THE MODEL PULSE ISSUE 20 · WEEK 36 OF 2026

THE BIG READ

Model governance now starts at the deployable package: entitlement, retention, runtime version, and price window

KEY SIGNALS THIS PERIOD

3	5	3	3
Models added to the tree	Vendors shipped frontier-class	Architectural patterns crossed multiple vendors	Benchmark moves

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

September 5, 2026

Issue 20 · 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.

The differentiated procurement move this week is not the already-circulating claim that the harness matters. It is that a model stock-keeping unit (SKU) now includes administrator enablement, retention route, provider-adapter version, reasoning effort, and promotion expiry. Claude Fable 5.1 is broadly available in GitHub Copilot but administrator-enabled for enterprises and follows a documented default-retention route outside approved zero-data-retention arrangements. Gemini 3.8 Flash enters the same surface under introductory provider pricing, while Astra enterprise access is off by default.

ARC Prize supplies the evidence for why runtime version belongs in that record: its controlled maximum-reasoning comparison produced a 35.9-point harness-associated spread. The best-observed adapter spread is larger but changes reasoning effort, so it is descriptive rather than causal. This launch-week finding belongs to ARC Prize and prior coverage; the operational extension here is to version the whole deployable SKU and rerun it without provider-private state before claiming portability.

Open weights were quiet after W35's five-model wave. NVIDIA's Apache-2.0 Muse-Glimmer-30B low-precision checkpoint reduced storage substantially, but it is a quantization update to an existing model, not evidence that the open frontier closed the gap this week. The model tree adds Astra, Fable 5.1, and Gemini 3.8 Flash without manufacturing a new open-frontier leader from a footprint optimization.

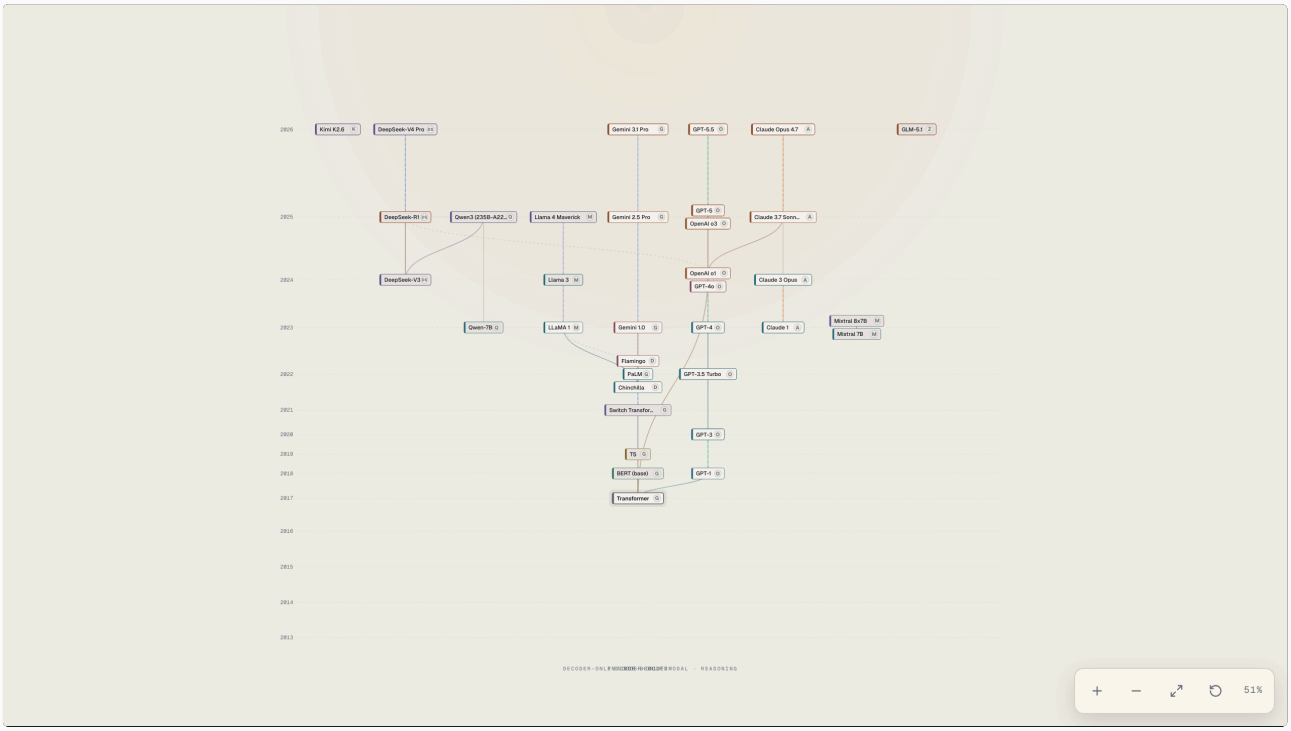
THE BAR

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.

3 model rows added: GPT-6 Astra, Claude Fable 5.1, and Gemini 3.8 Flash. No existing model row changed; Muse-Glimmer's low-precision checkpoint is an update, not a new lineage node.



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

ADDED (3)

- gpt-6-astra
- claude-fable-5-1
- gemini-3-8-flash

UPDATED (0)

No updates to existing rows this period.

W36's delta is closed and distribution-heavy. The durable change is that runtime state, administrator entitlement, retention, and promotional pricing now sit beside architecture in model diligence.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-09-03
RELEASE 01
OpenAI
FRONTIER
AGENTIC

GPT-6 Astra

1M-context closed agent model with a 35.9-point matched-effort harness-associated ARC-AGI-3 spread

Evaluate Astra as a model-plus-runtime release. ARC Prize's matched-effort comparison associates a 35.9-point spread with the harness; portability reviews should reproduce it without provider-private reasoning state.

SOURCE: OPENAI, ARC PRIZE

2026-09-01
RELEASE 02
Anthropic
FRONTIER
REASONING

Claude Fable 5.1 in GitHub Copilot

General availability across Copilot surfaces with administrator enablement and retention conditions

Engineering leaders get a broadly distributed coding model, not an automatically enabled enterprise SKU. Put administrator entitlement and the documented 30-day default retention route into the architecture record alongside model version and benchmark results.

SOURCE: GITHUB

2026-09-03
RELEASE 03
Google
FRONTIER
REASONING

Gemini 3.8 Flash in GitHub Copilot

Flash-tier agent model enters Copilot under introductory provider pricing through December 31

Treat current economics as a promotional observation, not a 2027 run rate. GitHub describes recovery from actionable terminal failures but publishes no reproducible benchmark or exact multiplier, so buyers should measure completed-task cost on their own harness.

SOURCE: GITHUB

OPEN WEIGHTS

Open-frontier and open-source drops.

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

2026-09-01

RELEASE 01

NVIDIA

OPEN FRONTIER

MULTIMODAL

Muse-Glimmer-30B NVFP4

Apache-2.0 Blackwell checkpoint uses NVIDIA's 4-bit floating-point format (NVFP4) to cut storage while preserving multimodal tools

This is a deployment-footprint update to an existing W33 model, not a new frontier row. Blackwell operators should test it for lower memory pressure, but NVIDIA's small Terminal-Bench gain over the 16-bit floating-point (BF16) baseline may be sampling noise and should not drive a capability claim.

SOURCE: NVIDIA ON HUGGING FACE

2026-09-03

RELEASE 02

H Company

SPECIALIST

MULTIMODAL

NeoMME-800M

Apache-2.0 single-tower multimodal encoder for shared text-token and raw-image-patch retrieval

Relevant for multimodal retrieval and agent memory, not generative-model substitution. Day-zero Transformers support makes it testable now, but no production latency, retrieval cost, or independent benchmark is public, so keep it out of production scorecards.

SOURCE: HUGGING FACE, NEOMME PAPER

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 version becomes a governed SKU property

`GPT-6 Astra Standard harness` `GPT-6 Astra Provider Adapter`

ARC Prize's 35.9-point same-effort spread is too large to leave the runtime as an implementation footnote. Evaluation records should version context compression, memory visibility, state persistence, tool surface, retry budget, entitlement, and retention route with the model.

SOURCE: ARC PRIZE, OPENAI

PATTERN 02

Distribution policy is part of the model SKU

`Claude Fable 5.1 administrator enablement`

`Gemini 3.8 Flash promotional pricing` `Astra enterprise access off by default`

Three closed-model movements arrived with material entitlement, retention, or price-window conditions. Procurement catalogs should stop recording only model family and API rate; they need surface, administrator state, retention route, promotion expiry, and provider adapter version.

SOURCE: GITHUB, OPENAI

PATTERN 03

Open-weight progress shifted from new lineage to deployment compression

`Muse-Glimmer-30B NVFP4` `NeoMME-260M` `NeoMME-800M`

W36 did not repeat W35's open-frontier release wave. NVIDIA cut an existing checkpoint's footprint 2.4× and H Company opened compact multimodal encoders; both lower implementation friction, but neither supplies an independent frontier comparison. Buyers should value deployability without relabeling it capability convergence.

SOURCE: NVIDIA ON HUGGING FACE, H COMPANY

BENCHMARK MOVES

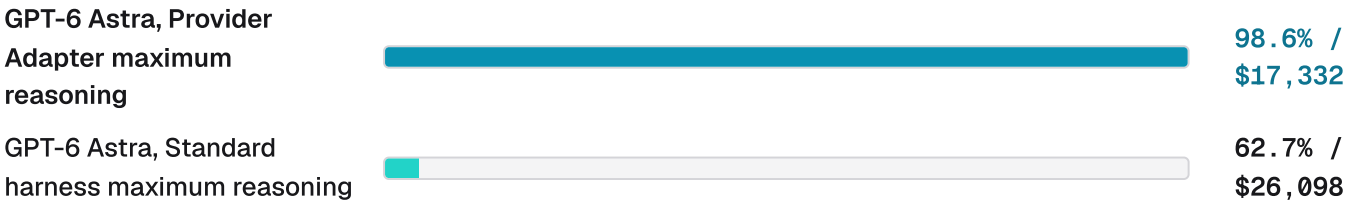
Where the leaderboard moved.

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

ARC-AGI-3 SEMI-PRIVATE

MOVE 01

At matched maximum reasoning, Astra moved from 62.7% on the Standard harness to 98.6% on OpenAI's Provider Adapter, a 35.9-point system spread.

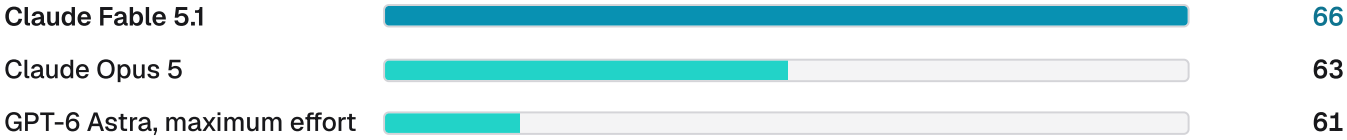


SOURCE: ARC PRIZE

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX V4.1.1

MOVE 02

Astra entered at 61 at maximum effort, reinforcing that one interactive benchmark should not define general capability.

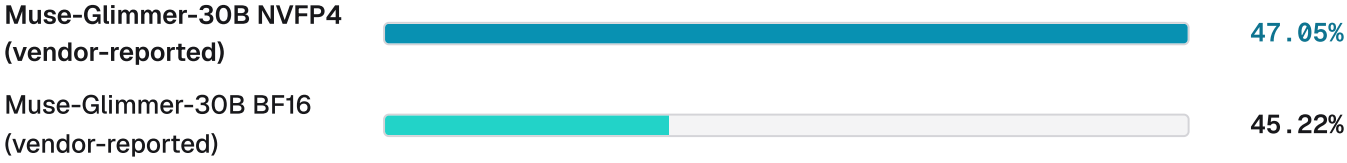


SOURCE: ARTIFICIAL ANALYSIS

TERMINAL-BENCH 2.1, VENDOR-REPORTED CHECKPOINT COMPARISON

MOVE 03

NVIDIA reports 47.05% for Muse-Glimmer NVFP4 versus 45.22% BF16, while cautioning that the small gain may be sampling noise.



SOURCE: NVIDIA ON HUGGING FACE

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Sep 5, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	**Claude Fable 5.1**	Claude Opus 5	Artificial Analysis Index remains the broad capability anchor; Astra's adapter result is harness-specific and should not replace it.
Open frontier	**GLM-5.3-Flash**	IBM Granite 4.2 30B	No new open-frontier release displaced W35's leaders; Muse-Glimmer NVFP4 changes footprint, not lineage or independent rank.
Reasoning	**Claude Fable 5.1**	GPT-6 Astra	Astra is the system-design challenger; persistent state materially changes ARC results, while Fable retains the broader independent Index lead.
Coding	**Claude Fable 5.1**	Gemini 3.8 Flash	Both gained GitHub distribution; compare completed-task cost after administrator policy and year-end Gemini pricing are included.
Multimodal	**Gemini 3.8 Flash**	GPT-6 Astra	Astra adds 1M context and computer use; Gemini's Copilot rollout broadens surface reach, but GitHub publishes no neutral task benchmark.
Edge / small	**NeoMME-800M**	NeoMME-260M	Specialist retrieval encoders, not generative leaders; test for multimodal memory only after measuring latency and retrieval quality.

VENDOR SIGNALS

Pricing, gating, deprecation.

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

2026-09-03
SIGNAL 01
OpenAI

Astra enterprise access off by default; Standard API priced at $10/$50 per million input/output tokens

Astra is a premium governed rollout, not an automatic fleet replacement. Require explicit enablement, provider-adapter versioning, reasoning-effort capture, and a portability test before using its adapter result in procurement.

SOURCE: OPENAI

2026-09-01
SIGNAL 02
Anthropic / GitHub

Fable 5.1 reaches general availability but Business and Enterprise administrators must enable it

Public availability does not equal enterprise entitlement. Catalog the selected safety and retention route, including the documented 30-day default outside approved zero-data-retention arrangements.

SOURCE: GITHUB

2026-09-03
SIGNAL 03
Google / GitHub

Gemini 3.8 Flash enters Copilot under introductory provider pricing through December 31

Do not annualize current economics into 2027. Measure completed-task cost now, then rerun the same workload when post-promotion pricing is published.

SOURCE: GITHUB

2026-09-01
SIGNAL 04
NVIDIA

Muse-Glimmer-30B NVFP4 ships Apache 2.0 for Blackwell with a 24.7 GB checkpoint

Blackwell self-hosters gain a smaller deployable artifact. Treat the vendor-reported benchmark delta as noise until reproduced, and value the release on footprint and runtime compatibility.

SOURCE: NVIDIA ON HUGGING FACE

WATCHLIST

On the radar next.

5 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.

SEP 9

WATCH 01

GLM-5.3-Flash promotion expires

Recalculate the W35 open-frontier cost case on steady-state list pricing and independent provider variance.

SEPTEMBER

WATCH 02

Neutral Astra memory-harness reproductions

A visible-state harness closing part of the 35.9-point matched-effort gap would make the runtime gain more portable.

Q4 2026

WATCH 03

Anthropic Enterprise Frontier Safeguards rollout

Look for precision, recall, rolling-window, alert-volume, audit, and customer-cloud cost evidence before approval.

DEC 31

WATCH 04

Gemini 3.8 Flash introductory pricing expiry

The post-promotion rate determines whether today's Copilot task economics survive into 2027.

NEXT AA REFRESH

WATCH 05

Astra and W35 open-model independent rankings

Use a common harness to test whether Astra's system advantage and W35's vendor-reported open scores generalize.

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 20

Week 36 of 2026 · September 5, 2026

WEB

brianletort.ai/industry/model

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