THE MODEL PULSE ISSUE 14 · WEEK 30 OF 2026

THE BIG READ

Opus-class intelligence just moved into the mid-tier price band, while Google reset the fleet-economics floor

KEY SIGNALS THIS PERIOD

2	3	3	4
Models added to the tree	Vendors shipped frontier-class	Architectural patterns crossed multiple vendors	Benchmark moves

AUTHOR

Brian Letort
BrianLetort.AI

PUBLISHED

July 25, 2026
Issue 14 · 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.

Anthropic's Jul 24 Claude Opus 5 release is not a conventional flagship upgrade. It delivers vendor-reported state-of-the-art coding and knowledge-work results at $5/$25 per million tokens, the same list price as Opus 4.8 and roughly half the price of Claude Fable 5. That compresses the gap between the premium and mid-tier on both capability and price: buyers no longer need to pay the Fable premium by default for hard knowledge work. The model adds adaptive thinking, a beta that lets applications change tool definitions during a conversation without invalidating the prompt cache, and beta automatic API fallbacks when a safety classifier blocks a request. Fast mode is roughly 2.5x faster at 2x the token price, so latency is now an explicit purchasable tier rather than a separate model.

Google attacked the same market from below on Jul 21. Gemini 3.6 Flash holds the $1.50 input price of 3.5 Flash, cuts output to $7.50, and Google says it uses about 17% fewer output tokens on comparable tasks. Gemini 3.5 Flash-Lite lands at $0.30/$2.50 and roughly 350 output tokens per second. The important number is effective completed-task cost, not the rate card: fewer generated tokens compound across every reasoning and tool loop in a fleet. Gemini 3.5 Flash Cyber is a different category — a security-specialist model limited to governments and select partners through CodeMender — and should not be mistaken for a generally available procurement option. Gemini 3.5 Pro still has not shipped, while Google says Gemini 4 pretraining has started; the roadmap is moving before the delayed flagship closes its launch.

DeepSeek supplied the migration lesson. The legacy deepseek-chat and deepseek-reasoner aliases were retired Jul 24 at 15:59 UTC with no redirect, forcing applications onto deepseek-v4-pro or deepseek-v4-flash and making thinking a parameter rather than a separate endpoint. That resolves only part of last week's prediction: the alias retirement occurred, but the clean GA and pricing evidence required for a full hit remains incomplete. Kimi K3 is the opposite watch item — weights and license remain unpublished as of this issue, with Moonshot's Jul 27 promise still ahead. Net/net: model procurement is becoming a routing

problem. Use Opus 5 when the work needs frontier judgment, Flash when fleet economics dominate, and treat availability, safety fallback behavior, token efficiency, and cache continuity as part of the model specification.

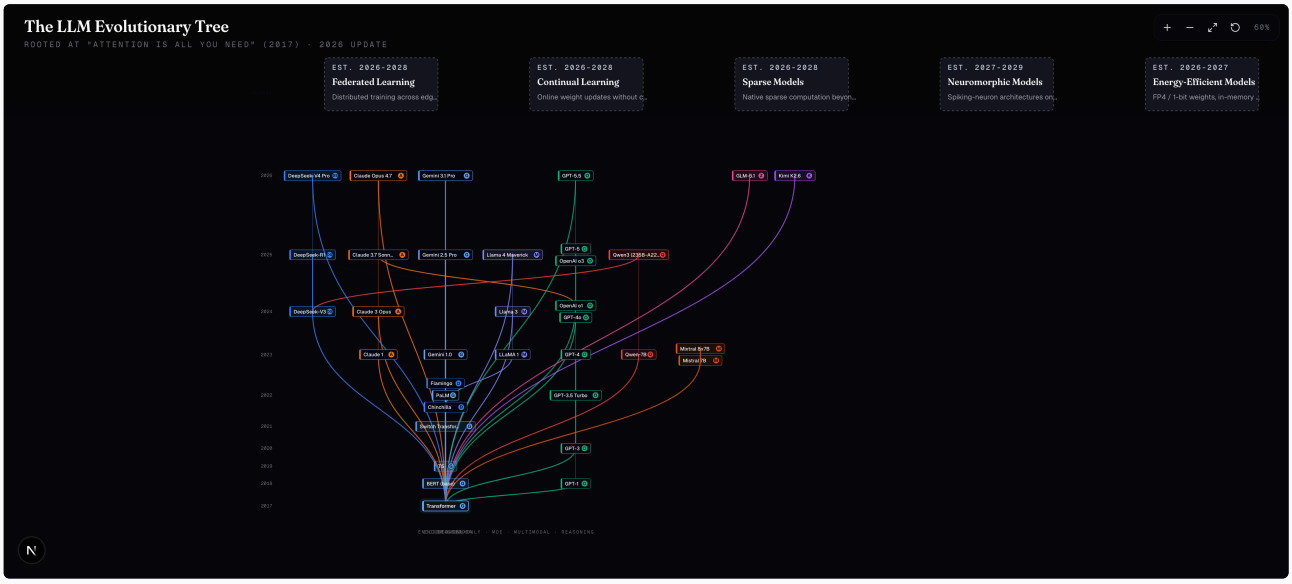
THE BAR

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.

Two closed-model rows added: Claude Opus 5 compresses near-frontier intelligence into the $5/$25 band, and Gemini 3.6 Flash resets high-volume agent economics through lower output pricing and lower token use.



LLM EVOLUTIONARY TREE — OVERVIEW LEVEL — BRIANLETORT.AI/INDUSTRY/TREE

ADDED (2)

- `claude-opus-5`
- `gemini-3-6-flash`

UPDATED (0)

No updates to existing rows this period.

Gemini 3.5 Flash-Lite is covered as a fleet-economics release but not added to the tree this week; the tree delta stays focused on the two models that materially moved the frontier or its price-performance boundary. Kimi K3 remains recorded as closed until weights and license text actually ship.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-07-24

RELEASE 01

Anthropic

FRONTIER

REASONING

Claude Opus 5

Near-Fable intelligence at $5/$25 per million tokens, with adaptive thinking, mutable tools, automatic safety fallbacks, and a paid fast lane

Default hard knowledge and coding workloads to Opus 5 before paying the Fable 5 premium: Anthropic reports state-of-the-art Frontier-Bench and GDPval-AA results while holding Opus 4.8 pricing. Keep Mythos 5 for cyber-sensitive frontier work, where Opus 5 still trails, and evaluate the two beta controls separately — tool mutation changes cache economics, while automatic fallbacks change which model may process a request.

SOURCE: ANTHROPIC

2026-07-21

RELEASE 02

Google DeepMind

FRONTIER

MULTIMODAL

Gemini 3.6 Flash

$1.50/$7.50 per million tokens with about 17% fewer output tokens than 3.5 Flash, aimed directly at high-volume agent fleets

Re-benchmark completed-task cost rather than comparing rate cards: Google's claimed 17% output-token reduction compounds over every tool loop and may matter more than the $1.50 output-price cut. The model is the practical fleet tier while 3.5 Pro remains delayed; buyers should require their own task-level token and latency traces before accepting the vendor efficiency claim.

SOURCE: GOOGLE

2026-07-21

RELEASE 03

Google DeepMind

SPECIALIST

AGENTIC

Gemini 3.5 Flash Cyber

A security-specialist Flash variant restricted to governments and select partners through CodeMender rather than a public API SKU

Treat the release as evidence of capability gating, not as a model available for ordinary security procurement. Security teams should ask whether benchmark or hardening benefits will flow into generally available Gemini models; until then, any comparison against public cyber models is structurally unequal because access is policy-limited.

SOURCE: GOOGLE

OPEN WEIGHTS

Open-frontier and open-source drops.

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

2026-07-16

RELEASE 01

Moonshot AI

FRONTIER MOE

Kimi K3

Weights and commercial license still pending ahead of Moonshot's Jul 27 commitment; hosted capability remains strong but openness remains promissory

Do not build a self-host plan around a dated promise. K3 remains the strongest open-weights candidate on last week's independent scores, but the procurement fact is unchanged as of Jul 25: no downloadable weights and no license text. Keep the evaluation environment ready, then review the license, serving footprint, and independent replications before moving it into the open-frontier column.

SOURCE: MOONSHOT AI

ARCHITECTURE WATCH

Patterns to track.

3 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

Runtime policy becomes part of the model contract

`Claude Opus 5 automatic safety-classifier fallbacks`

`DeepSeek thinking as a parameter rather than a separate endpoint`

`Gemini Flash Cyber access restricted by customer class`

A model ID no longer fully specifies runtime behavior. Providers are adding fallback selection, reasoning effort, access policy, and latency tiering around the weights, which means the same request can execute on a different model or policy path than the caller expected. Regulated buyers should log the effective model, fallback reason, reasoning mode, and tool schema version on every run rather than treating the requested model name as sufficient provenance.

SOURCE: ANTHROPIC; DEEPSEEK API DOCUMENTATION; GOOGLE

PATTERN 02

Token efficiency becomes a first-class benchmark

`Gemini 3.6 Flash: ~17% fewer output tokens than 3.5 Flash`

`Claude Opus 5 adaptive thinking`

`Claude Opus 5 fast mode at 2x price and ~2.5x speed`

The rate card is losing explanatory power because models now vary reasoning effort, token verbosity, and latency price independently. A 17% output-token reduction can erase more cost than a modest list-price cut across long agent loops, while a 2x fast-mode premium may still win where human wait time dominates. Evaluation harnesses should report completed-task cost, output tokens, wall-clock time, retries, and cache hits together.

SOURCE: GOOGLE; ANTHROPIC

PATTERN 03

Tool schemas become mutable conversation state

`Claude Opus 5 mid-conversation tool changes`

`Prompt-cache preservation across tool updates`

`Agent platforms adding tools after authentication or escalation`

Opus 5's beta tool mutation breaks the old assumption that an agent's complete tool surface must be known at session start. Platforms can expose privileged or task-specific tools only when needed without throwing away the cached prefix, reducing both permission exposure and repeated input cost. The control requirement rises with the efficiency gain: every tool-set change needs an auditable reason, policy check, and versioned schema.

SOURCE: ANTHROPIC

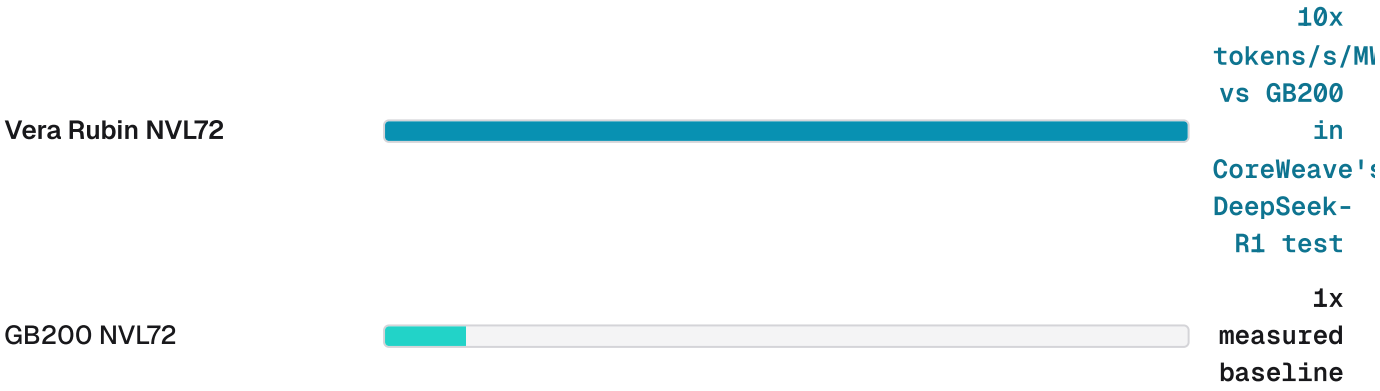
BENCHMARK MOVES

Where the leaderboard moved.

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

RACK-LEVEL INFERENCE EFFICIENCY MOVE 01

CoreWeave measured NVIDIA Vera Rubin NVL72 at 10x the DeepSeek-R1 tokens per second per megawatt of GB200



SOURCE: COREWEAVE

FRONTIER-BENCH AND GDPVAL-AA MOVE 02

Anthropic reports Opus 5 at state of the art on coding and knowledge work, placing a $5/$25 model near the Fable 5 capability tier

Claude Opus 5	Vendor-reported SOTA on Frontier-Bench / GDPval-AA
Claude Fable 5	Premium reference; roughly 2x Opus 5 list price
Claude Opus 4.8	Prior Opus baseline at the same $5/$25 price

SOURCE: ANTHROPIC

CYBER CAPABILITY

MOVE 03

Opus 5 improves the mainstream frontier but remains behind Claude Mythos 5 on cyber, preserving a meaningful specialist gap

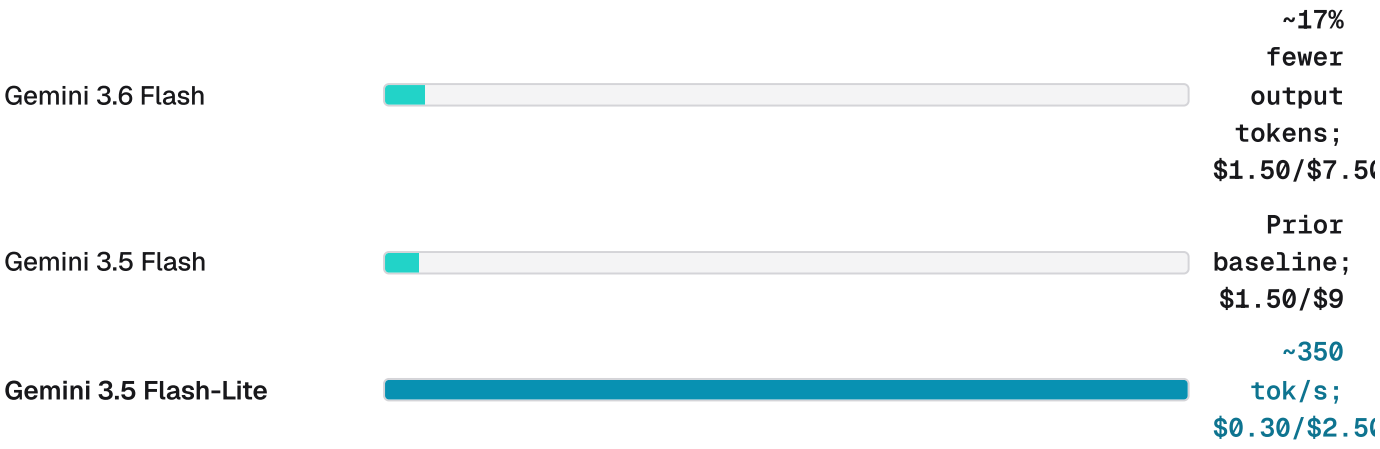
Claude Mythos 5	Anthropic's stronger cyber-capability reference
Claude Opus 5	Behind Mythos 5 on cyber; broadly available flagship
Gemini 3.5 Flash Cyber	Restricted government/partner access via CodeMender

SOURCE: ANTHROPIC; GOOGLE

FLEET EFFICIENCY

MOVE 04

Google claims Gemini 3.6 Flash cuts output-token use about 17% versus 3.5 Flash while lowering output price from $9 to $7.50



SOURCE: GOOGLE

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Jul 25, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	Claude Fable 5	Claude Opus 5	Fable retains the absolute reference, but Opus 5 is now the economic default for most frontier work at roughly half the price.
Open frontier	GLM-5.2	Kimi K3	K3 remains hosted-only as of publication; it cannot take the open-frontier lead until weights and a commercially usable license ship.
Reasoning	Claude Fable 5	Claude Opus 5	Adaptive thinking and flat Opus pricing narrow the operational gap; independent cost-per-task testing is the next checkpoint.
Coding	Claude Opus 5	GPT-5.6 Sol	Opus 5 takes the practical coding lead on Anthropic's launch evidence; independent harness comparisons still need to confirm it.
Multimodal	Gemini 3.6 Flash	Kimi K3	Gemini combines public availability, multimodal breadth, and fleet economics; K3's deployment-control advantage remains pending.
Edge / small	Gemini 3.5 Flash-Lite	Mellum2	Flash-Lite resets the hosted speed-cost tier at roughly 350 tok/s, while Mellum2 remains the local-control alternative.

VENDOR SIGNALS

Pricing, gating, deprecation.

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

2026-07-21
SIGNAL 01
CoreWeave

CoreWeave published a measured Vera Rubin NVL72 result of 10x more DeepSeek-R1 tokens per second per megawatt than GB200

This is stronger than a vendor projection because it is a workload-level operator measurement, but it remains one workload and one stack. Capacity planners should request reproducible methodology and their own workload results before carrying 10x into fleet economics.

SOURCE: COREWEAVE

2026-07-24
SIGNAL 02
Anthropic

Opus 5 holds Opus 4.8 pricing at $5/$25 while fast mode sells roughly 2.5x speed for a 2x token-price premium

Anthropic is segmenting by latency and workflow value rather than reserving higher intelligence for a higher base rate. Buyers should route only latency-sensitive work into fast mode and keep ordinary asynchronous agents on the standard tier.

SOURCE: ANTHROPIC

2026-07-21
SIGNAL 03
Google

Gemini 3.6 Flash and 3.5 Flash-Lite cut the fleet cost curve while Gemini 3.5 Pro remains delayed and Gemini 4 pretraining starts

Google is shipping around its delayed flagship. Procurement teams should evaluate the models available now rather than buying a roadmap story around 3.5 Pro or Gemini 4.

SOURCE: GOOGLE

2026-07-24
SIGNAL 04
DeepSeek

deepseek-chat and deepseek-reasoner aliases retired at 15:59 UTC with hard failure and no redirect; V4 Pro/Flash use a thinking parameter

Model aliases are not stable infrastructure. Pin explicit model IDs, monitor provider deprecation feeds, and test hard-failure behavior before the next forced migration rather than assuming an alias will redirect safely.

SOURCE: DEEPSEEK API DOCUMENTATION

2026-07-25

SIGNAL 05

Moonshot AI

Kimi K3 weights and license remain unpublished two days before the vendor's Jul 27 commitment

The open-weights claim remains a roadmap item. Hold the self-hosting conclusion until downloadable artifacts and explicit commercial terms are public.

SOURCE: MOONSHOT AI

WATCHLIST

On the radar next.

4 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.

JUL 27 - AUG 10

WATCH 01

Kimi K3 weights and commercial license

This decides whether K3 becomes a real self-hosted frontier option or remains a strong hosted model wrapped in open-language marketing.

BY JUL 31

WATCH 02

Gemini 3.5 Pro GA

The delayed flagship must ship a callable model ID and public pricing to resolve the standing prediction; another slip changes the roadmap credibility read.

BY AUG 15

WATCH 03

Independent Opus 5 cost-per-task results

Vendor benchmark leadership is not enough; the market needs task cost, token use, and latency against Fable 5 and GPT-5.6 Sol.

NEXT 30 DAYS

WATCH 04

Automatic fallback audit semantics

Watch whether Anthropic exposes the effective fallback model and classifier reason in logs; without both, the beta creates a provenance gap.

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 14

Week 30 of 2026 · July 25, 2026

WEB

brianletort.ai/industry/model

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