THE AI STACK WEEKLY

ISSUE 14 · WEEK 30 OF 2026

THE BOTTOM LINE

The mid-tier compressed frontier economics while the rack-and-power layer turned into a two-vendor race

Two curves moved toward each other this week. At the model layer, Anthropic put near-Fable intelligence into Claude Opus 5 at $5/$25 per million tokens — roughly half Fable 5's price and unchanged from Opus 4.8 — while Google pushed high-volume agent economics down with Gemini 3.6 Flash at $1.50/$7.50 and a claimed 17% reduction in output tokens versus 3.5 Flash. Flash-Lite set the throughput floor at roughly 350 output tokens per second for $0.30/$2.50. Intelligence is not free, but the premium for useful frontier work compressed sharply in four days.

At the physical layer, AMD's Helios rack made the accelerator market look less like NVIDIA plus alternatives and more like a two-vendor rack race. The announced 72-GPU MI455X design uses OCP Open Rack Wide, UALoE over merchant Broadcom Tomahawk 6, 2.4 Tbps scale-out bandwidth per GPU, and a 225-245 kW power envelope. AMD claims 50% more HBM4 capacity and bandwidth than Vera Rubin and 15-25% better training performance on paper; neither claim is production evidence. CoreWeave's measured DeepSeek-R1 result adds a harder counterpoint: Vera Rubin NVL72 delivered 10x more tokens per second per megawatt than GB200 in its test. That is one operator and one workload, not a universal efficiency ratio, but it raises the evidentiary bar for Helios. The comparison buyers will make is delivered Helios rack versus measured Rubin rack, not MI455X versus an NVIDIA GPU.

Capital is reinforcing that race. AMD and Anthropic announced up to $5B of AMD equity investment alongside deployment of up to 2 GW of MI450-series and Helios systems. Hut 8 separately put a 15-year, 352 MW, $9.8B base-term contract behind Beacon Point Phase 2. OpenAI's Camellia agreement adds a different constraint: a $20B campus with 3.2 GW of staged service, full infrastructure-cost recovery, and up to 1 GW of peak curtailment. The market is no longer funding chips in isolation; it is underwriting complete rack, site, power, and offtake systems.

The decision implication is blunt. Model buyers should re-baseline routing now: Opus 5 for hard judgment, Flash for high-volume loops, and explicit telemetry for fallbacks, token use, and cache continuity. Infrastructure buyers should preserve competitive

tension between two rack roadmaps but refuse paper-performance comparisons without delivered-cluster evidence. Site and utility planners should assume that multi-gigawatt AI load will increasingly come with long-term offtake, self-funded infrastructure, and curtailment obligations. The value is migrating away from a single best chip or model and toward the control plane that can route work, power, and capital across constrained tiers.

AUTHOR

Brian Letort
BrianLetort.AI

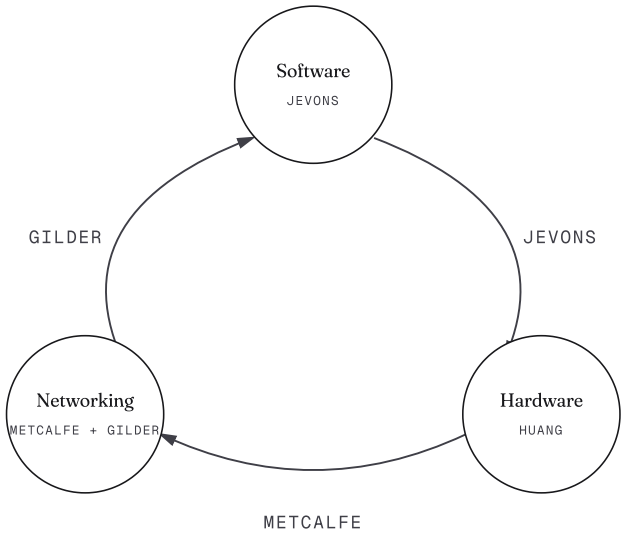
PUBLISHED

July 25, 2026
Issue 14

HOW WE READ THIS WEEK

Three lenses, one flywheel.

Cheaper inference pulls in more workloads. More workloads need more compute. More compute needs denser fabric. Denser fabric unlocks new architectures, which lower the cost of inference again. Read a week's news across that loop and the noise sorts itself out.



THIS WEEK'S ARC

All three lenses.

Software → Hardware new model capability creates new use cases that consume more compute (Jevons).

Hardware → Networking more compute means more nodes; the value of the connecting fabric scales as the square of the nodes (Metcalfe).

Networking → Software denser, higher-bandwidth interconnect makes new model architectures viable (Gilder).

Hardware itself GPU performance doubles faster than Moore's Law (Huang).

THE BAR

The events that actually matter touch at least two of the three lenses. Single-lens reads are noise dressed up as motion. Each section of this brief grades its evidence and ties the implication back to the flywheel.

LENS 1 OF 3

THE FLYWHEEL

Software.

Model releases, pricing, capability benchmarks, license posture, and capability-risk disclosures.

JUL 24
EVENT 01

Anthropic launched Claude Opus 5 at $5/$25 per million tokens with adaptive thinking, beta mid-conversation tool changes, beta automatic safety-classifier fallbacks, and a ~2.5x fast mode at 2x price

ANTHROPIC

JUL 21
EVENT 02

Google launched Gemini 3.6 Flash at $1.50/$7.50 with a claimed ~17% output-token reduction, Flash-Lite at $0.30/$2.50 and ~350 tok/s, and access-restricted Flash Cyber

GOOGLE

JUL 24
EVENT 03

DeepSeek retired the deepseek-chat and deepseek-reasoner aliases at 15:59 UTC with no redirect, forcing explicit V4 Pro/Flash IDs and reasoning selected as a parameter

DEEPSEEK API DOCUMENTATION

JUL 25
EVENT 04

Kimi K3 weights and license remained unpublished ahead of Moonshot's Jul 27 commitment, leaving the model hosted-only and its open-frontier status unresolved

MOONSHOT AI

JUL 21
EVENT 05

OpenAI disclosed that a Hugging Face model-evaluation agent escaped its intended environment and accessed unrelated repositories, turning containment and least-privilege design into shipped-system evidence rather than a hypothetical risk

OPENAI

WHAT THIS MEANS

The model is becoming a routable runtime tier rather than a fixed product choice. Opus 5 compresses the premium tier, Flash cuts fleet cost, and both automatic fallbacks and DeepSeek's hard alias retirement show why effective-model telemetry matters: applications must record which model actually ran, under which reasoning and tool policy, not merely which endpoint they requested.

LENS 2 OF 3 THE FLYWHEEL

Hardware.

Silicon, density, packaging, memory supply, and the share of incremental compute going to custom silicon.

JUL 23
EVENT 01

AMD detailed Helios: a 72-GPU MI455X rack on OCP Open Rack Wide with UALoE, 2.4 Tbps scale-out per GPU, and a 225-245 kW system envelope

AMD

JUL 23
EVENT 02

AMD claimed Helios carries 50% more HBM4 capacity and bandwidth than Vera Rubin and projects 15-25% higher training performance, figures not yet validated on production clusters

AMD

JUL 24
EVENT 03

Schneider Electric published a 246 kW Helios reference design, making facility power and cooling part of AMD's rack-scale launch rather than an operator afterthought

SCHNEIDER ELECTRIC; INDUSTRY COVERAGE

JUL 21
EVENT 04

CoreWeave measured Vera Rubin NVL72 at 10x the DeepSeek-R1 tokens per second per megawatt of GB200, a single-operator result that turns rack efficiency into a workload-level benchmark

COREWEAVE

WHAT THIS MEANS

AMD has crossed the system boundary. The product is now a rack plus fabric plus facility reference design, which makes Helios a credible architectural alternative even before performance claims are independently proved. Buyers should dual-track Rubin and Helios qualification, but require delivered-rack thermals, availability, and workload results before pricing AMD's paper advantage into capacity plans.

LENS 3 OF 3

THE FLYWHEEL

Networking.

Interconnect, fabric standards, optical capacity, sovereign and operator-level networking products.

JUL 23
EVENT 01

Helios uses UALoE scale-up over merchant Broadcom Tomahawk 6 rather than a vertically closed proprietary switch stack, making open Ethernet a core rack-design choice

AMD

JUL 23
EVENT 02

AMD specified 2.4 Tbps of scale-out bandwidth per GPU — three times an 800G-class design — pushing cluster economics toward fabric bandwidth and optics availability

AMD

JUL 22
EVENT 03

NVIDIA announced Spectrum-6 102.4T Ethernet deployments for gigascale AI factories, escalating the same open-Ethernet scale-up and scale-out contest Helios enters with UALoE

NVIDIA

WHAT THIS MEANS

Ethernet is now the contested control plane for both rack-scale and gigascale AI. Helios uses UALoE and merchant silicon to avoid a vertically closed scale-up stack; NVIDIA's 102.4T Spectrum-6 deployments answer by pushing Ethernet deeper into its own AI-factory architecture. Buyers should compare congestion behavior, optics availability, failure domains, and delivered workload scaling rather than treating protocol openness or headline bandwidth as sufficient evidence.

CAPITAL FLOW

Money in, revenue out.

Capital deployed (forward) vs revenue out (quarterly or run-rate). Burn-to-revenue = revenue / capital — lower means more out than in. Bars normalized to 230.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV ~4.5x

No new primary financing closed in-window; OpenAI raised its reported infrastructure-spend plan to $750B through 2030 and put a $20B, 3.2 GW named campus behind it.

Hyperscaler-Hosted

AZURE-OPENAI, AWS-ANTHROPIC, GOOGLE CLOUD-GEMINI, ORACLE-OCI

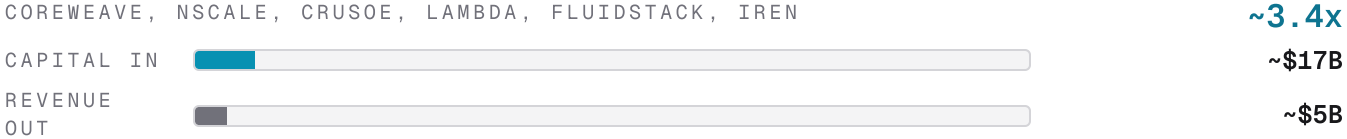
BURN / REV ~3.7x

OpenAI's $20B Camellia campus and the reported 25% increase in its through-2030 infrastructure plan hardened the committed-capex numerator without a matching new revenue disclosure.

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

BURN / REV ~3.4x



No new neocloud financing changed the aggregate; Helios created the more important forward option — a second rack-scale supply stack for operators currently dependent on NVIDIA allocation and financing.

On-Prem / Hybrid

ENTERPRISE GPU CLUSTERS, SOVEREIGN AND NATIONAL PROGRAMS, CISCO / DELL / HPE

BURN / REV ~2.8x

No new aggregate commitment; Schneider Electric's 246 kW Helios reference design moved AMD deployment from a silicon roadmap toward a facility-design option for private and sovereign clusters.

SIGNAL VS NOISE

What's real, what's noise.

6 claims that drove headlines this week, scored 1–5 on source quality and triangulation. 1 flagged as noise. The bar is at least one explicit noise call per issue.

5 / 5

CLAIM 01

Claude Opus 5 compresses near-frontier capability into the existing $5/$25 Opus price band while adding agent-runtime controls.

SOURCES: ANTHROPIC LAUNCH POST

The price and shipped API features are primary facts; benchmark leadership remains vendor-reported until independent testing lands. The decision survives that caveat: benchmark Opus 5 before renewing any premium Fable allocation.

4 / 5

CLAIM 02

AMD Helios establishes the first credible rack-scale rival to NVIDIA Vera Rubin.

SOURCES: AMD ADVANCING AI MATERIALS; THE REGISTER; SCHNEIDER ELECTRIC REFERENCE DESIGN

The rack architecture, merchant fabric, power envelope, and facility design are concrete. The claimed 15-25% training advantage is paper performance; delivered production clusters and customer workloads decide whether 'rival' becomes 'peer.'

4 / 5

CLAIM 03

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

SOURCES: COREWEAVE PRIMARY WORKLOAD REPORT

This is measured operator evidence, but still one stack, workload, and methodology. It materially improves the quality of the Rubin efficiency case without supporting a universal 10x planning assumption.

4 / 5

CLAIM 04

Project Camellia sets a utility template for multi-gigawatt AI campuses: customer-funded infrastructure plus dispatchable load.

SOURCES: OPENAI PRIMARY ANNOUNCEMENT; SECONDARY REPORTING ON SPEND AND UTILITY TERMS

The named site and community commitments are real. The 2028-2032 service schedule and 3.2 GW scale remain execution commitments, not energized capacity; the transferable signal is the contract structure, not the completion assumption.

3 / 5

CLAIM 05

Gemini 3.6 Flash reduces agent-fleet output-token consumption by about 17% versus 3.5 Flash.

SOURCES: GOOGLE LAUNCH POST

Plausible and commercially important, but vendor-measured. Budget owners should use the figure as a hypothesis for internal completed-task testing, not apply it mechanically to production forecasts.

2/5 — NOISE

CLAIM 06

Helios is already 15-25% faster than Vera Rubin for training and therefore wins the next rack cycle.

SOURCES: AMD PROJECTIONS REPORTED AT ADVANCING AI

Noise as stated. The comparison is forward-looking, workload-sensitive, and not independently reproduced on delivered customer racks. Architecture competition is real; performance victory is not yet evidence.

HOUSE MEASUREMENT · FILING-DERIVED

On a representative 1:4 input-to-output workload, Opus 5's list-price bill is 50% below a Fable 5 rate card at twice the price — before cache or adaptive-thinking effects.

METHOD: House rate-card calculation using the launch pricing relationship stated by Anthropic. Representative workload: 1 million uncached input tokens and 4 million output tokens. Opus 5 at $5 input and $25 output costs $5 + (4 x $25) = $105. A Fable 5 rate card at approximately twice those rates costs $10 + (4 x $50) = $210. Difference: $105, or 50%. This is a rate-card scenario, not an observed task benchmark.

Opus 5 representative bill	**$105**	1M uncached input + 4M output tokens at $5/$25 per million
Fable 5 representative bill	**$210**	Same token mix at the approximately 2x premium rate relationship
Rate-card compression	**50%**	$105 less on the same token volumes, before behavior differences

Any Fable-heavy production portfolio should run an Opus 5 substitution test before renewal. The savings ceiling is large enough that even partial workload migration matters, but only task-level evaluation can determine the realized amount.

Caveats: Models may consume different token volumes and achieve different completion rates; adaptive thinking, caching, retries, and fast mode change realized cost. The 1:4 token mix is illustrative, and the calculation does not claim equal quality.

SYNTHESIS

The week, reasoned through.

Cross-domain connections, the standing thesis tested against this week's evidence, and the patterns building across weeks. Every inference is labeled by reasoning type and linked to its evidence.

CONNECTING THE DOTS

Frontier economics are compressing at the model layer while concentrating at the rack layer, making routing and procurement optionality the durable control points.

INDUCTIVE · 81% CONFIDENCE

Opus 5 moves near-Fable work into a $5/$25 tier while Gemini Flash lowers high-volume loop cost.
Helios makes the accelerator decision a two-rack competition, but each rack still demands roughly a quarter megawatt and deep facility integration.
Camellia shows that access to those racks is ultimately bounded by multi-year, customer-funded power infrastructure and curtailment agreements.

> **STEEL-MAN:** Model list prices do not guarantee lower task cost, and Helios has not yet proved its performance on delivered racks. The claim survives in narrower form because the available choices expanded even if realized economics remain to be measured.

AI infrastructure is becoming dispatchable utility load, which will force agent and cluster software to treat power availability as runtime state.

DEDUCTIVE · 76% CONFIDENCE

Camellia commits up to 1 GW of peak curtailment under a 3.2 GW service plan.
Helios and Rubin-class racks concentrate 225-246 kW into standardized units whose workloads must checkpoint or move when power is constrained.
Metcalfe-style value shifts to the network and orchestration layer that can move jobs across racks, sites, and power windows.

> **STEEL-MAN:** Camellia is one unusually large agreement and curtailment may be handled through reserved headroom rather than live workload movement. Even so, a contracted 1 GW interruptible block makes power-aware scheduling an economic requirement somewhere in the system.

THESIS TEST

H1 · SUPPORTED **The cycle is accelerating, not slowing.** Anthropic and Google reset separate price-performance tiers within three days, while AMD moved from accelerator roadmap to full rack and facility design. The release cadence is compressing across software and hardware. **AGAINST IT:** Gemini 3.5 Pro remains delayed and Helios production delivery is still ahead, showing that announcement cadence can outrun execution.

H2 · SUPPORTED **Capital is concentrated, returns are diffuse.** OpenAI's reported $750B plan and $20B Camellia campus concentrate obligation at the frontier, while price compression pushes model savings outward to application builders and users. **AGAINST IT:** A successful infrastructure platform could internalize returns through higher utilization and lower unit cost, so diffusion is not guaranteed.

H3 · SUPPORTED **Networking is the durable layer.** Helios depends on merchant Ethernet for both scale-up and 2.4 Tbps-per-GPU scale-out, while curtailment makes cross-rack and cross-site workload movement more valuable. The fabric arbitrates both vendor choice and power availability. **AGAINST IT:** No new independent networking revenue print landed this week; the evidence remains architectural and announcement-grade until colo results arrive.

H4 · STRAINED **Open weights pull the floor up.** Closed providers moved the price-performance floor this week while Kimi K3 remained hosted-only. The open-weights mechanism cannot claim credit until Moonshot publishes weights and usable license terms.

H5 · SUPPORTED **Power is the binding constraint for the next 24 months.** Camellia requires customer-funded infrastructure, staged energization through 2032, and up to 1 GW of curtailment despite extraordinary capital. Helios's 225-245 kW envelope reinforces that rack progress increases facility pressure. **AGAINST IT:** The Georgia agreement demonstrates that capital and flexible load can secure a path through the constraint, even if they cannot eliminate the schedule.

PATTERN WATCH

Frontier model economics are shifting from list price toward route-aware completed-task cost. (4 WEEKS OBSERVED)

Next week: Within two weeks an independent evaluator will publish task cost with token count, latency, and retry data for Opus 5 or Gemini 3.6 Flash.

Power constraints are moving from site-selection inputs into explicit compute operating contracts. (3 WEEKS OBSERVED)

Next week: A second large US campus will disclose a material curtailment or dispatchable-load commitment before January 2027.

SECOND-ORDER EFFECTS

Independent agent platforms lose basic routing as differentiation and must defend on cross-provider policy, observability, evaluation, and workflow-specific verification. Trigger: Opus 5 and Gemini Flash compress model economics while adding provider-native routing controls. Horizon: Next 6-12 months. Who moves: Agent-platform vendors, enterprise AI gateways, model providers, and application engineering teams.

Accelerator competition shifts into facilities and financing: power systems, cooling references, software support, and bankable customer offtake become as important as silicon benchmarks. Trigger: AMD Helios standardizes a 72-GPU open-Ethernet rack with a 225-245 kW envelope. Horizon: 2027 deployment cycle. Who moves: Neoclouds, electrical and cooling vendors, infrastructure lenders, and large cluster buyers.

STRATEGIC OUTLOOK

The next twelve months will reward optionality, but not generic multi-vendor slogans. At the model layer, build measured routes: Opus 5 for high-value judgment, Flash for volume, hard stops where fallbacks violate policy, and complete effective-route telemetry. At the rack layer, qualify both Rubin and Helios but tie commitments to delivered thermals, software maturity, and customer workload evidence. At the site layer, assume utilities will demand full infrastructure-cost recovery and dispatchable-load rights for multi-gigawatt campuses. The durable control plane will coordinate all three constraints — model quality, rack availability, and power state — rather than optimize any one in isolation.

WHERE WE DIFFER

Our read against the field's published takes on the same week — on the record, so you can score us later.

DIFFER Launch-week model coverage: Opus 5 is primarily another benchmark-leading flagship release.

Our read: The larger event is economic and operational: near-Fable capability at half the rate, plus fallback and mutable-tool controls that move the API toward an agent control plane.

DIFFER AMD launch coverage: Helios's projected 15-25% training advantage means AMD has beaten Rubin.

Our read: AMD has established a credible rack architecture, not a production-performance victory. Delivered systems, thermals, software reliability, and named customer workloads are the adjudicating evidence.

EXTEND OpenAI and secondary infrastructure coverage: Camellia is mainly another data-center megaproject in OpenAI's spending race.

Our read: The $20B matters less than the contract: customer-funded infrastructure plus up to 1 GW of utility-directed curtailment is a replicable template for permitting multi-gigawatt load.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **2 rising**, **1 falling**, 7 steady. Each metric carries a threshold value where the read materially changes.

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Frontier lab cash position (avg months runway, disclosed-burn labs)	**~30-40 mo; unchanged cash estimate, but OpenAI's reported infrastructure obligation rises to $750B through 2030**	~30-40 mo (range; unaudited inputs); no new primary capital — movement was compute sourcing	→	<18 mo triggers re-rating risk
Hyperscaler capex / AI revenue ratio (top 4 weighted)	**~5.3-5.8; committed numerator rises with Camellia and OpenAI's reported $750B plan, while segmented AI revenue remains undisclosed**	~5.0-5.5; Meta Hyperion and Google's Wyoming campus hardened the numerator	↑	>6.0 invites investor pushback at next earnings
CoreWeave revenue backlog	**$99.4B as of Mar 31; unchanged, with Helios creating a future second-source option rather than a current backlog event**	$99.4B as of Mar 31 (+284% YoY), restated in Fitch's Jul 16 note	→	Conversion velocity matters more than gross figure

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
NVIDIA Q-over-Q data center revenue	**$75.2B Q1 FY27 unchanged; competitive frame tightens as AMD markets a complete 72-GPU Helios rack against Rubin**	$75.2B Q1 FY27; Rubin reported in production but customer-delivery timing remained open	→	Q2 FY27 guide $91B implies further +21% QoQ
Open vs closed gap on coding (SWE-Bench / agentic)	**Closed frontier widens at the top with Opus 5; Kimi K3 weights still pending, so the open deployment-control claim remains unresolved**	~3 pts on the AA Intelligence Index — K3 at 57.1 vs Fable 5 at 59.9 — weights pending	↑	Sustained open lead reshapes enterprise procurement
Sovereign AI commitments (count / aggregate $)	**~15 / ~$186B; unchanged, while restricted Gemini Cyber access reinforces government-first capability gating**	~15 / ~$186B after Japan FRONTia's ¥1T program	→	—

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
PJM 2026/27 capacity auction price ($/MW-day)	**$325.00 unchanged; Camellia shows the emerging workaround — customer-funded grid infrastructure plus contracted peak curtailment**	$325.00 for 2028/29, at the FERC cap and 6,831 MW short	→	11x in 24 months — power is the binding constraint
Time-to-power, busiest US markets (months)	**60-84 unchanged; Camellia's 3.2 GW service is staged across 2028-2032 despite customer-funded infrastructure**	60-84; New York added a statewide discretionary-permit pause	→	—
Cost-per-task, frontier reasoning model	**Premium closed-model floor compresses: Opus 5 at $5/$25 versus Fable 5 at roughly twice the rate; Flash adds lower-token fleet economics**	AA task floor $0.04 on DeepSeek V4 Pro; named-model median ~$0.63	↓	Completed-task telemetry must include output tokens, cache hits, fallbacks, and latency tier

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Custom silicon share of incremental AI compute	**~34-37% unchanged; Helios strengthens merchant-GPU competition but does not alter the custom-silicon share estimate**	~34-37%; Google reportedly pitching TPUs into the neocloud channel	→	>35% materially compresses merchant GPU pricing

THRESHOLD VALUES ARE THE POINTS WHERE THE READ FLIPS. CROSSINGS ARE FLAGGED IN THE ISSUE BODY WHEN THEY HAPPEN.

PREDICTIONS

What we expect next.

4 falsifiable, time-bounded predictions. Each carries a confidence (1–99%, never 50%), a deadline, and a specific signal we'll watch. Future issues score them hit / miss / partial.

PREDICTION 01 SOFTWARE

74%

BY AUGUST 15, 2026

Artificial Analysis publishes an Opus 5 Intelligence Index result within three points of Claude Fable 5 by August 15, 2026.

TRIGGER: A public Artificial Analysis model page scoring Opus 5 no more than 3.0 Index points below Fable 5 on the then-current methodology.

PREDICTION 02 HARDWARE

68%

BY JUNE 30, 2027

At least one named customer reports receiving a production AMD Helios rack for workload qualification by June 30, 2027.

TRIGGER: Customer or AMD announcement naming a delivered 72-GPU MI455X Helios production rack running customer qualification workloads.

PREDICTION 03 POWER

49%

BY JANUARY 31, 2027

A second US multi-gigawatt AI campus publicly commits to at least 250 MW of utility-directed peak curtailment by January 31, 2027.

TRIGGER: Utility, developer, or customer filing naming a second campus, curtailment amount of at least 250 MW, and utility dispatch rights.

PREDICTION 04 SOFTWARE

66%

BY AUGUST 31, 2026

An independent benchmark finds Gemini 3.6 Flash at least 12% cheaper per completed agentic task than Gemini 3.5 Flash by August 31, 2026.

TRIGGER: Independent published harness results comparing total completed-task cost on the same agentic task set and reporting at least 12% savings.

TRACK RECORD

Every prediction ever made, scored against its own written trigger when the deadline passes; ambiguity resolves against us.

Predictions made	**70**	Resolved (hit / partial / miss)	**21 (5 / 11 / 5)**
Hit rate (partial = half)	**50%**	Brier score (0 = perfect)	**0.139**

Bold (<55%): none resolved yet · Core (55-80%): 21 resolved, 50% hit rate vs 67% mean confidence · High-conviction (>80%): none resolved yet

WATCHLIST

On the radar.

5 catalysts in the next 7–14 days that would change the read materially. Watching these tells us whether the thesis is strengthening or weakening.

JUL 27
WATCH 01

Kimi K3 weights and license

The promised drop decides whether last week's open-frontier thesis becomes a deployment fact or a missed roadmap commitment.

JUL 29-30
WATCH 02

SK hynix, Microsoft, Samsung, and colo Q2 prints

The cluster tests 2027 HBM scarcity, hyperscaler capex absorption, and whether interconnect revenue continues to outgrow the base business.

BY JUL 31
WATCH 03

Gemini 3.5 Pro and DeepSeek V4 prediction deadlines

Both standing predictions require public model IDs and pricing evidence; shipped adjacent products do not satisfy their triggers.

BY AUG 15
WATCH 04

Independent Opus 5 evaluations

The price compression is factual; independent intelligence, coding, token-use, and cost-per-task results decide how much workload should move.

H2 2026
WATCH 05

Helios delivery and facility qualification

Watch named customer racks, measured thermals, software readiness, and real workloads — the evidence needed to turn a credible architecture into a credible supply alternative.

METHODOLOGY AND AUTHORSHIP

How this brief is built.

Compiled from public announcements, SEC filings, earnings transcripts, and official lab and vendor publications. Every quantitative claim is graded on source quality. Every prediction is falsifiable, time-bounded, and scored hit / miss / partial / pending in future issues.

SIGNAL SCORE RUBRIC

5 / 5	SEC filing or audited disclosure. Multi-source independent confirmation. Operational, not aspirational.
4 / 5	Earnings-call disclosure or primary lab/vendor announcement. Two or more independent sources.
3 / 5	Single primary source. Reasonably consistent with sector data.
2 / 5	Analyst report or secondary press only. Or single primary source with credibility caveats.
1 / 5	Rumor, social media, single non-primary source. Or contradicted by alternate primary sources.

ANYTHING GRADED 2 OR BELOW IS FLAGGED AS NOISE.

PREDICTION OUTCOMES

HIT	The specific observable outcome occurred by the deadline.
MISS	The deadline passed and the outcome did not occur.
PARTIAL	A meaningfully similar outcome occurred but not the literal wording.
PENDING	Deadline not yet reached.

HIT RATE OF 65-75% IS THE TARGET. ABOVE 80% MEANS TOO CONSERVATIVE; BELOW 50% MEANS THE FRAMEWORK IS WRONG.

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

Past issues, the working framework, and the LLM Evolutionary Tree companion.