietary Specialists**REKA (physical AI) · Cohere · Thinking Machines**

Win by owning a domain, its data and its customers — not by outscaling the giants.

Open-Weight Challengers**Kimi K3 · DeepSeek V4 · GLM · Qwen · Llama**

Closing the gap fast at a fraction of the price.

Open Efficiency Plays**Mistral · Phi · MiniMax**

Small, fast, cheap models for edge and cost-sensitive use.

TOP ROW: PROPRIETARY · BOTTOM ROW: OPEN-WEIGHT

READING THE MAP

The top half of the map is a capital war — the average cost of training a frontier model has reached roughly **\$200 million, ten times the 2022 level**. The bottom-left quadrant is won differently: by owning a vertical, its data, and its deployment surface. That is where Reka operates.

Who leads today

On independent evaluations **as of July 2026**, the closed US frontier still holds the top of the table — Anthropic's Claude Fable 5 and OpenAI's GPT-5.6 Sol lead most aggregate rankings, with Google's Gemini and xAI's Grok in the leading pack. But the story of the month is who sits immediately behind them.

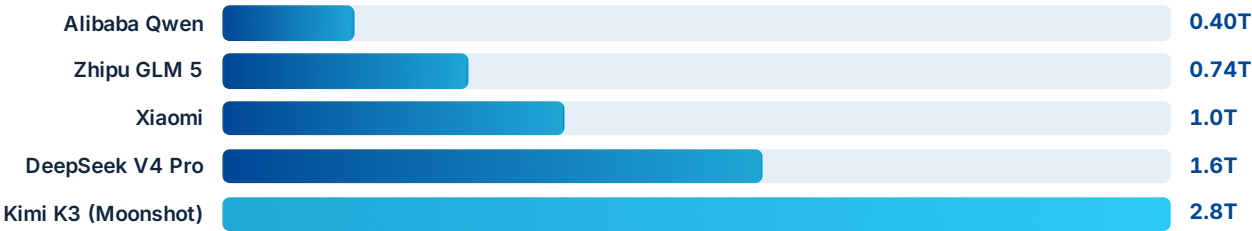
The challenger — at this issue's cut-off

Kimi K3 ranks **second on the Vals AI index and third on Artificial Analysis's Intelligence Index** — beaten only by Fable 5 and GPT-5.6 Sol — and first outright on frontend code benchmarks, ahead of models like Claude Opus 4.8. An open-weight challenger, built with far fewer resources under hardware export restrictions, is now trading blows with the best labs in the world.

EDITOR'S NOTE · SEPTEMBER 2026 REISSUE

The frontier has already moved past this snapshot — proof of this issue's own thesis that leads are now measured in months, not generations. Since the July cut-off: **Anthropic's Claude Opus 5** has taken the top of the aggregate indices, **xAI's Grok 4.6** (released August 12) landed just behind the leaders at roughly a third of their price, and **Claude Fable 5.1** followed in September. K3 remains the leading open-weight model, but the specific rankings on this page describe July 2026 — treat every leaderboard claim in this issue as time-stamped, not current.

OPEN-WEIGHT MODELS — HOW FAR SCALE HAS JUMPED (TOTAL PARAMETERS)



How K3 does it

A sparse mixture-of-experts architecture activates only **16 of 896 experts per token** (about 1.8% of the model), so inference costs behave like a much smaller model. Two in-house innovations — Kimi Delta Attention and Attention Residuals — plus a 1-million-token context window round out the package.

04 — MARKET SIGNAL

PRICING · JUL 2026

Intelligence is being commoditised in real time

K3 is priced at \$3 per million input tokens and \$15 per million output — the highest of any Chinese lab, yet still roughly **half the per-task cost of comparable US closed systems**, and about a third of comparable US output pricing. Industry-wide, the price of a fixed level of intelligence has been falling at a median of roughly **50x per year**.

FRONTIER-LEVEL ANSWERS ARE GETTING CHEAP (RELATIVE COST PER TASK, K3 = 1.0)**THE MESSAGE FOR ALLOCATORS**

Betting on "the smartest model" is now a bet against relentless, well-funded convergence from both sides of the Pacific. The strategic question for any lab is no longer "how large is your model?" but "**what do you own that scale cannot replicate?**" The enduring answers: proprietary data in a hard domain, physical deployment surfaces, embedded enterprise relationships, and regulatory or sovereign trust. Every projection in this report follows from that shift.

05 — DEMAND SIDE

Enterprises have crossed the tipping point

\$407B

Global enterprise AI spend, 2026

+35%

Growth vs 2025 (\$302B)

~\$632B

Projected global AI spend by 2028

72%

Enterprises with AI in production, 2026

From pilots to production

AI HAS CROSSED INTO EVERYDAY BUSINESS USE (% OF ENTERPRISES WITH AI IN PRODUCTION)



WHERE ENTERPRISE AI MONEY GOES IN 2026 (SPEND, US\$B)



Caveat one: value is concentrated

Most generative-AI pilots still fail to reach measurable P&L impact, and only a minority of enterprises run AI at scale across functions. Adoption has outpaced returns — boards are starting to notice.

Caveat two: budgets are rotating

Spend is moving from raw API access toward **agentic systems, domain-tuned models, and integration work**. Both trends favour specialists embedded in customer workflows over general-purpose chatbots.

Tokens: AI's unit of production, not yet its unit of value

Every figure in this report ultimately clears through one meter: the token. As Issue 01 established, the blended market price of a fixed unit of intelligence has already fallen from roughly **\$17 to \$2 per million tokens** since 2023. What is new in 2026 is that **volume is now compounding faster than price is falling** — which is why total AI spend keeps rising even as the unit price keeps dropping.

~\$1.10

Blended \$/1M tokens,
industry median, mid-
2026

14x

YoY growth in global
inference token volume

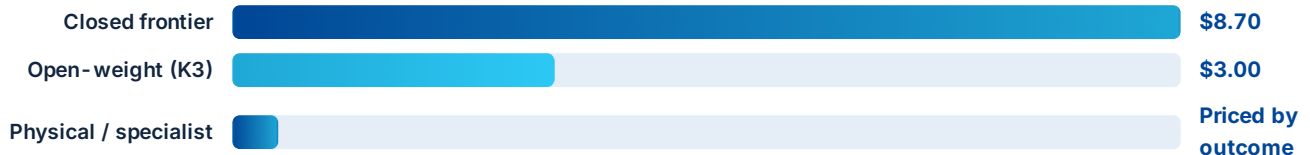
1.2–1.8

Price elasticity of token
demand

12 → 36

Tokens processed per
output token, chat →
agents

TOKEN PRICING BY CAMP (\$ PER 1M TOKENS, BLENDED, ILLUSTRATIVE)



Closed labs still command roughly a 3x per-token premium over K3 — compressing every quarter as open weights close the capability gap.

Who gains as tokens commoditise

- ✓ Compute owners and GPU clouds — volume compounds faster than price falls.
- ✓ Agentic-application builders — cheap tokens make multi-step workflows affordable.
- ✓ Specialists whose output isn't priced per token — Reka Vision bills for security outcomes, not tokens.

Who is exposed

- ✗ Pure API resellers with no proprietary data or workflow lock-in — margin compresses toward inference cost.
- ✗ Closed labs whose only moat is raw benchmark leadership, as the open-weight gap narrows.
- ✗ Anyone pricing on token volume alone, with no line of sight to the customer's actual outcome.

THE AGENTIC TAX

A reasoning model spends tokens thinking; an agent spends tokens planning, calling tools, checking its own work, and retrying — **12 tokens processed per output token for chat, 36 for agentic workflows**. Total tokens grew +39x vs output tokens +30x (Jan 2025 → Apr 2026); global inference volume runs 2.1Q → 30Q per month. Volume, not price, is now the swing factor: a 10% price cut pulls through 12–18% more token volume. Reka doesn't sell tokens — **Reka Vision is priced on deployed cameras, sensor feeds and detected events** — the clearest structural insulation from the price war.

07 — PROJECTIONS, 2026–2029

Where the frontier goes next

I · General intelligence becomes utility-like

Open-weight releases at frontier scale compress closed-model pricing power. Expect margins on raw token sales to keep thinning, with closed labs defending through agents, integrations, enterprise trust and safety differentiation rather than benchmark leads alone.

II · Agentic AI moves from pilots to processes

Analysts expect roughly **a third of enterprise applications to embed agentic AI by 2028**, up from under 1% in 2024. Multi-step, low-supervision workflows — coding, supply chain, finance operations — are the growth engine of model demand.

III · Physical AI becomes the new scaling race

World models — systems that understand motion, physics and consequence from video and sensor data — are the least commoditised frontier. NVIDIA's Cosmos platform and labs like Reka are building here; robotics, autonomy, security and smart infrastructure are the deployment surfaces.

IV · Small, specialised models eat targeted workloads

Specialised small models are projected to capture roughly **35% of enterprise inference by 2028 at 60–80% lower cost** than frontier LLMs — reinforcing the split between commodity general intelligence and premium domain intelligence.

V · Sovereign AI hardens as a category

Governments increasingly treat model access, data residency and infrastructure intelligence as national capabilities. Open weights accelerate this — states can host and adapt frontier-class systems while trusted specialist vendors win security-critical work.

VI · Compute and energy remain the binding constraint

AI infrastructure spending reached ~\$98B in 2026, with AI power demand heading toward several percent of US electricity. Regions that can deliver energy, land and capital at scale — the Gulf prominent among them — gain durable strategic leverage.

08 — INDUSTRY IMPLICATIONS

What this means, sector by sector

Financial services

The biggest spender (**\$68B in 2026**, ~79% adoption; roughly \$3,200 per employee). Cheap frontier-level reasoning compresses the cost of research, compliance and client servicing; the edge shifts to proprietary data and distribution. Expect consolidation of vendor stacks and rapid rollout of agentic middle-office automation.

Healthcare

~\$45B spend and adoption past 60%, anchored by imaging AI now matching specialist accuracy on narrow tasks. Regulatory trust and clinical data access are the moats; open models lower cost but raise the bar on validation and liability.

Manufacturing & logistics

The fastest-growing spender (**+48% YoY**), driven by predictive maintenance and computer-vision quality control. This is the front line of physical AI: world models and video understanding convert directly into fewer line stoppages and defects.

Infrastructure, security & smart cities

The purest physical-AI market. Camera and sensor networks at giga-project scale need systems that interpret the physical world in real time — mission-critical, recurring, sovereign-trust revenue. **This is Reka's wedge** (see Section 09).

Energy & compute

AI is a load-growth story for utilities and a demand story for chips (NVIDIA ~78–82% of training GPUs). Gulf economies sit on both sides of the trade — supplying energy and capital, and buying intelligence for the world's largest infrastructure build-out.

Software & knowledge work

Long-horizon coding is the frontier's proving ground — K3 and the top closed models now complete extended engineering tasks with minimal supervision. Expect 40% of working hours to be touched by LLMs, with the deepest impact where workflows are digitised end-to-end.

09 — COMPANY FOCUS · SINGULAR LINK PORTFOLIO

Reka AI: the physical-world specialist

REKA IN ONE PARAGRAPH

Founded 2022 in Sunnyvale by researchers from DeepMind, Google Brain and Meta FAIR — led by CEO **Dani Yogatama** and Chief Scientist **Yi Tay** — Reka is an AI research lab building foundational intelligence for the physical world: world models, multimodal reasoning, and AI that perceives and acts through cameras, sensors and video. Strategic backers include **NVIDIA**, **Snowflake** and **DST Global**.

I · The Moonvalley combination

The June 2026 merger brings researchers from DeepMind, Meta, Amazon, Microsoft, Wayve and Runway — including key contributors to the models behind Google's Veo — into Reka's research leadership. In one move, Reka acquired world-model and video-generation depth that would take years to hire organically.

II · A live commercial wedge: Reka Vision

Reka's multimodal physical-security platform interprets live camera and sensor networks for municipal and infrastructure buyers. The company reports investigation speeds improved up to **20x** and false alarms cut by **95%** — a concrete enterprise value proposition, not a consumer chatbot competing on benchmarks.

III · A research engine pointed one way

The combined team is focused on the **World Language Action Model (WLAM)** — an omni model trained on egocentric and physical-world data that can simulate motion, physics and consequences before acting, spanning robotics, wearables and real-time decision-making. Roadmap, product and team all point the same direction — a coherence rare among mid-sized labs.

09 — COMPANY FOCUS · CONTINUED

Why Reka matters for Gulf capital

Reka's chosen quadrant maps almost one-to-one onto GCC national priorities. The Kingdom and its neighbours are building the world's largest concentration of new physical infrastructure — giga-projects, smart cities, ports, and critical facilities — all requiring exactly what Reka builds: AI that understands and secures the physical world at scale.

Sovereign fit

Physical security and infrastructure intelligence are domains where governments prefer trusted, dedicated partners over consumer platforms.

Deployment surface

Every smart-city and giga-project rollout in the region is a potential Reka Vision deployment — recurring, mission-critical revenue.

Ecosystem leverage

Backing from NVIDIA and Snowflake connects Reka to the compute and enterprise-data stack already adopted across the Gulf.

Insulation from the price war

As open models commoditise general intelligence, Reka's value rests on domain data, deployed systems and customer trust — assets a cheaper general model cannot replicate by scale alone.

THE BOTTOM LINE

While the giants fight a capital war for general intelligence, Reka has chosen a quadrant it can lead — and is consolidating it with the right team, a live product, and a research engine all pointed at the physical world. For Gulf investors, it is one of the cleanest pure-play exposures to physical AI in the private market today.

10 — RISKS & WATCH ITEMS

What could change the picture

Convergence risk for closed leaders

If open models keep closing the gap, pricing power at the closed frontier erodes faster than agent and enterprise revenue can replace it.

Big-tech entry into physical AI

NVIDIA's Cosmos world-model platform and hyperscaler robotics programs could crowd the physical-AI quadrant; specialists must convert their head start into deployed, sticky systems quickly.

The ROI reckoning

With most pilots still failing to show measurable P&L impact, boards are demanding returns. A funding pause for underperforming projects would slow demand — though surveys suggest very few executives plan to cut AI budgets even so.

Geopolitics and export controls

Three years of hardware restrictions have not stopped Chinese labs reaching the frontier. Policy responses — in either direction — could reshape model availability, pricing and sovereign procurement overnight.

Compute and energy bottlenecks

GPU scarcity and power constraints remain real; delays already push out enterprise projects and could cap the pace of the next scaling wave.

Verification of vendor claims

Performance figures across the industry — including company-reported product metrics — are largely self-reported and should be validated in diligence.

Reissued September 2026 in the unified Singular Link Intelligence design. Model rankings, capability comparisons and token pricing throughout are a **point-in-time snapshot as of the July 2026 cut-off** and are labeled as such; releases after the cut-off — including Claude Opus 5, xAI's Grok 4.6 (August 12, 2026) and Claude Fable 5.1 (September 2026) — have shifted the leaderboard since, as noted in the Section 03 editor's note.

Figures on frontier models, pricing, token economics and enterprise adoption are drawn from public model evaluations, vendor pricing pages, and industry infrastructure-spending research current as of July 2026, and build on Singular Link Intelligence Issue 01's token-economy baseline. Company-reported performance metrics (including Reka Vision speed and accuracy figures) are as stated by the companies concerned and have not been independently verified. Analysis, sector framing, and the Gulf-capital lens are Singular Link's own.

This material is provided by Singular Link for informational purposes only. It is not investment advice, nor an offer to sell or a solicitation of an offer to buy any security or interest in any fund. Company names shown, including current portfolio positions, are illustrative of Singular Link's thematic focus and should not be taken as a recommendation, a complete list of holdings, or an indication of future portfolio composition. Past performance is not indicative of future results.