

THE BIG READ

Same list price, three different bills

KEY SIGNALS THIS PERIOD

6

Models added to the tree

8

Vendors shipped frontier-class

4

Architectural patterns crossed multiple vendors

6

Benchmark moves

AUTHOR

Brian Letort
BrianLetort.AI

PUBLISHED

October 3, 2026
Issue 24 · Weekly read

WEB

brianletort.ai/industry/model
The Model Pulse archive

THE BIG READ

The thesis this issue defends.

Independent industry analysis. Compiled from public model cards, vendor blogs, and leaderboards. The body sets the read for the rest of the brief.

A 'model' is now a policy-wrapped system, and its price is a set of meters rather than a list rate: that is what this week's three mid-tier releases show, and it changes how you evaluate and how you budget. Start with the gates, because each lab shipped a different kind. Google's is an access gate: Gemini 4 Argon exists, scores on independent indices, and cannot be bought. Anthropic's is a routing gate: Sonnet 5.5 can hand a higher-risk cyber request to Sonnet 5 mid-session, visibly, through classifiers in the request path. OpenAI's is a classification gate: Sol inherits GPT-6 Astra's safeguards stack with a Critical rating in cybersecurity (OpenAI's highest internal capability tier, which triggers its strictest deployment controls), and the planned Astra successor was withheld with no numbers. The September 25 pause covers an unreleased most-capable tier; GPT-6 Astra itself still serves dots and the Ultrafast tier. An evaluation harness (the software that wraps the model, supplies its tools and runs the benchmark) that assumes a fixed model behind an endpoint will produce numbers that do not reproduce in production.

The meters are the second half. For a session that rereads a large context many times, the cache-read meter alone can make the Sol and Sonnet 5.5 bills diverge by 2x behind the same \$2 / \$10 list price; for a single long-document pass, Sol's long-context repricing does the opposite. Argon's \$2 / \$10 is a 50% introductory discount for at least one month, and TokenCost noted that its post-intro card is Opus 5.5's card line for line; DataCamp and TokenCost both published the meters-over-list-price read before this issue, and synthorai added the multiple the meter table does not show, that Sonnet 5.5 writes 1.7-3.1x the output tokens of Sol on comparable tasks. The full per-meter table, with the cache-read rates and the 272K threshold, is in the architecture watch below. Above the mid tier on the independent index, Opus 5.5 at \$4 / \$20 is the only buyable model that outscores Sonnet 5.5, and the question for most agent budgets is whether its two extra index points are worth doubling the bill; GPT-6 Astra at \$10 / \$50 is buyable and scores below Sonnet 5.5, which is the stronger point for a buyer.

One number to carry: Anthropic's own Terminal-Bench 4.0 harness puts Sonnet 5.5 at 70.6% and Artificial Analysis's independent run (its September 30 Argon article) at 64%, a 6.6-point gap that is the size of the vendor-versus-independent difference to expect in your own evaluation; the independent table, with the effort settings each model was run at, is in the benchmark moves.

What to do: rerun the agent cost model with cache-read rates, long-context thresholds and output tokens per task as inputs, not list price; benchmark Sonnet 5.5 against Sol on your own task set rather than theirs, with the real request mix including the requests the gates are designed to catch; and treat Argon as a Q4 evaluation item, not a Q4 deployment option. The capital and power consequences are in AI Stack Weekly; the permission-layer consequences for agents are in Agent Techniques.

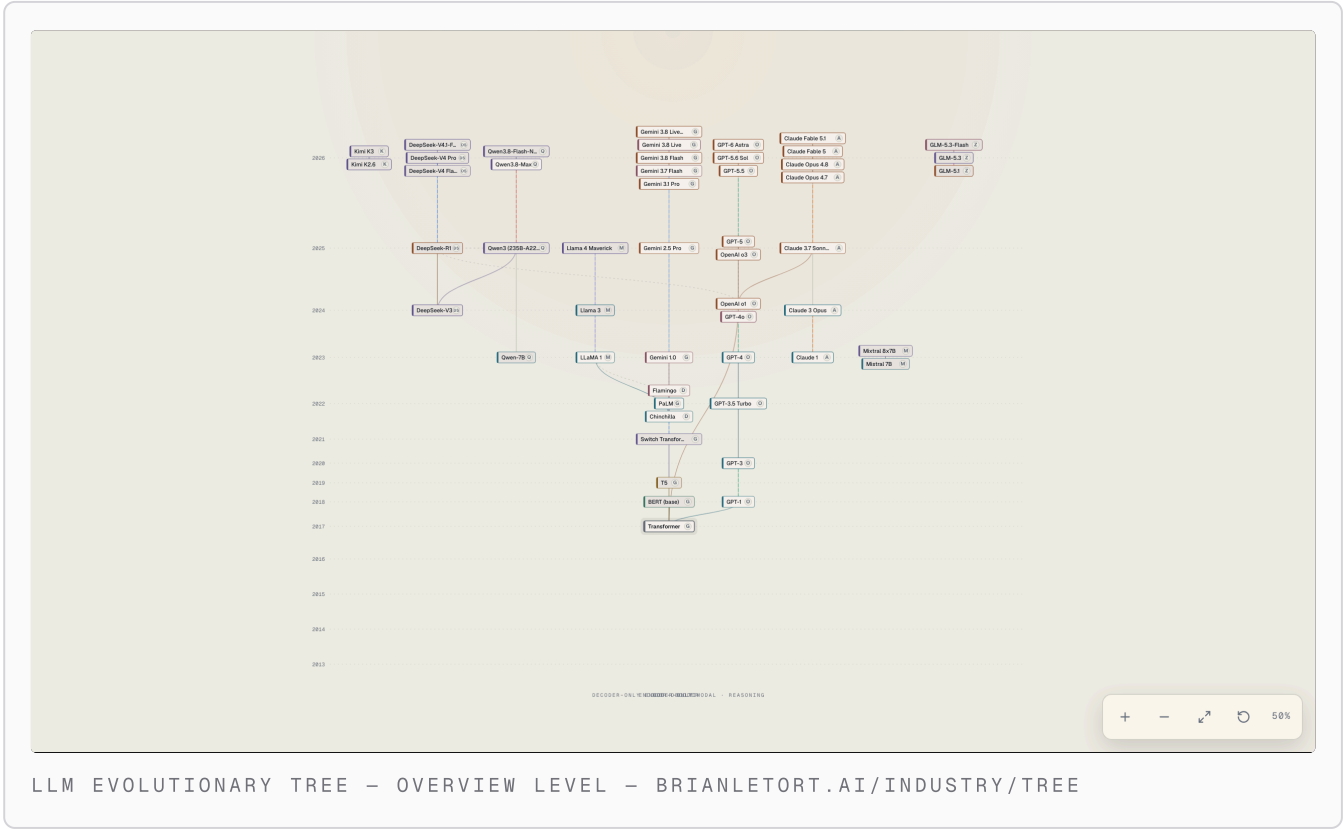
THE BAR

A model release matters when it changes one of three things: which workload runs on-prem, which tier of license is viable for a deployment, or which capability ceiling the closed frontier enforces. Single-vendor releases that change none of these are noise.

TREE DELTA

What changed in the tree.

Six rows added, one updated. Three closed frontier-adjacent models at \$2 / \$10 (Claude Sonnet 5.5, GPT-6.1 Sol, Gemini 4 Argon as gated), two open-weight MoE releases (mixture-of-experts, a model that activates only a few of its sub-networks per token; MiMo-V2.6-Pro now verified, Kolibri-1), and Holo4 under the Holo3.1 precedent (W23 placed H Company's computer-use family as a tree node because it ships in multiple sizes with quantized checkpoints, reduced-precision weights, for local inference). The decision-model fine-tunes are reviewed and deferred rather than placed: the tree's 3-of-6 rule creates a new branch only when at least three of six tests hold (durable architectural distinction, three or more independent vendors, changed infrastructure behaviour, changed user-facing capability, not representable as a tag, and three to five deployed models). This week supplied the vendor count; the architectural-distinction test is the open one, because a classifier head on a shared base may be a fine-tuning recipe rather than a new build.



ADDED (6)

claude-sonnet-5-5

gpt-6-1-sol

gemini-4-argon

mimo-v2-6-pro

holo-4

kolibri-1

UPDATED (1)

grok-4-7

Gemini 4 Argon carries status gated and placement confidence medium because the public cannot run it. MiMo-V2.6-Pro lifts last week's deferral: the Hugging Face repo now carries MIT license metadata and a model card. Grok 4.7's row now carries a 500K context, which the Amazon Bedrock listing prints citing xAI's September 21 launch announcement; last week's row had no context figure because the house had not recorded one.

FRONTIER MOVEMENTS

Flagship-class releases.

4 releases this period. Vendor-stated frontier capability. The releases that reset the closed-source ceiling.

2026-09-28

RELEASE 01

Anthropic

FRONTIER

REASONING

Claude Sonnet 5.5

Second on the independent index at half the price of first, and the first Sonnet with frontier cyber safeguards

Artificial Analysis places Sonnet 5.5 (max) at 56, two points behind Opus 5.5 at \$2 / \$10 against \$4 / \$20. It leads Terminal-Bench 4.0 on both the vendor harness and Artificial Analysis's independent run, 6.6 points apart (the table is in the benchmark moves). The visible fallback to Sonnet 5 on higher-risk cyber requests, driven by classifiers in the request path, is a behaviour your security team should test, because it changes which model answers mid-session; it is also the week's first-party example of a classifier sitting in a production approval path, which Agent Techniques picks up.

SOURCE: ANTHROPIC

2026-09-29

RELEASE 02

OpenAI

FRONTIER

REASONING

GPT-6.1 Sol

Near-Astra on OpenAI's tables at one-fifth the price, with the cheapest cache meter at the tier

Sol has the cheapest cache meter at the tier and the only long-context surcharge (the meter table is in the architecture watch). OpenAI's DeepSWE v1.1 75.2% and \$5.47 per Terminal-Bench Science task are vendor-run, and the cost-per-task advantage over Astra in OpenAI's table follows from Sol's list price being one-fifth of Astra's. Artificial Analysis's independent 52 puts Sol one point below Astra. Sol is classified Critical in cybersecurity (OpenAI's highest internal capability tier, which triggers its strictest deployment safeguards) and ships under Astra's safeguards stack, so expect the same refusal surface as the top tier. It replaced GPT-6 Sol after seven days; pin model ids.

SOURCE: OPENAI

2026-09-30

RELEASE 03

Google DeepMind

FRONTIER

REASONING

Gemini 4 Argon

Google's first above-Flash model in over seven months (Artificial Analysis's line) ties GPT-6 Astra on the independent index and cannot be bought

Artificial Analysis scored Argon (high) at 53, equal to Astra, at about 60% of Astra's cost per task, and attributes the gain to two things: lower hallucinations and stronger agentic results (AutomationBench-AA 78%, first). The hallucination half rests partly on abstention. Per AA's September 30 article, Argon's 15% rate on AA-Omniscience (AA's test of whether a model answers wrongly or declines when it does not know) is the lowest of the three models AA compared (GPT-6 Astra 51%, GPT-6.1 Sol 54%; Sonnet 5.5 was not in that comparison and Astra is not a week's release), but its accuracy is 50% against Astra's 63%, so it declines more rather than knowing more. Google's table leads DeepSWE v1.1 at 77.9% (vendor-run) and trails Opus 5.5 on Terminal-Bench 4.0. The \$2 / \$10 rate is a 50% discount for at least one month, after which the card is \$4 / \$20. Access is limited to Fairwind Program cyber defenders; developers are 'next' with no date; note the 1M-token output limit for long-document generation when it arrives.

SOURCE: GOOGLE

2026-09-28

RELEASE 04

SpaceXAI

FRONTIER

REASONING

Grok 4.7

Lands on Amazon Bedrock with a 500K context and four reasoning-effort levels, the first hyperscaler listing to print the context figure

For AWS-committed enterprises this is a way to run a SpaceXAI model under existing Bedrock IAM, logging and private-link controls. AWS's listing prints the 500K context figure citing xAI's September 21 launch announcement, which the house had not recorded; the price on Bedrock should be checked against the \$2 / \$6 API rate before assuming parity.

SOURCE: AWS MACHINE LEARNING BLOG

OPEN WEIGHTS

Open-frontier and open-source drops.

4 releases this period. Open-weights drops that change procurement options. Pull these into pilot when score parity meets license parity.

2026-09-21

RELEASE 01

Xiaomi

OPEN FRONTIER

MOE

MiMo-V2.6-Pro

Weights verified under MIT; independent trackers confirm it as the top open-weights model

Last week's deferral lifts: the repo carries MIT metadata and a model card for a 1.02T / 42B-active multimodal MoE ('active' is the parameters used per token). Artificial Analysis lists it at 46, ahead of GLM-5.3 (45) and Kimi K3 (44) and tied with Grok 4.7; its cost and minutes per task are in the benchmark moves. For a buyer the number to weigh is the latency: the cheapest open frontier model is also one of the slowest, which decides whether it fits interactive or batch work.

SOURCE: HUGGING FACE MODEL CARD

2026-10-03

RELEASE 02

Aleph Alpha

OPEN FRONTIER

MOE

Kolibri-1

A 78B / 3.46B-active English-German MoE under Apache 2.0 with a 1M context, trained on 768 B200s

The fully open training recipe (20T pretraining tokens in 21 days, about 392k GPU-hours, roughly 6.4e23 FLOPs) is the useful artifact for anyone budgeting a sovereign model; the benchmark table (AIME 2025 96.9%, SWE-Bench Verified 66.4%) is vendor-run with no independent evaluation. At about 78 GB in FP8 (8-bit floating-point weights, half the memory of 16-bit) it fits a single H200 or B200; among European open-weight releases it is the first the house has recorded that combines a single-GPU footprint with a 1M context, though Mistral has shipped Apache 2.0 models with on-prem footprints at shorter contexts.

SOURCE: ALEPH ALPHA

2026-09-28

RELEASE 03

H Company

SPECIALIST

AGENTIC

Holo4

Computer-use family at 27B dense and 35B-A3B with every benchmark trajectory published, under a non-commercial license

H Company's own figures are 45.4% on AutomationBench (business-workflow automation in simulated apps) at \$0.05 per task for the 27B, read from its chart, and 61.7% on OSWorld 2.0 (desktop control) at \$1.22 per task against Opus 5.5's 81.8%. The per-task costs are cheap in absolute terms; no frontier cost per AutomationBench or OSWorld task exists in the sources, so there is no frontier cost comparison to draw, and the vendor tables are from different versions. CC BY-NC 4.0 is more restrictive than the Apache 2.0 Qwen base, so enterprises need a commercial license from H Company before production use. Publishing all trajectories is the right precedent and lets you audit the score.

SOURCE: HUGGING FACE BLOG (H COMPANY)

2026-10-01

RELEASE 04

Cloudflare

SPECIALIST

DENSE

Clef

Open-weight decision model on Workers AI at \$0.24 per million input tokens, with a 9B Clef-flash sibling

Clef returns calibrated probabilities over 1-64 typed questions in one forward pass, which is the shape an agent's allow/flag/block gate needs. Cloudflare's benchmark table (BANKING77 94.20 macro-F1, the average of per-class accuracy scores, against Jev's 79.74) is vendor-run, and the one community head-to-head posted on Hacker News, a 250-sample test, found Clef 5.2x more expensive and roughly five times slower than Jev for a 1.2-point gain. Read it as a reviewer model you can self-host under Apache 2.0, not as a leaderboard result; the technique read and the decision-model price list are in Agent Techniques.

SOURCE: CLOUDFLARE BLOG

ARCHITECTURE WATCH

Patterns to track.

4 architectural patterns that crossed multiple vendors this period. Each pattern names the trend, the exemplar releases, and what it changes for deployment, cost, or capability.

PATTERN 01 List-price convergence with meter divergence

Claude Sonnet 5.5 (\$2 / \$10, cache \$0.20, no long-context tier)

GPT-6.1 Sol (\$2 / \$10, cache \$0.10, \$4 / \$15 above 272K)

Gemini 4 Argon (\$2 / \$10 intro, \$4 / \$20 list, cached 95% off)

Three labs chose the same headline number and set every other meter independently. This is the fact home for the mid-tier meter packet. Cache reads (re-sending context the provider already holds): GPT-6.1 Sol \$0.10 per million, Sonnet 5.5 \$0.20, Argon 95% off list. Long-context repricing: Sol moves to \$4 / \$15 above 272K tokens; Sonnet 5.5 has no surcharge across its 1M context. Introductory pricing: Argon's \$2 / \$10 is a 50% discount for at least one month, reverting to \$4 / \$20, and nobody outside the Fairwind Program can pay either rate yet. As a heuristic rather than a measurement (the house has no token-mix data for production agents), the cache-read rate is the meter most likely to dominate agent sessions, which tend to reread context more than they emit tokens; the long-context threshold catches document and codebase workloads, and only Sol has it; the introductory reversion catches budget planning, and only Argon has it. A router that keys on list price will pick the wrong model for most agent workloads; key it on cache-read rate, context threshold and effective price after the intro period ends, and measure your own token mix to replace the heuristic.

SOURCE: OPENAI API PRICING; ANTHROPIC PRICING; GOOGLE GEMINI 4 ARGON ANNOUNCEMENT

PATTERN 02 Task heads on a shared 27B base

Cloudflare Clef (Qwen3.8-27B)

Perplexity pplx-decider-v1-27b (Qwen3.8-27B)

AutoTrust JEV-27B (Qwen3.8-27B)

Holo4-27B (Qwen3.8-27B)

Four releases in one week post-trained the same Qwen3.8-27B checkpoint into a specialist: three decision heads and one computer-use agent. The pattern is that a 27B dense open base has become the default substrate for task-specific heads, in the way BERT-base once was for classifiers, and that the differentiation is in the head, the loss and the data rather than the trunk. For buyers this means provenance and training-data disclosure matter more than parameter counts, and the Perplexity case, where the published shards match a community checkpoint byte for byte, shows how thin the provenance can be. For the tree it raises the question of whether decision models earn a branch, which the placement rules answer with a 3-of-6 independent-vendor test that this week put in play.

SOURCE: CLOUDFLARE, PERPLEXITY, AUTOTRUST AND H COMPANY MODEL CARDS ON HUGGING FACE

PATTERN 03

Gating as a release stage

Gemini 4 Argon (Fairwind Program only)

Claude Sonnet 5.5 (visible fallback to Sonnet 5 on higher-risk cyber tasks)

GPT-6.1 Sol (Critical in cyber, Astra's safeguards stack)

GPT-6.1 Astra (withheld)

Each of the three labs shipped or announced a model this week with a safety gate that changes what a buyer actually gets. Google's is an access gate: the model exists, scores on independent indices, and is unavailable. Anthropic's is a routing gate: the model you called may hand the request to its predecessor mid-session, visibly, on the decision of reasoning-extraction classifiers in the request path. OpenAI's is a classification gate: Sol inherits Astra's Critical-cyber safeguards, and the planned Astra successor was withheld entirely. The architectural consequence is that 'model' now denotes a policy-wrapped system whose behaviour depends on request content, and evaluation harnesses that assume a fixed model behind an endpoint will produce numbers that do not reproduce in production. Test with your real request mix, including the requests the gate is designed to catch. The same architecture, a classifier deciding what the acting model may do, is the week's technique in Agent Techniques.

SOURCE: GOOGLE GEMINI 4 ARGON ANNOUNCEMENT; CLAUDE SONNET 5.5 SYSTEM CARD; OPENAI DEPLOYMENT SAFETY HUB

PATTERN 04

Verbosity and latency as hidden costs of open frontier scores

MiMo-V2.6-Pro (46 on the index, 19.5 minutes per task, \$0.13 per task)

GLM-5.3 (45 on the index, \$2.01 per task)

Kolibri-1 (3.46B active, vendor-run only)

The top open-weights model on the independent index is also one of the slowest: Artificial Analysis clocks MiMo-V2.6-Pro at 19.5 minutes per Intelligence Index task at \$0.13, roughly a fifteenth of GLM-5.3's \$2.01 per task for one more index point. AA's cost per task is price times tokens emitted, so a model that scores by generating long reasoning traces can be cheap per token and slow per task at the same time, and the comparison that matters is cost and minutes per task, not list price. Kolibri-1's 3.46B active on 78B total is the extreme case of a small active footprint, with no independent latency or tokens-per-task figure yet. An architect choosing an open model for interactive work should weight tokens per second and tokens per task as heavily as the index score; for batch work the index-per-dollar figure is the right one.

SOURCE: ARTIFICIAL ANALYSIS MODEL PAGES

BENCHMARK MOVES

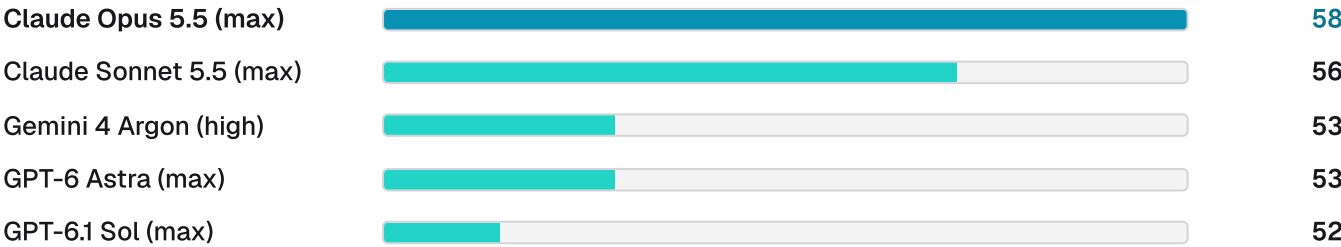
Where the leaderboard moved.

6 benchmark deltas that change a procurement read. Scores reflect public leaderboards or vendor model cards as of publication.

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX V4.3.2

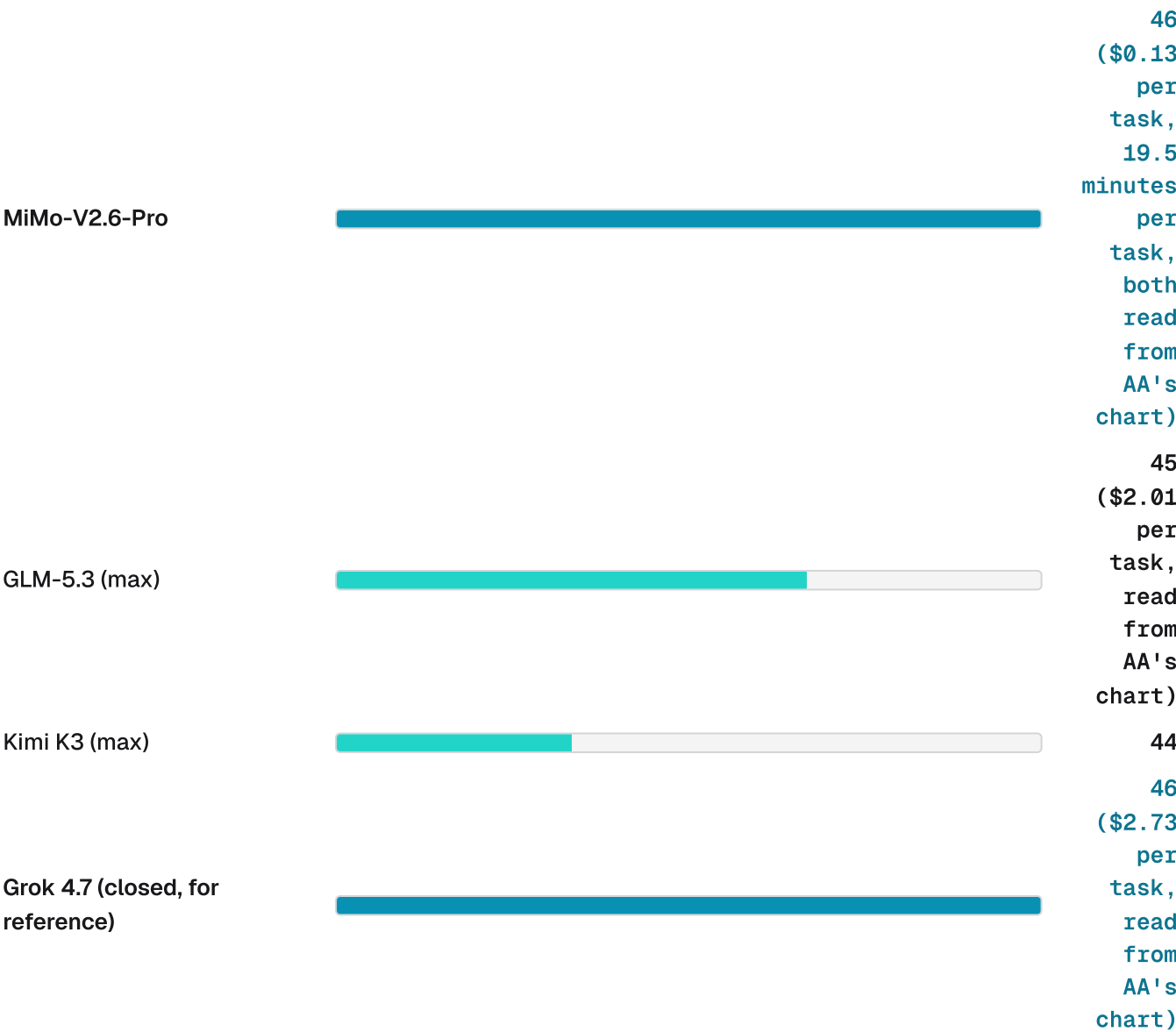
MOVE 01

Four new entries in one week; Opus 5.5 holds first, Sonnet 5.5 enters second, Argon ties Astra, MiMo-V2.6-Pro confirmed as top open weights. The setting in parentheses is the reasoning-effort level the model was run at; scores are not comparable across settings



SOURCE: ARTIFICIAL ANALYSIS

MiMo-V2.6-Pro confirmed first among open-weights models, tying Grok 4.7 at a fraction of the cost per task; this table is the fact home for the open-weights cost and latency figures

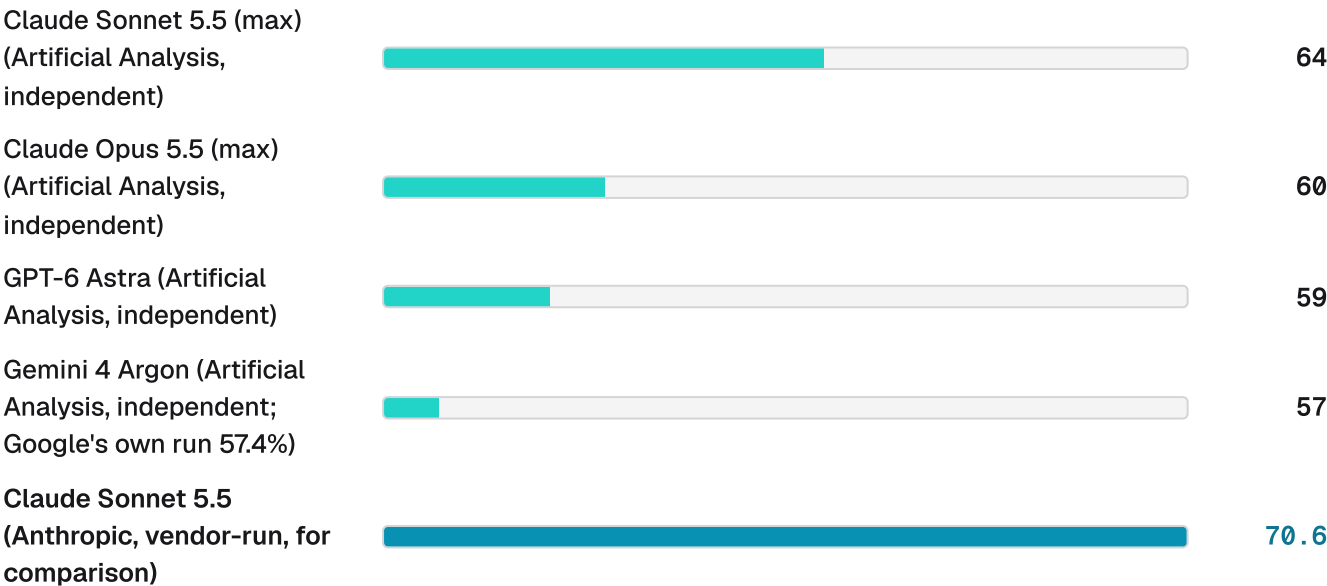


SOURCE: ARTIFICIAL ANALYSIS

TERMINAL-BENCH 4.0 (INDEPENDENT ARTIFICIAL ANALYSIS RUNS, WITH THE VENDOR FIGURE FOR COMPARISON)

MOVE 03

Artificial Analysis's independent run puts Sonnet 5.5 first at 64%, 6.6 points below Anthropic's own 70.6%; Opus 5.5, Astra and Argon follow within seven points. The AA figures are read from AA's September 30 Argon article and sit outside the week's graded research set; the vendor figures are graded



SOURCE: ANTHROPIC; GOOGLE DEEPMIND; ARTIFICIAL ANALYSIS

DEEPSWE V1.1 (VENDOR-RUN TABLES)

MOVE 04

Argon claims the top score; Sol matches Astra at one-fifth the price; every figure is from the vendor's own table

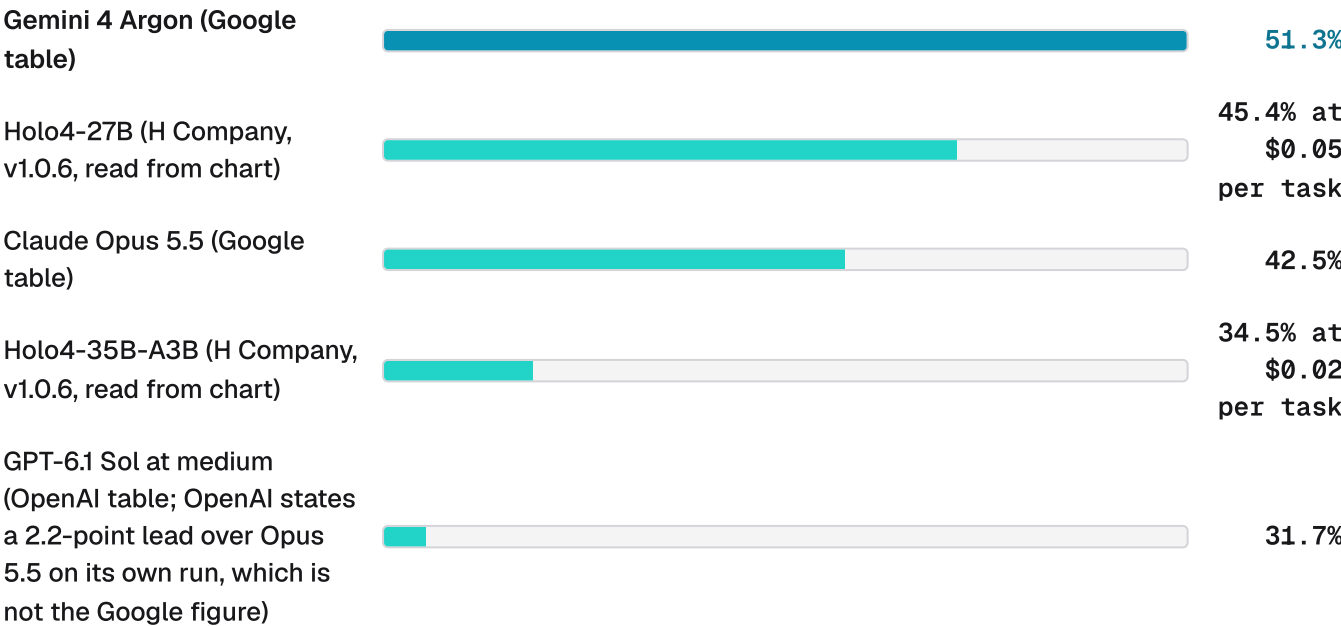


SOURCE: OPENAI; GOOGLE DEEPMIND

AUTOMATIONBENCH (VENDOR-RUN, MIXED VERSIONS)

MOVE 05

Argon leads on Google's table, 8.8 points ahead of Opus 5.5 on the same table; Holo4-27B reaches 45.4% at \$0.05 per task on H Company's chart; Sol's 31.7% on OpenAI's table is a different version and is not comparable with Google's rows

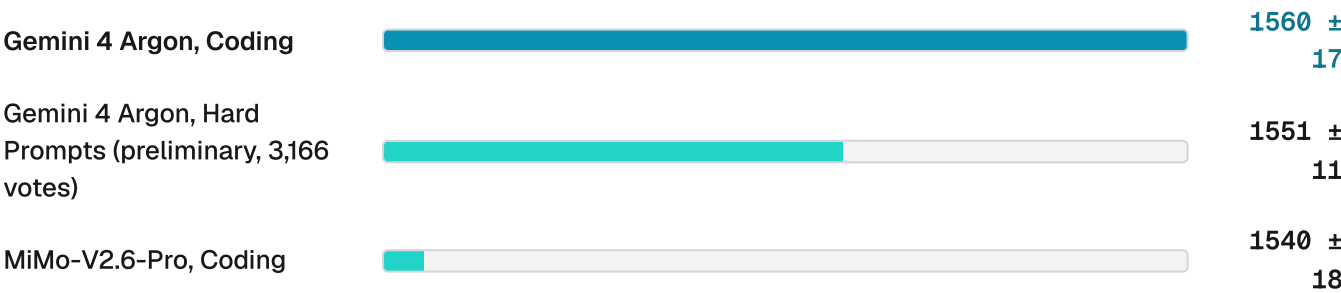


SOURCE: GOOGLE DEEPMIND; H COMPANY; OPENAI

LMARENA TEXT LEADERBOARDS (PUBLISHED SEP 30 AND OCT 2)

MOVE 06

Argon debuts first on Hard Prompts (preliminary) and Coding; MiMo-V2.6-Pro enters Coding at 1540



SOURCE: LMARENA

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Oct 3, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	Claude Opus 5.5	GPT-6 Astra (tied on the index by Gemini 4 Argon, which cannot be bought)	Opus 5.5 leads the independent index at 58, five points clear of Astra and Argon at 53. Fable 5.1 held the slot through W39 on buyer-trace grounds (the house kept it as the default until a buyer-side trace showed Opus 5.5 matching it on real work); no such trace has arrived, and the house has no independent index score for Fable 5.1 in this window, so the handover rests on Opus 5.5's index lead and on Anthropic positioning Opus as its flagship, not on a measured Fable gap. Argon's tie with Astra is real and unpurchasable.
Open frontier	MiMo-V2.6-Pro	GLM-5.3	Verified weights under MIT and an independent 46 make MiMo the open leader; GLM-5.3 at 45 is the challenger. DeepSeek-V4.1-Flash gives up the slot it held on serving cost; Kolibri-1 enters the watch column pending any independent score.
Reasoning	Claude Opus 5.5	Claude Sonnet 5.5	Two points separate them on the independent index at a 2x price difference. Sonnet 5.5 is the default for most reasoning workloads this quarter; Opus 5.5 is the ceiling. GPT-6.1 Sol at 52 is the cross-vendor alternative with the cheaper cache meter.
Coding	Claude Sonnet 5.5	GPT-6.1 Sol	Sonnet 5.5 leads Terminal-Bench 4.0 on both the vendor and the independent run (table above) and takes the slot from Opus 5.5 at half the price; Sol replaces GPT-6 Astra as challenger because it matches Astra on OpenAI's DeepSWE table at one-fifth the cost. Argon's 77.9% DeepSWE is the highest claim and the least testable.

TIER	LEADER	CHALLENGER	READ
Multimodal	Gemini 3.8 Live Extended Thinking	Gemini 4 Argon	The buyable Google multimodal model holds the slot; Argon's 1M-token output and LVBench 91.7% (vendor-run) would take it when developers can call it. MiMo-V2.6-Pro's native text, image, video and audio input makes it the open alternative.
Edge / small	4B decision heads via llama.cpp (Kev-4B and peers)	Clef-flash (9B)	This tier covers models meant to run on one device or one CPU, judged on a shipped task they do well rather than on the index. The day-0 decision heads, from 144 million parameters to 4B, answering allow/flag/block in 3-43 ms per question (author-reported; 3 ms is Julia-1 at 144 million parameters, and the 4B heads are 12-36 ms) through llama.cpp's typed endpoint take the slot from last week's leader, the vendor claim that a 30B mixture-of-experts model runs locally on the Snapdragon 8 Elite Extreme Gen 6 handset, which still has no independent throughput figure; Clef-flash at 9B under Apache 2.0 replaces Nex-N2.5 mini as challenger. Holo4-27B at \$0.05 per AutomationBench task (vendor-run) is noted, but a 27B model served through an API is not an edge footprint.

VENDOR SIGNALS

Pricing, gating, deprecation.

6 non-release moves that shift vendor risk — pricing, deprecations, gating decisions, license changes — each with a one-line procurement read.

2026-09-28

SIGNAL 01

OpenAI

Shelves GPT-6.1 Astra after internal alignment tests; the pause on an unreleased most-capable tier remains; publishes a 'towards safety cases' post with no numbers

GPT-6 Astra remains in production behind dots and the Ultrafast tier; what OpenAI withheld is its successor, and what it paused is a tier above Astra that never shipped. Architects should keep the cross-vendor fallback for top-tier tool use through Q4 because the roadmap above Astra is now empty. The post's absence of quantified results means the external evidence for the decision is the UK AI Security Institute's evaluation of GPT-6 Astra, which is about the model that stayed; the three-model table and its caveats are in Agent Techniques.

SOURCE: OPENAI

2026-09-29

SIGNAL 02

OpenAI

Adds an Ultrafast service tier for GPT-6 Astra at \$60 / \$300 per million (6x) for up to 8x token speed in Codex and 6x in the API, per OpenAI; introduces Pro 500 at \$500 a month; halves the Pro 200 Codex allowance from October 30

This is the first retail price increase on a frontier coding tier since the mid-tier convergence, and it prices speed as a premium product rather than a model property. Teams that depend on Codex throughput should model the move to Pro 500 or Sol Ultrafast against moving the workload to Sonnet 5.5 before October 30.

SOURCE: OPENAI API PRICING

2026-09-30

SIGNAL 03

Google DeepMind

Announces Gemini 4 Argon with access limited to Fairwind Program cyber defenders; publishes introductory pricing with no end date and no developer availability date

Google chose to claim the frontier on independent indices before letting anyone buy the model. For a buyer the signal is that Google's Q4 roadmap is public and its Q4 product is not. The reversion to \$4 / \$20 after the introductory month means the \$2 / \$10 figure in comparison pieces is a promotional rate, not Google's price.

SOURCE: GOOGLE

2026-09-28

SIGNAL 04

Anthropic

Ships Sonnet 5.5 to all platforms including free Claude.ai; sets retirement no sooner than September 28, 2027; Haiku 5.5 slips to 'coming weeks' for a second week

A one-year retirement floor is a deployment guarantee most labs do not give and should go into vendor-selection scorecards. The Haiku slip matters for the cheap tier: there is still no Anthropic model below \$2 / \$10 in the 5.5 family, which leaves the sub-dollar tier to GPT-6 Luna, Gemini Flash and the open models.

SOURCE: CLAUDE PLATFORM DOCS

2026-10-01

SIGNAL 05

Perplexity

Launches a Decisions API at \$0.04 per million input tokens with free output, backed by pplx-decider-v1-27b; community analysis finds the weights byte-identical to a 12-day-old community checkpoint that the launch post does not mention

The price is the lowest in the decision-model category and the provenance is the weakest. Enterprises evaluating per-action review should treat the vendor benchmark table as describing someone else's model until Perplexity explains the match in its launch materials, and should prefer a vendor whose training provenance is documented.

SOURCE: HUGGING FACE (PERPLEXITY-AI/PPLX-DECIDER-V1-27B);
MODELSYSTEM.ONE**2026-10-01**

SIGNAL 06

Black Forest Labs

Launches FLUX 3 Image via API and a commercial-weights license, with open weights promised 'in the coming weeks'

The open-weights promise is a calendar item, not a release; FLUX 3 is tracked here only because its weights, when published, would be the first open image model of the generation. Until then it is an API product.

SOURCE: BLACK FOREST LABS

WATCHLIST

On the radar next.

6 model-side catalysts in the next 7–30 days that would change the read materially. Watching these tells us whether the canopy is widening or thinning.

OCTOBER

WATCH 01

Claude Haiku 5.5 model id on the Claude Platform models page

Two weeks of 'coming weeks'. Haiku 5.5 would be Anthropic's first sub-\$2 model with the 5.5 safeguards and the obvious reviewer model for per-action gating.

Q4 2026

WATCH 02

Gemini 4 Argon developer availability and the introductory pricing end date

Independent benchmarks on a model nobody can run are a preview, not a result. Developer access resets the top-tier comparison and the \$4 / \$20 reversion tests the mid-tier convergence.

OCTOBER

WATCH 03

A second independent Terminal-Bench 4.0 harness on Claude Sonnet 5.5

Anthropic's 70.6% and Artificial Analysis's 64% are 6.6 points apart on the number most likely to move coding-agent defaults this quarter. A second independent harness would show whether the gap is AA's setup or the vendor's.

OCTOBER

WATCH 04

Perplexity response on pplx-decider-v1-27b provenance

Byte-identical shards to a community checkpoint, unmentioned in the launch post, is either an oversight or a disclosure problem. Which one it is decides whether the Decisions API belongs in an enterprise evaluation.

BY OCT 30

WATCH 05

Nvidia InferenceX submission for Vera Rubin NVL72

The lapsed Q3 commitment and its consequences are tracked in AI Stack Weekly; a Rubin submission would give the first independent cost-per-million figures for serving this week's models.

Q4 2026

WATCH 06

First independent evaluation of Kolibri-1 and FLUX 3 open weights

Kolibri-1 pairs a single-GPU footprint with a 1M context, a combination no other European open-weight release in the house record has; its table is vendor-run. FLUX 3's weights decide whether the open image tier moves this generation.

METHODOLOGY AND AUTHORSHIP

How this brief is built.

Compiled from public model cards, vendor blogs, leaderboards, and official lab announcements. The publication is anchored to the LLM Evolutionary Tree at brianletort.ai/industry/tree, which the brief annotates each period.

WHAT EACH SECTION IS FOR

TREE DELTA	Model rows added or updated in content/llm-tree/models.yaml since the prior issue. Every id is real and clickable on the web view.
FRONTIER AND OPEN WEIGHTS	Releases that reset the closed ceiling or move the open-frontier line. Each card cites a single primary source.
ARCHITECTURE WATCH	Patterns that crossed multiple vendors in the period. Named once, exemplified by recent releases.
BENCHMARK MOVES	Public leaderboards and vendor model cards. Bars reflect the score range across the rows shown, not zero-baselined.
TIER SCORECARD	Leader vs challenger by tier. A snapshot for procurement shortlists; refreshed every issue.

AUTHORSHIP

Brian Letort

BrianLetort.AI · Independent analysis. All sources public. Not investment guidance.

THIS ISSUE

Issue 24

Week 40 of 2026 · October 3, 2026

WEB

brianletort.ai/industry/model

The Model Pulse archive, the LLM Evolutionary Tree, and the AI Stack Weekly companion publication.