

THE BIG READ

Three frontier-scale open checkpoints in 72 hours, and the model card is no longer the number to plan on

KEY SIGNALS THIS PERIOD

5

Models added to the tree

6

Vendors shipped frontier-class

5

Architectural patterns crossed multiple vendors

6

Benchmark moves

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

October 17, 2026

Issue 26 · 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 downloadable artifact arrived this week and the model card stopped being the number a buyer can plan on. Three open checkpoints became downloadable between October 15 and October 17: Meta's Muse Spark Open (816B total parameters) under the Muse Model License, Reflection's Beam 1.0 (501B) under Apache 2.0 after W41's deferral, and DeepSeek's V4.1-Flash-1017 (763B) under MIT. All three are mixture-of-experts models, which activate a few expert sub-networks per token, so the active counts (42B, 23B and roughly 8-16B) are what per-token compute looks like. Each card's agentic rows are high: Terminal-Bench 2.1 (a benchmark of multi-step tasks completed in a terminal) of 76.4 and 80.1 for the first two, Terminal-Bench 4.0 of 42.3 for the third. The independent runs that exist land far lower on the same family of tests. Artificial Analysis (outside this week's graded packet, meaning gathered after the graded source set closed and not re-verified) scores Beam 30 on its Intelligence Index (its composite score across several evaluations) with 3.0% on its own Terminal-Bench 4.0 run, puts Muse Spark Open at 8.1% on the same benchmark, and Latent Space reads its tables as placing Spark Open last among the top nine open models on the four agentic evaluations. DeepSeek's 42.3 was measured inside the Claude Code harness; no independent run of the 1017 snapshot exists yet. The circulating 'US open weights are back' frame (Meta, Artificial Analysis, Interconnects, Latent Space) is accurate about the best US open score changing hands twice in about 36 hours and silent about the gap to the card. The gaps are unreproduced, not refuted: the task sets differ (2.1 versus 4.0), effort settings are unmatched and the harnesses differ, and a broken run is one live explanation. The reading here is that the gap has moved from 'the weights are not downloadable' (W41) to 'the downloadable weights have not reproduced their card', and the procurement consequence is that an independent harness run, not the repository, is now the acceptance test.

The second thing the week settled is where the best US open score sits. On Artificial Analysis's index it moved from 26 (Nemotron 3.5 Super VL, the morning of October 15) to 39 (Muse Spark Open, that afternoon), with Beam at 30 the

next day; all three finished behind MiMo-V2.6-Pro (46), GLM-5.3 (45), Qwen3.8 Max (45) and Kimi K3 (44), and level with the DeepSeek V4.1 Flash snapshot that preceded this week's update. The tree adds all five in-window checkpoints (Muse Wonder, Muse Spark Open, Beam, Nemotron 3.5 Super VL, LFM2.5-24B-A2B), four at medium confidence, and updates two rows; the tree delta has the detail. The licence table is the part to read before the scores: MIT (DeepSeek), Apache 2.0 (Beam), OpenMDW-1.1 (Nemotron) and the Muse Model License; the architecture watch carries the Muse terms in full. The AI Stack Weekly's synthesis observes that this week's licences line up with who finances the releaser; this issue is the fact home for the terms themselves.

The closed tier moved too, on price and availability rather than on score. Gemini 4 Argon reached every paid Gemini API developer on October 13 at an introductory \$2 / \$10 per million tokens (cached input 95% off), with Google's pricing page dating the rate through November 30 and listing \$4 / \$20 from December 1; Argon is first on LMArena's October 15 Text board at 1524 and scores 53 on the independent index, level with GPT-6 Astra. Muse Wonder arrived the day before at \$4 / \$20, Claude Opus 5.5's list price, scoring 52 on the index at \$4.52 per index task, which Artificial Analysis calls higher than most similarly intelligent models; its cost per task is 2.8x Muse Spark 1.3's and its input tokens per task rose 28% (6.03M to 7.69M). For a router, the change this week is that the \$2 / \$10 band now has three purchasable models (Sonnet 5.5, GPT-6.1 Sol and Argon) and one of them has a published doubling date; the scorecard moves Argon to closed-frontier challenger and multimodal leader on availability. If you are sizing an on-prem deployment this quarter, budget it against the independent row, treat the card as the vendor's upper bound, and carry a reversion date per hosted model in your routing table.

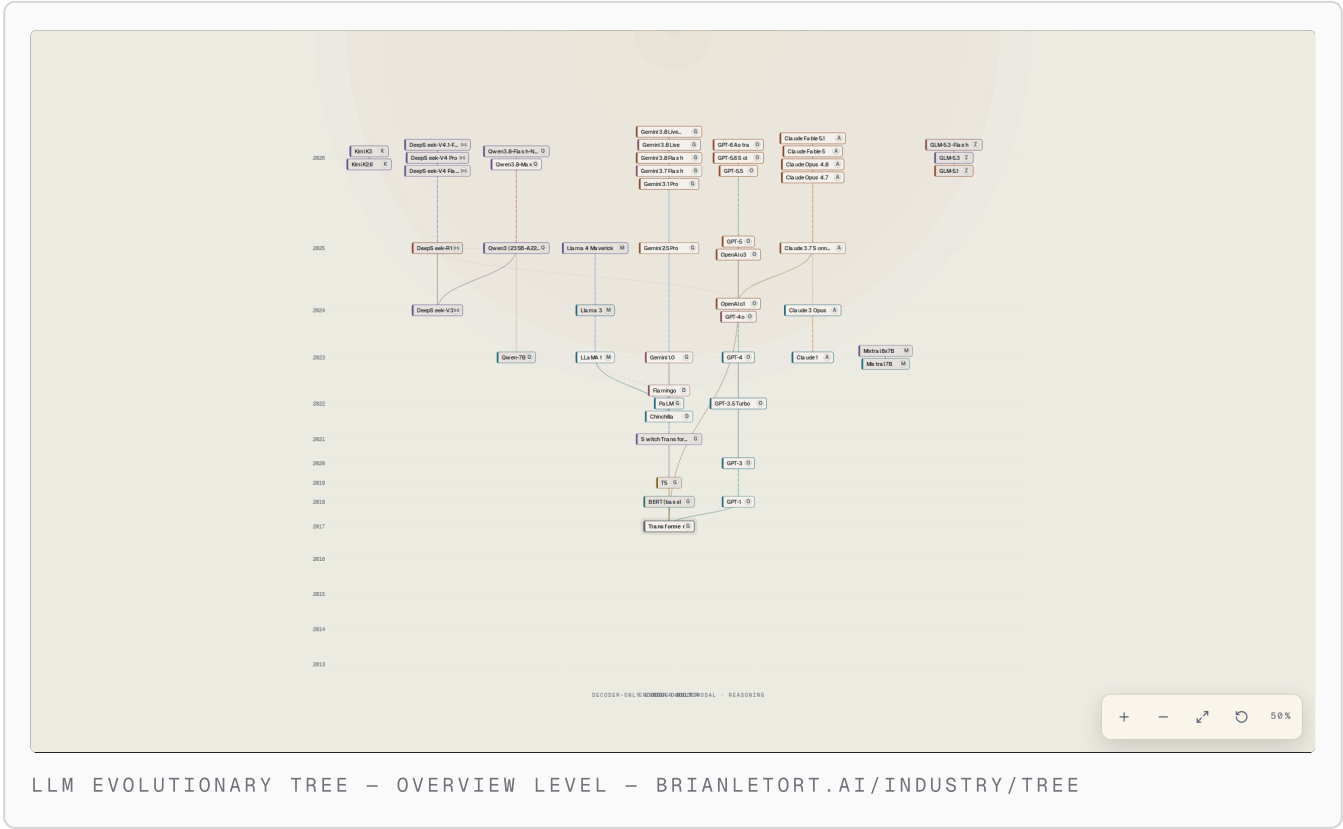
THE BAR

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.

Five rows added (Muse Wonder on reasoning under Muse Spark 1.2; Muse Spark Open on reasoning under Muse Spark 1.2 with the commercial_use_restricted tag; Reflection Beam 1.0 on reasoning with no parent, created now that the repository exists; Nemotron 3.5 Super VL on multimodal under Nemotron 3 Nano Omni; LFM2.5-24B-A2B on reasoning under LFM2.5-2.6B), two updated (DeepSeek V4.1 Flash for the 1017 snapshot and the 763B MIT checkpoint; Gemini 4 Argon from gated to closed and active on October 13 general availability), four deferred and five excluded. Four of the five additions are at medium confidence (muse-wonder, muse-spark-open, reflection-beam, lfm2-5-24b-a2b) because an independent harness scores them far below their cards, because reviewers have withheld a verdict, or because the licence file was not in the graded packet; nemotron-3-5-super-vl is at high confidence on a fully documented architecture.



ADDED (5)

- muse-wonder
- muse-spark-open
- reflection-beam
- nemotron-3-5-super-vl
- lfm2-5-24b-a2b

UPDATED (2)

- deepseek-v4-1-flash
- gemini-4-argon

Placement notes. Muse Spark Open goes on reasoning rather than mixture_of_experts because Meta positions it as a Spark-family model and the Spark line sits on reasoning; the MoE structure is a secondary

tag, as it is for DeepSeek-R1. Beam's five effort levels are the architectural primitive Reflection markets, so it joins reasoning without a parent (a new lab with no prior row). LFM2.5-24B-A2B is an always-think model trained in agent harnesses, which is why it is placed on reasoning rather than beside its dense 2.6B sibling on decoder_only. The vendor on Muse rows is Meta, matching muse-spark-1-2 and muse-glimmer-30b. DeepSeek V4.1 Flash is updated rather than forked because DeepSeek serves 1017 under the same deepseek-flash alias and describes it as a point update; the row now records 763B as the parameter count on the repository's evidence. The medium-confidence flags on muse-spark-open, reflection-beam and muse-wonder record that no harness their vendor did not choose has reproduced the card or that reviewers have withheld a verdict; the flag on lfm2-5-24b-a2b records a licence file not in the packet.

FRONTIER MOVEMENTS

Flagship-class releases.

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

2026-10-13

RELEASE 01

Google DeepMind

FRONTIER

REASONING

Gemini 4 Argon

Generally available to every paid Gemini API developer at an introductory \$2 / \$10, with a published \$4 / \$20 reversion and the top LMArena Text score

Teams that held Argon out of their router because nobody outside the Fairwind Program (Google's invitation-only early-access programme) could call it can now put it in, and the decision to make is the date, not the model: the \$2 / \$10 rate (cached input 95% off, so \$0.10 per million cached) runs through November 30 and doubles to \$4 / \$20, which is Claude Opus 5.5's list price, from December 1, both dates on Google's pricing page. At the introductory rate Artificial Analysis measures \$1.99 per index task for a score of 53, level with GPT-6 Astra; at the reverted rate that becomes \$3.98, still below Astra on the tracker's own cost-to-run measure. LMArena's October 15 Text board puts Argon first at 1524 ± 9 . Google's launch claims (a 2.85x vLLM inference speedup on Gemma 4 E2B, a Chromium refactor across a roughly 6M-line legacy folder, first on the Vals Index) are vendor-reported. The model is not on the free API tier. Budget any Q1 2027 workload at \$4 / \$20 and treat the next six weeks as the discount.

SOURCE: GOOGLE (THE KEYWORD, DEVELOPERS); ARTIFICIAL ANALYSIS; LMARENA

2026-10-12

RELEASE 02

Meta

FRONTIER

REASONING

Muse Wonder

Meta's first frontier-class model prices at Opus 5.5's list and costs 2.8x Spark 1.3 per task to run the independent index

For a buyer the number that matters is not the index score (52 at max effort, 56 for the subscription-only Wonder Full, against 48 for Muse Spark 1.3) but the cost per index task: \$4.52 against \$1.60 for Spark 1.3, 2.8x, with input tokens per task up 28% (6.03M to 7.69M) and the rest of the increase from the higher rate card; Artificial Analysis calls the cost higher than most similarly intelligent models. A model that reads that much context per task at \$4 / \$0.20 cached / \$20 per million is expensive in exactly the agentic loops it is trained for, and the Active Memory and Active Focus behaviours Meta describes (cross-conversation retrieval; checking in before consequential actions) are model behaviours rather than harness guarantees, which Agent Techniques covers. LMArena's October 15 Text board puts Wonder (Max) second at 1509 behind Argon. Zvi Mowshowitz's 'benchmaxxing' caution and Interconnects' 'vibe checks solid' are the two reviewer positions in circulation, and both are withholding a full verdict; the placement here is medium confidence for that reason. Wonder Full has no per-token API and is deferred on the tree. If you evaluate Wonder, instrument input tokens per task before anything else.

SOURCE: META AI RESEARCH; ARTIFICIAL ANALYSIS; LMARENA

OPEN WEIGHTS

Open-frontier and open-source drops.

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

2026-10-15

RELEASE 01

Meta

OPEN FRONTIER

MOE

Muse Spark Open

The most capable US open-weight model on the independent index (39), under a licence with an attribution clause and a \$20M model-as-a-service gate

Platform teams get an 816B / 42B-active MoE with text, image, audio and video input, a 1,048,576-token context, BF16 and NVFP4 weights, and day-zero llama.cpp, vLLM and SGLang support (SGLang's PR was still open at week's end), with Meta demonstrating Muse Code workflows on two M5 Ultra Mac Studios via MLX. Two things to read before the benchmarks. First, the licence, whose terms the architecture watch carries in full: attribution, an acceptable-use policy that applies to derivatives and outputs (so synthetic-data use inherits it), and a \$20M model-as-a-service revenue gate; Interconnects calls it an unnecessary strategic risk, and the tree carries the commercial_use_restricted tag. Second, the independent agentic numbers: 8.1% on Artificial Analysis's Terminal-Bench 4.0 and, per Latent Space's reading of the tables, last among the top nine open models on its four agentic evaluations (outside the graded packet), against Meta-reported Terminal-Bench 2.1 of 76.4 and OSWorld Verified of 80.4 (a benchmark of desktop-computer tasks). The research packet reads the card as sharing Spark 1.3's base; Meta's card does not name the base, and the nine-point index gap to Spark 1.3 (48) is why Interconnects reads it as a safety-tuned Spark 1.2, and why confidence is medium. The Hacker News thread's top-voted comparison was to Qwen3.8 Flash Next and GLM 5.3 Flash, both runnable on a single B300.

SOURCE: META AI RESEARCH; HUGGING FACE MODEL CARD

2026-10-16

RELEASE 02

Reflection AI

OPEN FRONTIER

MOE

Reflection Beam 1.0

Beam's Apache 2.0 weights land as promised, and the independent run lands at 30 on the index with 3.0% on Terminal-Bench 4.0

W41 deferred Beam until the repository existed; it now does, with 501B / 23B-active text-only weights in BF16, FP8 and NVFP4, a DFlash2 speculative drafter (a small model that proposes tokens for the large one to verify), day-zero transformers, vLLM and SGLang support, a Rust harness and a Safety Technical Report, so the row is created. The card's final figures rose slightly from the announcement (81.6% SWE-bench Verified, 76.5% SWE-bench Pro v2-Hard, 80.1% Terminal-Bench 2.1, 97.7% AIME 2026, 90.6% GPQA Diamond). Artificial Analysis's run, published the same day (outside the graded packet), scores Beam 30 on the Intelligence Index, 3.0% on Terminal-Bench 4.0 and a 47% hallucination rate on AA-Omniscience (its knowledge-and-abstention test); the 2.1-versus-4.0 difference is unlikely to account for a 77-point gap on its own, but no matched run exists and a broken run (wrong chat template, wrong effort setting) is one of the live explanations, so the card is unreproduced rather than refuted. The FT reported on October 10 that NVIDIA, with \$800M invested, is in early talks on a transaction from full acquisition to acqui-hire (outside the graded packet); the licence on the next checkpoint is the thing to watch. For a buyer the Apache 2.0 licence is the cleanest of the week and the independent number is the one to plan on until a matched run exists; Hugging Face is offering \$10,000 in Hub credits for community Beam projects, which may produce the third-party runs the card needs.

SOURCE: REFLECTION AI; HUGGING FACE; ARTIFICIAL ANALYSIS (OUTSIDE THE GRADED PACKET)

2026-10-17

RELEASE 03

DeepSeek AI

OPEN FRONTIER

MOE

DeepSeek V4.1-Flash-1017

A point update under the same alias: MIT weights as a 763B checkpoint, Terminal-Bench 4.0 from 31.2 to 42.3 on the card, price unchanged at \$0.15 / \$0.60

This is the release the week's other open checkpoints are measured against, and it changed nothing about its own price: \$0.15 / \$0.60 per million off-peak (\$0.30 / \$1.20 peak), 384K max output, served under the deepseek-flash alias on DeepSeek's API and published on Hugging Face and ModelScope under MIT as a 763B checkpoint in FP8 with FP4 experts (about 510 GB), which settles the 552B-versus-763B discrepancy the W37 row recorded in favour of the repository. The card's Terminal-Bench 4.0 of 42.3 (31.2 for the prior snapshot; DeepSeek's table lists GPT-5.6 Sol at 39.9 and Claude Opus 5 at 51.8) was measured in the Claude Code harness, which is not the harness Artificial Analysis uses, so 'highest published for a downloadable model' compares numbers measured differently; no independent run of 1017 existed at publication and the-decoder notes the same. DeepSWE v1.1 77.0, Automation-Bench 60.5, Agents' Last Exam 35.4 and HLE 42.6 are card figures. Operational notes: the release omits a Jinja chat template (the file that tells a serving stack how to format a conversation) in favour of a reference encoder and a Rust 'deepseek-recipe' toolkit, and third-party NVFP4, MLX and GGUF quantizations (210-432 GB) appeared within hours. The 'DOA' verdict on Muse Spark Open in circulation (@teortaxesTex, via Latent Space) is anchored on this model's price; the independent score is still the 39 of the prior snapshot until the tracker runs 1017.

SOURCE: HUGGING FACE MODEL CARD; DEEPSEEK API NEWS

2026-10-15

RELEASE 04

NVIDIA

OPEN FRONTIER

MULTIMODAL

Nemotron 3.5 Super VL

A 120B / 12B-active hybrid Mamba-Transformer vision-language MoE under OpenMDW, NVFP4 on day zero, with a 24 GB Bolt derivative

This is the week's clearest example of a model sized to the hardware its publisher sells: NVFP4 checkpoints on day zero, up to 4,856 output tokens per second per GPU on GB200 (vendor figure), a Bolt 74B-A9B derivative compressed to fit 24 GB and measured by NVIDIA at 84 tokens per second on the RTX Spark laptops that went on sale the next day, and day-one support in llama.cpp, Ollama 0.40.5, LM Studio, Unsloth and vLLM. The technical report claims 51.78% on SWE-bench Pro with about 4.6x fewer active parameters than Nemotron 3 Ultra, 71.10% SWE-bench Verified, 66.70% Terminal-Bench 2.1 and 99.62% median NVFP4 accuracy recovery, all vendor-run; Artificial Analysis scores it 26 on the Intelligence Index (outside the graded packet), and the 'up to 5x higher throughput' claim names no comparison model or operating point. OpenMDW-1.1 is a permissive licence with no revenue gate. For an architect deciding what runs on a 128 GB notebook or a single workstation GPU, Bolt is the first in-window artifact that was built for that envelope rather than quantized down to it. The row is placed at high confidence because the architecture, training recipe and licence are fully documented in the technical report, and the confidence flag records how well the tree knows what the artifact is, not how well its card has been reproduced; the scores are labelled vendor-run where they are.

SOURCE: NVIDIA BLOG; NEMOTRON 3.5 TECHNICAL REPORT

2026-10-13

RELEASE 05

Liquid AI

EDGE / SMALL

MOE

LFM2.5-24B-A2B

Liquid's largest LFM2.5: a 24B / 2B-active always-think MoE with 131K context at 184 tokens per second on a laptop, vendor-measured

Edge and on-device teams get the first LFM with a 131K context (from 32K), a 128K vocabulary, a hybrid of gated short convolutions and grouped-query attention under 64 experts with top-4 routing, and base plus post-trained checkpoints with GGUF, MLX, ONNX, vLLM and SGLang support. Liquid's table has it winning five of eight agentic benchmarks against Qwen3.6-35B-A3B (48.09 versus 38.19 on LiveClaw, 27.69 versus 8.72 on AutomationBench) and running at 184 tokens per second on an Apple M5 Max and up to 21K output tokens per second on a single B300; every figure is vendor-run and no independent index placement is in the graded packet. The licence file was not in the packet either, and the 2.6B row's LFM Open License v1.0 carried a \$10M revenue gate with automatic termination on non-compliance, so the row is open_weights at medium confidence pending that check; teams above the gate should read the licence before the benchmarks. Liquid's other event this week was d1-ultra taking the Decision Index lead at 65.2 (outside the graded packet), which keeps the decision-model category deferred but moves its leader away from the startup that defined it.

SOURCE: LIQUID AI BLOG; HUGGING FACE MODEL CARD

ARCHITECTURE WATCH

Patterns to track.

5 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 Card versus harness: the open-weight gap moves from downloadability to reproducibility

Reflection Beam: card 80.1% Terminal-Bench 2.1 / 81.6% SWE-bench Verified; Artificial Analysis index 30, Terminal-Bench 4.0 3.0%, 47% hallucination (outside the graded packet)

Muse Spark Open: card Terminal-Bench 2.1 76.4, OSWorld Verified 80.4; Artificial Analysis Terminal-Bench 4.0 8.1%, last of top nine open models on agentic evals per Latent Space's reading (outside the graded packet)

DeepSeek V4.1-Flash-1017: card Terminal-Bench 4.0 42.3 in the Claude Code harness; no independent run yet

Nemotron 3.5 Super VL: card 51.78% SWE-bench Pro; Artificial Analysis index 26 (outside the graded packet)

Single reference point: Haiku 5.5 in W41, vendor 39.2% versus independent 32.8% on Terminal-Bench 4.0, a 6.4-point gap on one model

Last week the open-weight problem was that the label came without an artifact; this week every artifact arrived and the scores came apart instead. The independent Artificial Analysis figures for Beam, Muse Spark Open and Nemotron sit outside the graded packet and are labelled at every use; the vendor figures are from the cards. Four mechanisms could explain the gap and none of them is fraud: Terminal-Bench 2.1 and 4.0 are different task sets; DeepSeek measured inside the Claude Code harness, which supplies tool scaffolding the tracker's harness does not; reasoning-effort settings are rarely matched; and an independent run can itself be broken (wrong chat template, wrong effort level), which for a two-day-old checkpoint is a live possibility. The one same-benchmark reference point, Haiku 5.5's 6.4-point vendor-versus-tracker gap in W41, is a single model and not a distribution, so it bounds nothing; it says only that a harness effect of that size exists. The figures are therefore unreproduced, not refuted. The procurement consequence is the same whatever the mechanism: a model card's agentic row is the vendor's upper bound under the vendor's harness, and the acceptance test for an on-prem deployment is a run in the harness you will actually use. The AI Stack Weekly's pattern watch tracks the label-artifact gap as a seven-week inductive pattern; the confirming condition it names for next week is an independent 1017 run below 35 on Terminal-Bench 4.0 (p138). Until that run exists, the open-frontier scorecard holds the models whose independent scores are known.

SOURCE: REFLECTION AI; META AI RESEARCH; HUGGING FACE (DEEPSEEK);
ARTIFICIAL ANALYSIS (OUTSIDE THE GRADED PACKET)

PATTERN 02

Four licences for four open checkpoints, and the restrictive one came from the lab that does not sell silicon

DeepSeek V4.1-Flash-1017: MIT

Reflection Beam 1.0: Apache 2.0 (NVIDIA \$800M investor; acquisition talks reported by the FT, outside the graded packet)

Nemotron 3.5 Super VL and Bolt: OpenMDW-1.1

Muse Spark Open: Muse Model License ('Built with Muse Spark Open' attribution; acceptable-use policy on derivatives and outputs; separate agreement for model-as-a-service providers with over \$20M in revenue across the preceding 12 months, as the Hugging Face card's licence section states it)

LFM2.5-24B-A2B: licence file not in the graded packet; the 2.6B sibling's LFM Open License v1.0 gated commercial use above \$10M revenue

FLUX 3 Image Klein (excluded from the tree): Kommunity License, commercial output below \$5M revenue only

China, October 16 (outside the graded packet): a draft export catalogue proposing licences for exporting model weights, checkpoints and source code above 2T parameters, 10^{26} training FLOPs or a global top-10 ranking, with automatic 'report and export' licences for open-sourcing and an exemption for weights already public; DeepSeek's 763B release sits under the thresholds

This is the fact home for the week's licence terms; the AI Stack Weekly's synthesis observes that the terms line up with who finances the releaser, and this entry supplies the table it rests on. Two of the three permissive frontier-scale licences this week (Apache 2.0 on Beam, OpenMDW on Nemotron) came from a lab NVIDIA has invested \$800M in and from NVIDIA itself; the third (MIT) came from DeepSeek, as every DeepSeek release has. The one restrictive licence came from Meta, whose financier is its own advertising business. For a platform team the operative clauses are concrete: Muse's attribution requirement travels with derivatives, its acceptable-use policy covers outputs (so a post-training set distilled from Muse Spark Open inherits it), and the \$20M gate is on the provider's revenue across the preceding 12 months, not revenue from the model; secondary summaries differ on the period (Interconnects writes 'per month', Latent Space 'annual'), and the licence file in the repository governs. Interconnects' observation that Western open models historically out-adopt Chinese models of equal capability is the strongest argument that the licence will be ignored in practice, as Llama's 700-million-user clause largely was; Interconnects' open question is whether the output clause chills synthetic-data use, which is the fastest-growing consumer of open weights. The China draft catalogue is the week's other licence-adjacent fact: it is a supply-side restriction on the open models that currently lead the independent index, it is a draft, and none of this week's releases crosses its thresholds. The tree records the Muse gate with the `commercial_use_restricted` tag so the filter on `/industry/tree` shows it. The weekly's open question on NVIDIA and Reflection is the test for the Apache 2.0 leg: if a transaction closes and Beam's next checkpoint is relicensed, the pattern breaks.

SOURCE: HUGGING FACE MODEL CARDS (META, REFLECTION, DEEPSEEK); NVIDIA; LIQUID AI

PATTERN 03

NVFP4 as the day-zero checkpoint format, and a serving stack that absorbed five frontier-scale releases in a week

NVFP4 (NVIDIA's 4-bit floating-point weight format) checkpoints on day zero: Beam, Muse Spark Open, Nemotron 3.5 Super VL and Bolt, FLUX 3 Image Klein; DeepSeek ships FP8 with FP4 experts and third-party NVFP4 within hours

vLLM v0.32.0 (Oct 18, added at the research window's close): 814 commits from 363 contributors, NVFP4 KV cache (the cache of attention state that holds a conversation's context in GPU memory) on SM8x / SM90 / SM12x GPU generations, FP8 KV cache by default for GLM-5.x on SM100, default-on POST /v1/systemone structured-decision endpoint, generic 'hf' tool and reasoning parsers, Python 3.10 dropped, default CORS restricted

llama.cpp b11671: Muse Spark Open text, vision and audio (PR #30616), Q8_0 / Q4_K_S / Q2_K GGUF builds validated against the full-precision model; some layers held at 8-bit because lower-bit versions produced looping output

Ollama 0.40.5 (Nemotron 3.5 Super and Bolt) and 0.40.6 (q2_0 tensors); SGLang PR #44530 (+41,020 lines) still open

Muse Spark Open on two M5 Ultra Mac Studios via MLX; Bolt 74B-A9B at 84 tokens per second on RTX Spark (vendor)

The deployability half of the open-weight story is now measured in hours rather than weeks, and the format the checkpoints arrive in is NVIDIA's. Four of the week's five open releases shipped an NVFP4 checkpoint on day zero, vLLM's 814-commit release added an NVFP4 KV cache, and the NVIDIA-generation numbering of the supported architectures (SM8x, SM90, SM12x) tells an operator which hardware the format was built for; the hyperscaler-silicon and AMD paths run the BF16 or FP8 weights and their own quantizers. The llama.cpp detail is the one for anyone planning a 2-bit deployment: the Muse Spark Open quantizer keeps some layers at 8-bit because the low-bit types caused looping, so the headline Q2_K size understates the memory a working build needs. vLLM's breaking changes (Python 3.10 dropped, default CORS restricted, three GitHub security advisories fixed) are the week's upgrade-gating items for platform teams, and the default-on structured-decision endpoint is the serving layer catching up with the decision-model category the tree still defers. The AI Stack Weekly's hardware lens carries the notebooks and workstation GPUs these artifacts were sized for; the point here is that the serving stack is no longer the bottleneck between a repository and a deployment, which moves the bottleneck to the independent evaluation the first architecture-watch entry describes.

SOURCE: VLLM GITHUB RELEASES; LLAMA.CPP PR #30616; OLLAMA RELEASES; SGLANG PR #44530

PATTERN 04

Introductory rates with published reversions: the \$2 / \$10 band now has a calendar

Gemini 4 Argon: \$2 / \$10 introductory through November 30, cached input 95% off; \$4 / \$20 from December 1, both dates on Google's pricing page

Muse Wonder: \$4 / \$0.20 cached / \$20, no promotion; Muse Wonder Full subscription-only

DeepSeek V4.1-Flash-1017: \$0.15 / \$0.60 off-peak, \$0.30 / \$1.20 peak, unchanged through a capability update

Reference rows from W41: Sonnet 5.5 \$2 / \$10 with \$0.10 cache reads; GPT-6.1 Sol \$2 / \$10; Mistral Large 4 \$0.68 / \$2.09 on an undated 50% discount

This is the fact home for the week's API rate cards (OpenAI's ChatGPT Enterprise per-token card is a seat-plus-usage contract and lives in The Application Layer). Three purchasable models now sit at \$2 / \$10 (Sonnet 5.5, GPT-6.1 Sol, Argon) and one of them carries a published doubling: Google's pricing page runs the \$2 / \$10 rate through November 30 and lists \$4 / \$20 from December 1. That makes Argon's effective price a function of when the workload runs, which is the seventh week in a row that the AI Stack Weekly's pattern watch has recorded a list price moving through a non-list mechanism (dated promotions, multipliers, credit caps). Muse Wonder's \$4 / \$20 is the first new frontier release in months to launch at Opus 5.5's list with no promotion, and its \$4.52 per index task says the list price is not where its cost sits. DeepSeek is the control: a capability update with no price change, at roughly a thirteenth of the band's input price and a sixteenth of its output price. For a router the rule stays what it was in W41, key on prompt-length band and call type, with one addition: carry a reversion date per model, and re-run the routing table on December 1.

SOURCE: GEMINI API PRICING; META DEVELOPERS; DEEPSEEK API NEWS

PATTERN 05

The Decision Index flips in 48 hours and the LLMs still win on knowledge

Artificial Analysis Decision Index v1.0 (Oct 13): AA-Omniscience-Decision, HLE-MultiChoice, When2Call, WildGuardTest at 25% each

Launch leader TypeSafe Jev 1.13 at 58; GLiDE and Microsoft-Decision-1 at 54; Perplexity Decider v1.1 27B cheapest at \$17 per million tasks

Liquid d1-ultra at 65.2 by the Oct 15-16 snapshot at \$0.042 per million input tokens, output free; Jev 1.14 at 57.1, below 1.13 (outside the graded packet)

Claude Haiku 5.5 (Low) scores 61 at 1.4 s per task and \$184 per million tasks against 0.13 s and \$42 for the decision models

The decision-model category (small models that return a classification or approve/flag decision rather than generated text) got its first independent index this week and the index's own numbers make the architectural argument the tree has been waiting on. A general LLM at its lowest effort setting (Haiku 5.5 at 61) still outscores every decision model at launch on the knowledge-heavy components, at roughly 11x the latency and 4x the cost per million tasks; the decision models win on speed and price, which is a serving property, not an architectural one. That is consistent with the fine-tuning-recipe reading of the 3-of-6 test's architectural-distinction condition, so the category stays deferred. The leadership change is the newer fact: the index debuted with Jev 1.13 at 58 and within 48 hours Liquid's d1-ultra posted 65.2 at the same \$0.042 input price (outside the graded packet), with an 85% non-hallucination rate against 35-37% for Jev, while Jev 1.14 scored below 1.13. The graded packet's software-04 still carries the launch-day leader. For an architect routing approve, flag and classify calls, the index gives a first public basis for choosing a decision model on accuracy rather than on price alone, and the first evidence that the cheap tier's ranking is not stable week to week. Artificial Analysis also added FrontierCode 1.1 (leading models fail more than 40% of maintainer-mergeability tasks) and AA-3D the same week; neither moves a scorecard row yet.

SOURCE: ARTIFICIAL ANALYSIS (DECISION INDEX LAUNCH); ARTIFICIAL ANALYSIS CHANGELOG

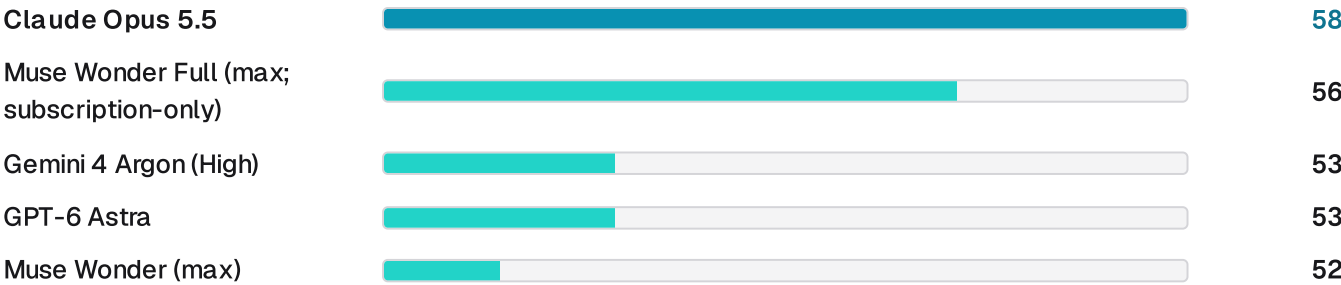
BENCHMARK MOVES

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 (INDEPENDENT), CLOSED TIER MOVE 01

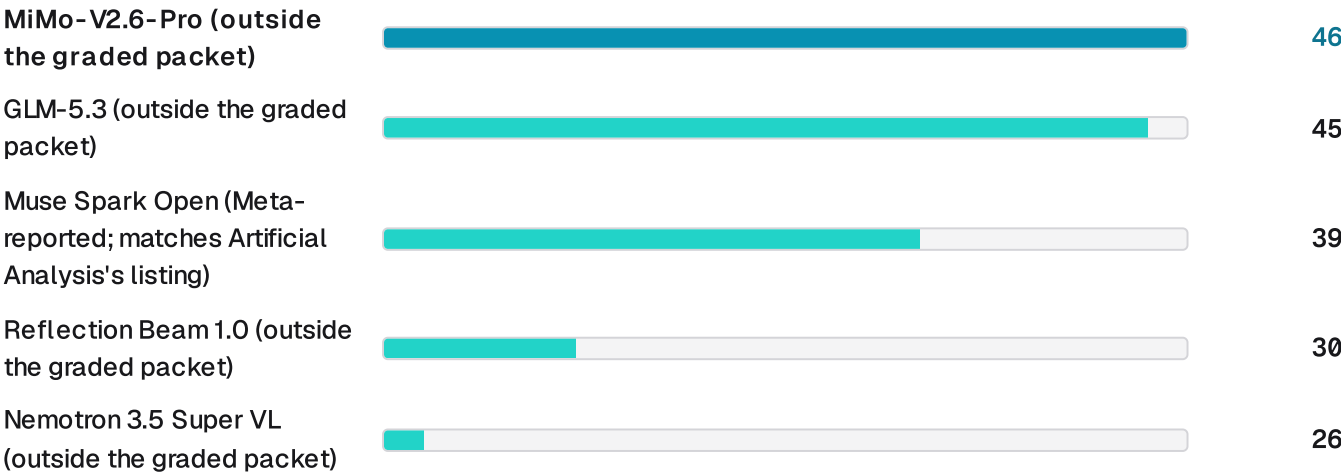
Two new entries from Meta enter the top five; Argon's score is unchanged by general availability. Setting in parentheses is the reasoning-effort level and scores are not comparable across settings



SOURCE: ARTIFICIAL ANALYSIS (MUSE WONDER AND GEMINI 4 ARGON ARTICLES)

ARTIFICIAL ANALYSIS INTELLIGENCE INDEX (INDEPENDENT), OPEN-WEIGHT MOVE 02
TIER

The best US open score moves from 26 to 39 inside 36 hours and every US entrant finishes behind the Chinese models already on the tree; the Beam, Nemotron and MiMo/GLM rows are Artificial Analysis figures outside the graded packet

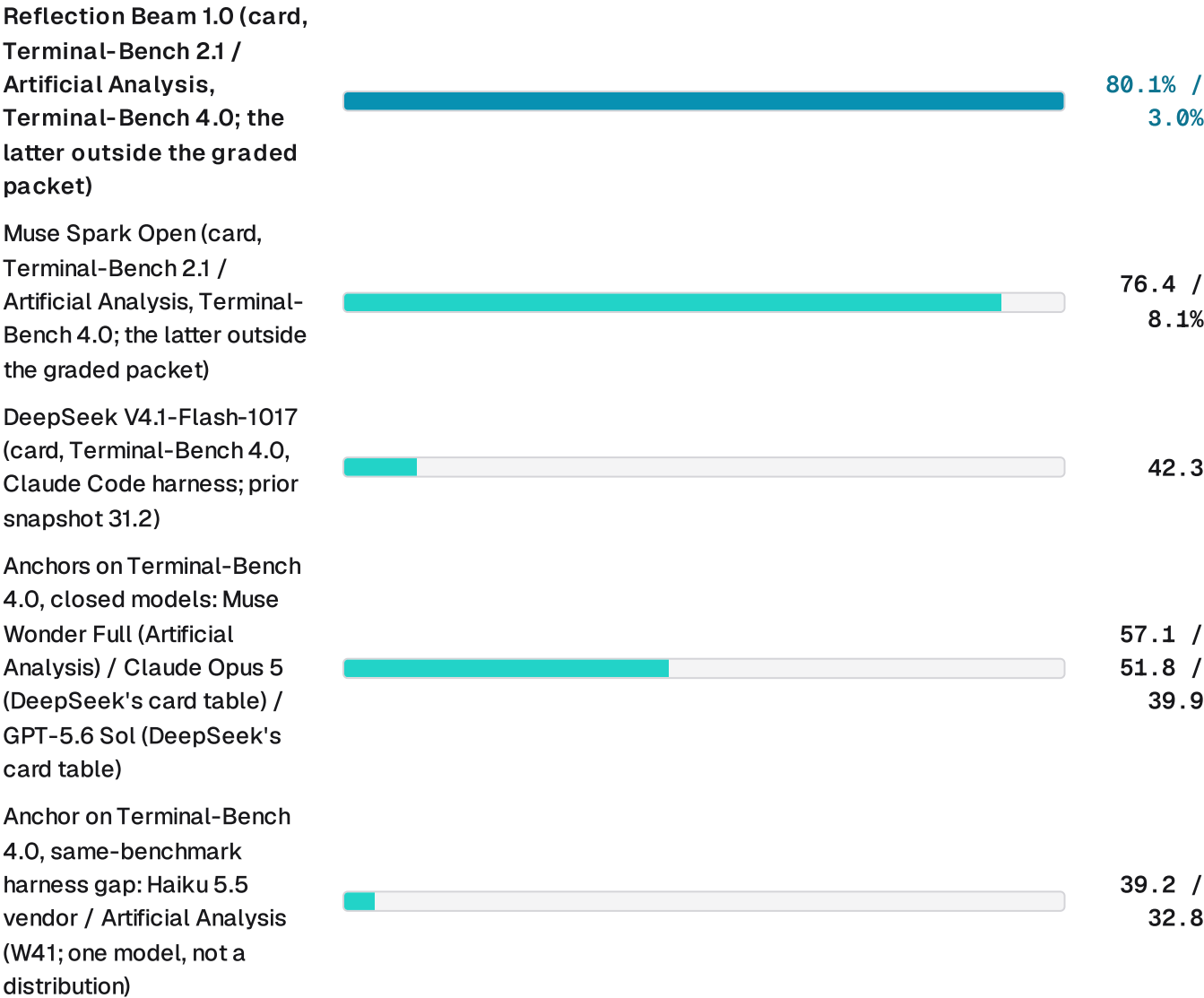


SOURCE: META AI RESEARCH; ARTIFICIAL ANALYSIS (OUTSIDE THE GRADED PACKET)

TERMINAL-BENCH, VENDOR CARD VERSUS INDEPENDENT HARNESS, WITH ANCHORS

MOVE 03

The card-to-independent gap for the week's open checkpoints is 68-77 points where both exist; the task set (2.1 versus 4.0), the harness (DeepSeek measured in Claude Code) and effort settings differ, nobody has yet shown how much each accounts for, and the figures are unreproduced rather than refuted. The anchor rows show where known models sit on 4.0



SOURCE: REFLECTION AI; META AI RESEARCH; HUGGING FACE (DEEPSEEK); ARTIFICIAL ANALYSIS (OUTSIDE THE GRADED PACKET)

LMARENA TEXT ARENA, OCTOBER 15 BOARD (INDEPENDENT, PREFERENCE-BASED)

MOVE
04

Argon takes first on general availability and Muse Wonder debuts second; arena scores measure pairwise human preference, not task completion

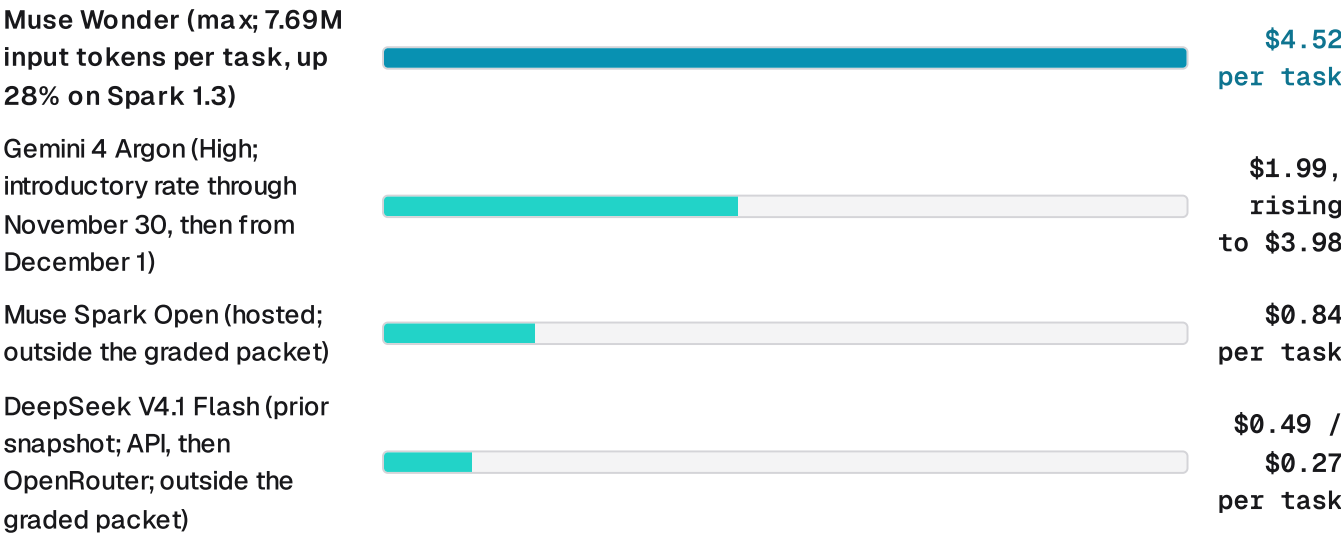


SOURCE: LMARENA

ARTIFICIAL ANALYSIS COST TO RUN THE INTELLIGENCE INDEX (INDEPENDENT)

MOVE
05

Muse Wonder's per-task cost is higher than most similarly intelligent models (Artificial Analysis's wording) and 2.8x Muse Spark 1.3's \$1.60; Argon's cost doubles on reversion; the Muse Spark Open and DeepSeek Flash rows are Artificial Analysis figures outside the graded packet

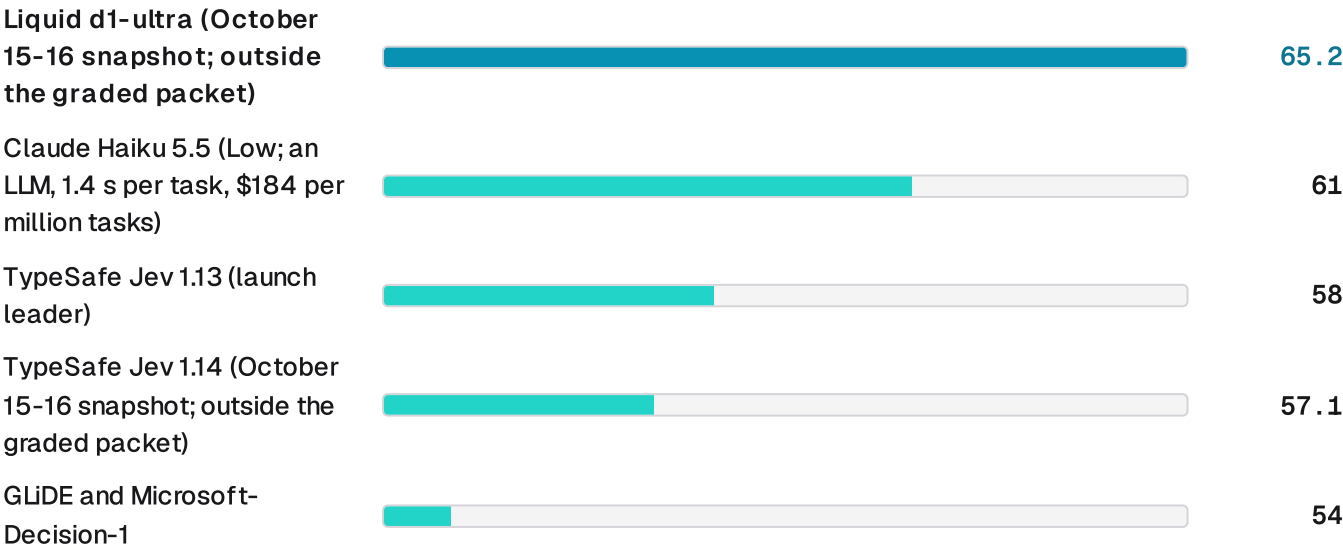


SOURCE: ARTIFICIAL ANALYSIS (MUSE WONDER AND GEMINI 4 ARGON ARTICLES); OPEN ROWS OUTSIDE THE GRADED PACKET

ARTIFICIAL ANALYSIS DECISION INDEX V1.0 (INDEPENDENT), LAUNCHED
OCTOBER 13

MOVE
06

The launch leader was displaced within 48 hours; the d1-ultra and Jev 1.14 rows are from the October 15-16 snapshot (outside the graded packet), the rest from the launch article



SOURCE: ARTIFICIAL ANALYSIS (DECISION INDEX LAUNCH ARTICLE; OCTOBER 15-16 SNAPSHOT OUTSIDE THE GRADED PACKET)

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class.
As of Oct 17, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	Claude Opus 5.5	Gemini 4 Argon	Challenger changes from GPT-6 Astra to Argon. The two tie at 53 on the independent index; Argon became purchasable by every paid Gemini API developer on October 13 and leads LMArena Text at 1524, while GPT-6.1 Astra remains withheld. Muse Wonder Full scores 56 but is subscription-only with no API; Muse Wonder (52) sits below both at \$4.52 per task. Opus 5.5 holds at 58.
Open frontier	MiMo-V2.6-Pro	GLM-5.3	Unchanged. Three frontier-scale open checkpoints shipped and none displaces the slots: Muse Spark Open scores 39, Beam 30 and Nemotron 3.5 Super VL 26 on the independent index (the latter two outside the graded packet), against 46 and 45 for the holders. DeepSeek V4.1-Flash-1017's card figures are vendor-run in a different harness; the row is re-examined when the tracker runs the snapshot.
Reasoning	Claude Opus 5.5	Claude Sonnet 5.5	Unchanged. Muse Wonder's Active Focus and Active Memory are behavioural claims evaluated on Meta's own benchmarks; Beam's five effort levels ship with an independent index of 30. Neither moves the row.
Coding	Claude Sonnet 5.5	GPT-6.1 Sol	Unchanged. DeepSeek's 42.3 on Terminal-Bench 4.0 is a card figure measured in the Claude Code harness; Google's Chromium refactor and 57.4% Terminal-Bench 4.0 for Argon are vendor-run; Beam's 81.6% SWE-bench Verified sits against a 3.0% independent Terminal-Bench 4.0. The row waits for independent coding runs of the three.
Multimodal	Gemini 4 Argon	Gemini 3.8 Live Extended Thinking	Leader and challenger swap. Argon (text, image, audio and video input, 1M output) is now generally available, which was the condition W40 and W41 set for it to hold a leader slot; 3.8 Live

TIER	LEADER	CHALLENGER	READ
			Extended Thinking keeps the live-audio case. Muse Spark Open is the first frontier-scale open model with four input modalities and is tracked on the open-frontier line; Nemotron 3.5 Super VL's vision scores are vendor-run.
Sub-dollar hosted	Claude Haiku 5.5	GPT-6 Luna	Unchanged, with a pending challenge. DeepSeek V4.1-Flash-1017 at \$0.15 / \$0.60 carries a card Terminal-Bench 4.0 (42.3) above Haiku 5.5's independent 32.8%, but the figures are from different harnesses and the tracker has not run the snapshot; it is tracked on the open-frontier line until an independent number exists, as GLM-5.3-Flash was in W41.
Edge / small	LFM2.5-24B-A2B	Nemotron 3.5 Bolt 74B-A9B 24GB	Both slots change, provisionally. The tier covers models meant to run on one device; LFM2.5-24B-A2B takes the slot on a 131K context and a vendor-measured 184 tokens per second on an M5 Max with open weights, and Bolt takes challenger as the first frontier-derived model compressed to a 24 GB envelope (84 tokens per second on RTX Spark, vendor-measured). Both placements rest on vendor figures with no independent index placement, which the other tiers do not allow; the exemption is that no independent tracker yet scores on-device models at these sizes, so the tier is scored on documented architecture and measured throughput until one does, and the slots are re-examined when an independent on-device number exists. EmbeddingGemma 2, last week's leader, is an embedder rather than a generator and moves to the note; its row on the tree is unchanged.

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.

2026-10-15

SIGNAL 01

Meta

Publishes Muse Spark Open under the Muse Model License (terms in the architecture watch) with 'additional safety post-training' on the card; the research packet reads the card as sharing Muse Spark 1.3's base, and Meta's card does not name the base

Meta has reopened a US open-weight line at frontier scale for the first time since Llama 4 and kept its most capable models closed, which Alexandr Wang stated as policy (The Next Web, outside the graded packet); the licence is more restrictive than Muse Glimmer's Apache 2.0 in August. For any provider near the \$20M gate or any team building synthetic data from outputs, the licence is the deployment decision, and the unresolved 1.2-versus-1.3 base question is why the tree's row is medium confidence.

SOURCE: META AI RESEARCH; HUGGING FACE MODEL CARD

2026-10-13

SIGNAL 02

Google DeepMind

Opens Gemini 4 Argon to all paid Gemini API developers and AI Ultra subscribers in AI Studio and Antigravity at \$2 / \$10 with cached input 95% off through November 30; the pricing page lists \$4 / \$20 from December 1; not available on the free tier

The thirteen-day Fairwind gate was shorter than the W40 and W41 watchlists expected, and the price is a promotion with a published destination and date rather than a list. Budget holders should model Argon at \$4 / \$20 for anything running from December 1 and treat the \$2 / \$10 window as the evaluation period; at the reverted rate Argon is priced level with Opus 5.5 and Muse Wonder.

SOURCE: GOOGLE (THE KEYWORD); GEMINI API PRICING

2026-10-17

SIGNAL 03

DeepSeek AI

Ships V4.1-Flash-1017 under the existing deepseek-flash alias with MIT weights (763B, about 510 GB), unchanged \$0.15 / \$0.60 off-peak pricing, a reference encoder in place of a Jinja chat template and a Rust 'deepseek-recipe' toolkit

A capability update with no price change is the behaviour that anchors the 'DOA' verdict on Muse Spark Open in circulation (@teortaxesTex via Latent Space); the alias reuse means API customers were moved to the new snapshot without a model-id change, so platform teams that pin behaviour should re-run their evaluations against the alias and pin to a dated snapshot where the provider offers one. The missing Jinja template is a small operational break for serving stacks that assume one, and the 763B figure resolves the parameter-count discrepancy the tree carried since W37.

SOURCE: DEEPSEEK API NEWS; HUGGING FACE MODEL CARD

2026-10-15

SIGNAL 04

NVIDIA

Releases Nemotron 3.5 Super VL and Bolt under OpenMDW-1.1 with NVFP4 checkpoints on day zero, one day before RTX Spark laptops (128 GB unified memory, from \$2,599) go on sale; the FT reports NVIDIA in early talks to acquire or acqui-hire Reflection AI, whose Apache 2.0 Beam shipped October 16 (FT, outside the graded packet)

NVIDIA is simultaneously the week's second-largest open-weights publisher, the investor behind the third, and the vendor whose number format (NVFP4) four of five open releases shipped in on day zero. For buyers the practical effect is that the open-weight stack is being packaged for NVIDIA hardware first; for the tree the open question is the licence on Beam's next checkpoint if a transaction closes. The 'up to 5x throughput' claim for Super VL names no comparison model or operating point and is carried as vendor-reported.

SOURCE: NVIDIA BLOG; NVIDIA NEWSROOM; FT VIA YAHOO FINANCE (OUTSIDE THE GRADED PACKET)

2026-10-16

SIGNAL 05

Reflection AI

Delivers Beam 1.0 weights under Apache 2.0 in BF16, FP8 and NVFP4 with a DFlash2 drafter, a Safety Technical Report and a Rust harness, with final card figures slightly above the October 5 announcement; Hugging Face offers \$10,000 in Hub credits for community Beam projects

Reflection kept the 'later this month' promise W41 deferred on, which moves it out of the label-without-artifact column; the card figures remain unreproduced, and the independent run that exists is far below them. The Hub-credit programme is the mechanism most likely to produce third-party harness runs, which is what the tree's medium-confidence flag is waiting for. Buyers get the cleanest licence of the week on a model whose independent number is 30.

SOURCE: REFLECTION AI; HUGGING FACE BLOG

2026-10-15

SIGNAL 06

Mistral AI

Releases Timestral, four open zero-shot time-series forecasters (5M to 209M parameters, 21 quantiles, 32,768-step context, CPU-capable) claiming first place among pretrained entries on GIFT-Eval, TIME and fev-bench from a single checkpoint, with fine-tuning through Mistral Forge; Mistral Large 4's weights remain due at the end of October

Mistral's first non-language family lands as open weights two weeks before its trillion-parameter language model's weights are due, which is the sequence a lab follows when the small release carries no competitive cost and the large one does. Timestral is outside the tree's scope; the Large 4 repository and its licence (the AI Stack Weekly bets neither Apache 2.0 nor MIT by November 15) are the Mistral items that move a row. Leaderboard claims are vendor-reported.

SOURCE: MISTRAL AI; GITHUB MISTRALAI/TIMESTRAL

WATCHLIST

On the radar next.

7 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.

NEXT TWO WEEKS

WATCH 01

Artificial Analysis publishes Intelligence Index and Terminal-Bench 4.0 runs for DeepSeek V4.1-Flash-1017

The first independent number for the week's most-cited open checkpoint; a Terminal-Bench 4.0 result below 35 (more than seven points under the card's 42.3) is the AI Stack Weekly's p138 and the confirming condition for its label-artifact pattern, and the result decides the sub-dollar and open-frontier rows.

OCT 23

WATCH 02

Nemotron 3.5 Bolt E4B 4GB release

The second Bolt derivative, sized for 4 GB; it is reviewed for the edge row and for whether NVIDIA's compression recipe holds at a size where the independent index has no comparable open model.

OCT 27-31

WATCH 03

Mistral Large 4 weights and licence on Hugging Face

The tree's closed-at-medium-confidence row flips to open_weights if the files land; the licence resolves the weekly's p132 (neither Apache 2.0 nor MIT by November 15) and adds a fourth data point to the licence-and-financing observation the architecture watch tabulates.

OPEN; FT
REPORTED A
DECISION
POSSIBLE 'WITHIN
WEEKS' ON
OCTOBER 10

WATCH 04

NVIDIA-Reflection transaction terms and the licence on Beam's next checkpoint

A closed transaction with Beam relicensed restrictively would break the week's licence pattern and change the reflection-beam row's openness; a transaction that leaves Apache 2.0 in place confirms it. The talks are single-sourced to the FT and unconfirmed by either company.

DEC 1 (GOOGLE'S
PRICING PAGE
DATES THE \$2 /
\$10 RATE THROUGH
NOVEMBER 30)

WATCH 05

Gemini 4 Argon introductory pricing reverts to \$4 / \$20

The first published doubling in the \$2 / \$10 band; usage share on OpenRouter and the Gemini API before and after is the test of whether a promotional frontier price carries volume through its reversion, and the scorecard's challenger slot is re-examined at the new price.

OPEN

WATCH 06

Independent harness runs for Muse Spark Open and Reflection Beam on any Terminal-Bench version their cards report; Meta's answer on the Spark Open base (1.2 or 1.3)

Both rows are at medium confidence because no harness their vendor did not choose has reproduced the card; a run within five points on the same benchmark version would lift the confidence (and a run that reproduces the gap on a matched version would turn 'unreproduced' into 'refuted'), and Meta's base statement decides the parent on the tree.

Q4 2026

WATCH 07

Artificial Analysis Decision Index second snapshot and any OpenAI entry

The index flipped in 48 hours; a stable second snapshot and OpenAI's absence or presence (the Decisions API GA is the weekly's p133) decide whether the decision-model category gets the third technical report the 3-of-6 test needs.

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 <code>content/llm-tree/models.yaml</code> 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 26

Week 42 of 2026 · October 17, 2026

WEB

brianletort.ai/industry/model

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