THE AI STACK WEEKLY ISSUE 11 · WEEK 27 OF 2026

THE BOTTOM LINE

The sell-side of AI compute gained two hyperscale balance sheets in 48 hours — just as the model layer started cutting prices.

Within two days, Bloomberg reported Meta is building "Meta Compute" to sell excess AI capacity (and possibly hosted Muse Spark access), knocking 12-15% off CoreWeave and Nebius in a morning, and SoftBank answered with SB Neo, a Delaware neocloud targeting 10GW by roughly 2030. The compute-rental market just gained two entrants with balance sheets larger than the entire listed neocloud category — exactly as the Epoch analysis showing hyperscaler free cash flow hitting zero around Q3 2026 became the consensus macro frame. The software lens ran the same economics in the other direction: Anthropic launched Claude Sonnet 5 at $2/$10 introductory pricing, and OpenAI's gated GPT-5.6 preview promises a Terra tier at half GPT-5.5's cost — the model layer is cutting prices into the very demand that is supposed to fill all this capacity. Meanwhile the frontier itself became a regulatory variable: GPT-5.6 previewed to only ~20 government-vetted organizations, while Claude Fable 5 returned globally on July 1 after export controls were withdrawn. Net/net: the industry's central question moved from "who has the best model" to "who can profitably fill capacity they have already financed" — boards should stress-test compute-vendor concentration, investors should separate scarce-input owners (memory, power) from capacity resellers, and architects should use the price war to renegotiate inference contracts now.

AUTHOR

Brian Letort
BrianLetort.AI

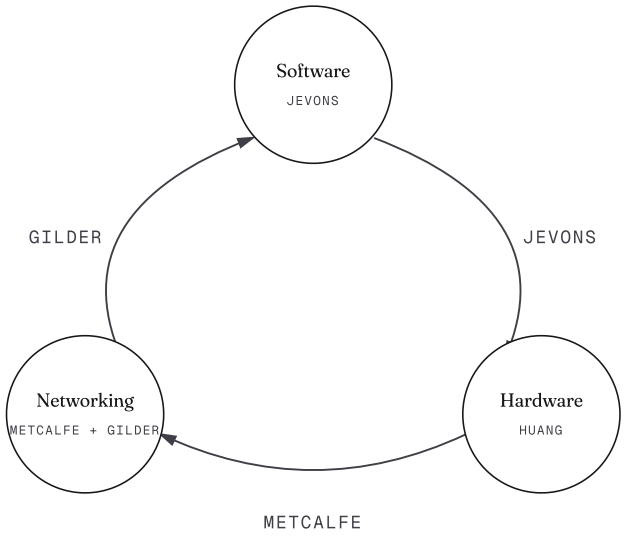
PUBLISHED

July 4, 2026
Issue 11

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.

JUN 30
EVENT 01

Anthropic launched Claude Sonnet 5 — its most agentic Sonnet, the new Free/Pro default — at $2/$10 per MTok introductory pricing through Aug 31 (then $3/$15)

ANTHROPIC NEWSROOM; TECHCRUNCH

JUN 30-JUL 1
EVENT 02

Claude Fable 5 was restored globally after export controls were withdrawn; Mythos 5 access returned for ~100 US critical-infrastructure organizations on the Commerce 'Annex A' list

ANTHROPIC NEWSROOM

JUN 26
EVENT 03

OpenAI previewed the GPT-5.6 family (Sol $5/$30, Terra $2.50/$15 at ~GPT-5.5 performance, Luna $1/$6) — limited to ~20 vetted partner organizations at the US government's request, broad GA 'in the coming weeks'

OPENAI; REUTERS SYNDICATION; THE BATCH

JUL 2
EVENT 04

GitHub shipped Copilot agent session streaming to public preview — enterprise admins can stream all agent session data (cloud agents, CLI, IDEs) to SIEM endpoints including Microsoft Purview

GITHUB CHANGELOG

JUN 30-JUL 2
EVENT 05

Gemini 3.5 Pro missed its implied June GA window; no public API model ID as of Jul 2, with reporting pointing to a July target while 3.5 Flash carries the load

TOKENMIX STATUS CHECK; BUSINESS INSIDER VIA PONDERO

WHAT THIS MEANS

The software story is a price-and-permission story: closed labs are cutting rates into open-weight cost pressure while Washington now sits inside the release pipeline itself. Architects should use Sonnet 5's intro pricing and the promised Terra tier to renegotiate inference contracts this quarter — and treat model availability as a regulatory risk to engineer around, with the Model Pulse carrying the full architecture read.

LENS 2 OF 3 THE FLYWHEEL

Hardware.

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

JUN 30
EVENT 01

SK hynix filed for a Nasdaq ADS listing seeking ~$29.4B (ticker SKHY, ~Jul 10 debut expected) — potentially the largest ADR offering in history, with proceeds earmarked for fab and EUV capex

SEC FILING VIA IBTIMES; YAHOO FINANCE

JUL 3
EVENT 02

Reports say SK hynix removed price caps from long-term memory supply agreements while Micron kept ceilings anchored to Q2 market prices — 2027 memory pricing power being contractually locked in

PCCENTRAL; CRYPTO BRIEFING (SECONDARY; NO PRIMARY DISCLOSURE YET)

JUN 28
EVENT 03

Korean industry reporting puts SK hynix at roughly two-thirds of NVIDIA's 2026 Vera Rubin HBM4 allocation, with Samsung expected to begin supplying NVIDIA in July after final quality tests

MAEIL BUSINESS

JUL 1
EVENT 04

Oxmiq Labs (Raja Koduri) raised $35M led by Fundomo and Samsung Catalyst Fund to license OxCore GPU IP — an 'Arm for AI silicon' play lowering the entry cost of custom accelerators

SILICONANGLE

WHAT THIS MEANS

Memory is completing its move from allocation story to capital-markets story: a ~$29.4B SEC-filed listing and reported cap-free long-term contracts mark the scarcity to market. Operators should lock HBM-backed capacity pricing before 2027 contracts reprice, and investors should watch the ~Jul 10 SKHY debut as the cleanest public-market referendum yet on the memory supercycle.

LENS 3 OF 3

THE FLYWHEEL

Networking.

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

JUN 30
EVENT 01

PJM opened the 2028/29 capacity auction (bids close Jul 7, results Jul 14 — a week later than previously expected) and advanced a proposal for a September-October backstop procurement against the anticipated capacity shortfall

PJM INSIDE LINES; AMERICAN PUBLIC POWER ASSOCIATION

JUL 9 / AUG 17
EVENT 02

FERC's large-load show-cause clock ticked into July: motions to intervene in all six RTO/ISO dockets are due Jul 9; tariff responses due Aug 17

HUSCH BLACKWELL REGULATORY UPDATE

JUN 24 (IN-WINDOW COVERAGE)
EVENT 03

CoreWeave's Stockholm expansion (Conapto colocation, 32MW live) pairs Blackwell and Vera Rubin with Quantum-X800 InfiniBand — a flagship Vera Rubin deployment shipping on InfiniBand, not Spectrum-X CPO Ethernet

COREWEAVE; BEBEEZ

WHAT THIS MEANS

No new optics event landed this week — the networking lens is power-process dominated, and the calendar now has three hard dates (Jul 9, Jul 14, Aug 17) that will reprice siting economics before any fabric decision does. Architects should note the CoreWeave data point: flagship Vera Rubin clusters can still ship on InfiniBand, so treat CPO Ethernet as an option to be priced, not an assumption.

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 187.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV ~1.3x

No new financing closed; the posture turned to price competition (Sonnet 5 at $2/$10, Terra promised at half GPT-5.5 cost) and IPO sequencing — OpenAI leaning 2027, Anthropic holding October 2026.

Hyperscaler-Hosted

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

BURN / REV ~3.0x

Meta Compute (reported Jul 1) and SoftBank's SB Neo (announced Jul 2) opened a new lane: hyperscale balance sheets selling excess compute — a monetization response to the free-cash-flow crossover now framed for ~Q3 2026.

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

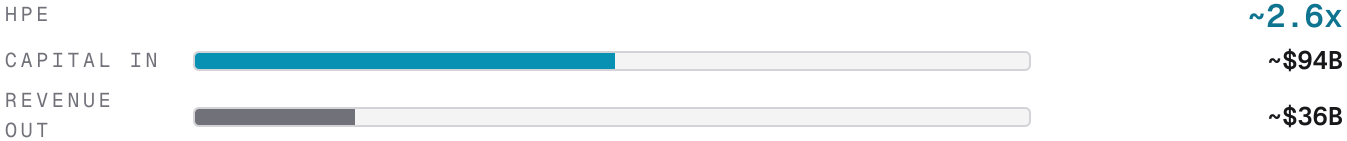
BURN / REV ~2.7x

Together AI closed $800M at $8.3B (the week's largest transaction) — but Meta Compute repriced the whole category, with CoreWeave and Nebius down 12-15% intraday and Crusoe reportedly in talks at ~$30B.

On-Prem / Hybrid

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

BURN / REV ~2.6x



No new drawn sovereign commitment; SB Neo's 10GW US pipeline (DOE-land Portsmouth, Ohio campus, 800MW first phase in 2028) is corporate but state-adjacent, and the FERC/PJM process calendar remains the binding variable.

SIGNAL VS NOISE

What's real, what's noise.

5 claims that drove headlines this week, scored 1–5 on source quality and triangulation. 2 flagged as noise. The bar is at least one explicit noise call per issue.

5/5

CLAIM 01

SK hynix's ~$29.4B Nasdaq ADS filing marks the memory supercycle to public markets.

SOURCES: SEC REGISTRATION STATEMENT; IBTIMES; YAHOO FINANCE

The cleanest audited-document signal of the week. An at-range or above-range pricing around Jul 10 validates memory scarcity with public-market money rather than analyst notes; a weak debut would be the first real crack in the thesis.

4/5

CLAIM 02

Claude Sonnet 5's $2/$10 launch is a closed-lab price response to open-weight and efficiency pressure.

SOURCES: ANTHROPIC ANNOUNCEMENT; TECHCRUNCH

The first flagship-adjacent price cut since the open-weight cost gap widened. Watch whether OpenAI's Terra GA (promised at half GPT-5.5 cost) confirms a repricing cycle rather than a one-off — buyers should time contract renegotiations to that confirmation.

3/5

CLAIM 03

Meta Compute repriced the neocloud category ~15% in a morning.

SOURCES: BLOOMBERG (ANON-SOURCED); CNBC CONFIRMATION; OBSERVABLE MARKET MOVES

The business plan is early and could change, but the market reaction is the durable fact: neocloud equity now embeds hyperscaler-competition and customer-concentration risk it did not price on Jun 30. Investors should treat the repricing, not the plan, as the signal.

1/5 — NOISE

CLAIM 04

Grok 4.5 performs 'close to or exceeding' Claude Opus, per xAI.

SOURCES: MUSK POSTS ON X; TECHTIMES SECONDARY COVERAGE

Noise until proven otherwise: private beta confined to SpaceX and Tesla, no public API, no independent benchmark, and a claimed monthly from-scratch-model cadence with no documented precedent. Procurement teams should not move on this claim.

2/5 — NOISE

CLAIM 05

Meta's 'Watermelon' Muse Spark update has internally caught GPT-5.5.

SOURCES: BUSINESS INSIDER (ANONYMOUS TOWNHALL SOURCING); WANG'S PUBLIC POST PROMISES ONLY 'BIG IMPROVEMENTS'

Internal parity claims from labs trailing the frontier have a poor verification record. The verifiable part is that a Muse Spark update with a new API is coming; wait for third-party evals before treating parity as real.

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

Meta Compute and SB Neo are best explained as monetization responses to the hyperscaler free-cash-flow crossover — meaning the compute glut is arriving as a financing problem before it arrives as a price problem.

ABDUCTIVE · 68% CONFIDENCE

The Epoch analysis — top-5 hyperscaler cash capex crossing aggregate operating cash flow around Q3 2026, on ~$725B of 2026 top-4 capex — became the week's consensus macro frame.
Bloomberg reported Jul 1 that Meta is building a cloud business to sell excess AI compute; Meta rose 10% while CoreWeave and Nebius fell 12-15% on customer-concentration risk.
SoftBank announced SB Neo on Jul 2 — a Delaware neocloud targeting 10GW by ~2030 — the second hyperscale balance sheet to enter the compute-rental market in 48 hours.

With the model layer cutting prices and the capacity layer adding sellers, margin is migrating to scarce physical inputs — memory and power — and the two hard-number tests of that migration land within ten days (SKHY debut ~Jul 10, PJM results Jul 14).

DEDUCTIVE · 71% CONFIDENCE

Anthropic cut Sonnet 5 to $2/$10 intro pricing and OpenAI promised Terra at half GPT-5.5's cost — the model layer is repricing downward.
Meta Compute and SB Neo add hyperscale sellers to compute rental, pressuring capacity margins from the supply side.
SK hynix filed a ~$29.4B Nasdaq listing and reportedly removed price caps from long-term memory contracts, while PJM capacity has cleared at the FERC cap two auctions running.

The US government has become a standing variable in the frontier release pipeline: availability of the best models is now set by regulatory process at least as much as by lab readiness.

INDUCTIVE · 66% CONFIDENCE

OpenAI's GPT-5.6 preview was limited to ~20 vetted partner organizations at the US government's request, with GA timing left open.
Claude Fable 5 was restored globally Jul 1 only after export controls were withdrawn; Mythos 5 remains gated to an ~100-organization 'Annex A' allowlist.
Anthropic's government-ID verification policy takes effect Jul 8 — the likely mechanism for staged, citizenship-verified frontier access.

THESIS TEST

H1 · SUPPORTED **The cycle is accelerating, not slowing.** The hypothesis predicts compressing cadences across the flywheel. This week's evidence is on the release-and-pricing side: three frontier availability events in five days (Sonnet 5 GA, GPT-5.6 preview, Fable 5 restoration) and a price structure — Sonnet 5 at $2/$10, Terra promised at half GPT-5.5 — that pulls next-generation economics forward before the prior generation has amortized. Price-performance is compounding faster than deployment cycles can absorb it.

H2 · SUPPORTED **Capital is concentrated, returns are diffuse.** The crossover frame sharpened the concentration side: ~$725B of 2026 top-4 capex is now consuming the spenders' free cash flow, forcing Meta and SoftBank to monetize excess capacity. The diffusion side got direct evidence too — model-layer price cuts hand value to buyers, and the neocloud selloff shows the market repricing whether concentrated capital earns its return at all. The spread between where capital pools and where returns land widened this week.

H3 · STRAINED **Networking is the durable layer.** No new optics or fabric event landed in-window, and the one relevant data point cuts against the strong version of the hypothesis: CoreWeave's flagship Stockholm Vera Rubin deployment ships on Quantum-X800 InfiniBand rather than the CPO Ethernet fabric NVIDIA has positioned as the AI-factory default. One deployment is not a refutation, but a framework that is never strained is not being tested — and this week strained it.

H4 · SUPPORTED **Open weights pull the floor up.** The mechanism the hypothesis describes — open-weight cost pressure forcing closed-lab repricing — visibly operated this week: Sonnet 5 launched at $2/$10 into a market where GLM-5.2 sits at ~1/6 frontier cost, and Bridgewater reported a tuned open Qwen3-235B beating its strongest frontier model on finance tasks at 13.8x lower cost. The closed price cuts are the floor pulling up, not the ceiling falling.

H5 · SUPPORTED **Power is the binding constraint for the next 24 months.** PJM opened the 2028/29 auction under an at-cap setup and simultaneously proposed a September-October backstop procurement against an anticipated capacity shortfall — the system operator itself is now planning for scarcity. Time-to-power held at 60-84 months, and the FERC docket calendar (Jul 9, Aug 17) is where siting economics get decided. Nothing this week suggested power is loosening before 2028.

PATTERN WATCH

Grid process — not hardware supply — is the recurring headline constraint on AI buildout. (3 WEEKS OBSERVED)

Next week: The PJM 2028/29 result on Jul 14 clears within 5% of the ~$325 cap (formalized as prediction p54). A materially lower print would break the pattern and reopen the case that power scarcity is regional, not structural.

HBM supply is tightening generation-over-generation into a seller's market. (3 WEEKS OBSERVED)

Next week: SK hynix's SKHY debut (~Jul 10) prices at or above range and its late-July earnings call confirms allocation exhaustion or 2027 price increases. A weak debut plus cautious earnings language would be the first counter-evidence.

US government action is now a standing gate on frontier-model availability. (3 WEEKS OBSERVED)

Next week: Anthropic's government-ID verification policy ships on schedule Jul 8, and GPT-5.6's broad GA timing is visibly set by the government process rather than OpenAI's product calendar. A clean, ungated GA within two weeks would weaken the pattern.

SECOND-ORDER EFFECTS

Neocloud cost of capital rises as equity markets embed hyperscaler-competition risk — marginal neocloud buildouts get costlier financing, slowing independent capacity additions even as total sell-side capacity grows. Enterprise buyers gain a stronger negotiating position on multi-year GPU contracts than at any point since 2024. Trigger: Meta Compute and SB Neo enter compute rental. Horizon: H1 2027. Who moves: Listed neoclouds and their lenders; enterprises negotiating multi-year compute contracts.

Agent-workload unit economics improve enough to pull forward enterprise automation deployments that were marginal at prior pricing — which generates the inference demand the new sell-side capacity needs. The price cut is simultaneously margin compression at the model layer and demand generation for the compute layer the same companies are financing. Trigger: Closed-lab price cuts (Sonnet 5 at $2/$10; Terra promised at half GPT-5.5 cost). Horizon: Q4 2026. Who moves: Enterprise AI budget owners, inference clouds, and SaaS vendors defending seat pricing.

Enterprise access to the best models becomes eligibility-dependent: organizations on government allowlists gain a structural capability edge, and multi-model routing shifts from a cost optimization to a continuity requirement. Procurement teams start writing regulatory-availability clauses into model contracts. Trigger: Government gating of frontier releases (GPT-5.6's vetted-partner preview; Annex A allowlists; ID verification from Jul 8). Horizon: Q4 2026. Who moves: Enterprises outside allowlists, model-routing platform vendors, and procurement/legal teams.

STRATEGIC OUTLOOK

If you allocate capital or run infrastructure, this week reframed the 12-month posture: the scarce assets are memory and power, the crowded assets are model capability and — for the first time — raw compute capacity. Lock inference pricing while the closed-lab price war runs (Sonnet 5 intro rates end Aug 31; Terra GA is the confirmation signal), but diversify compute-vendor exposure before hyperscaler sell-side capacity resets the market's structure — anchor-customer concentration just became the neocloud category's defining risk. Treat the Jul 10 SKHY debut and the Jul 14 PJM print as the two highest-information events of the month: together they will tell you whether the physical-scarcity trade is still underpriced. And build model-availability contingency into architecture reviews now — the government is in the release pipeline, and routing flexibility has become a continuity control, not a cost lever.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **0 rising**, **2 falling**, 8 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, top 3)	**~34-37 mo; OpenAI leaning 2027 IPO, Anthropic holding Oct 2026**	~34-37 mo (flat; no new in-window round)	→	<18 mo triggers re-rating risk
Hyperscaler capex / AI revenue ratio (top 4 weighted)	**~5.0-5.3; free-cash-flow crossover framed for ~Q3 2026**	~5.0-5.3 (flat; next top-4 earnings catalyst pending)	→	>6.0 invites investor pushback at next earnings
CoreWeave revenue backlog	**~$100B reported; Meta Compute repriced concentration risk (stock -12-15% Jul 1)**	~$100B reported / ~$131B analyst-estimated by end-Q2	→	Conversion velocity matters more than gross figure
NVIDIA Q-over-Q data center revenue	**$75.2B Q1 FY27; Q2 guide $91B, reports Aug 26**	$75.2B Q1 FY27; Q2 guide $91B, reports Aug 26	→	Q2 FY27 guide $91B implies further +21% QoQ

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Open vs closed gap on coding (SWE-Bench / agentic)	**Narrowing from both sides: closed prices down (Sonnet 5 $2/$10; Terra promised at half GPT-5.5), open pressure sustained**	Open pressure sustained: GLM-5.2 / DeepSeek V4 Pro remain the cost challengers	↓	Sustained open lead reshapes enterprise procurement
Sovereign AI commitments (count / aggregate $)	**~14 / ~$180B+ (flat; SB Neo's 10GW is corporate, not sovereign)**	~14 / ~$180B+ (flat; no new drawn sovereign mega-commitment)	→	—
PJM 2026/27 capacity auction price ($/MW-day)	**$329.17; 2028/29 BRA bids close Jul 7, results Jul 14 (slipped ~1 week)**	$329.17; 2028/29 BRA results expected around Jul 7	→	11x in 24 months — power is the new binding constraint
Time-to-power, busiest US markets (months)	**60-84; FERC intervenor deadline Jul 9, tariff responses due Aug 17**	60-84; FERC 60-day tariff-response clock is the next catalyst	→	—

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Cost-per-task, frontier reasoning model	**~$0.08-$0.13 effective; Sonnet 5 $2/$10 intro, GPT-5.6 tiering Sol $5/$30 / Terra $2.50/$15 / Luna $1/$6**	~$0.10-$0.15 effective; inference-cloud competition rising	↓	DeepSeek V4 official (mid-July) adds peak/off-peak surge pricing — the first major time-of-day price discrimination
Custom silicon share of incremental AI compute	**~33-36%; silicon-IP layer forming (Oxmiq) to lower custom-ASIC entry cost**	~33-36%; HBM and CPO now more binding than raw accelerator demand	→	>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 CAPITAL

SK hynix's Nasdaq ADS offering prices at or above its indicated ~$166/ADS level and closes its first trading week above the offer price, by July 31, 2026.

67%

BY JULY 31, 2026

TRIGGER: SKHY pricing announcement and first-week Nasdaq trading data (~Jul 10 debut expected).

PREDICTION 02 POWER

The PJM 2028/29 base residual auction clears within 5% of the ~$325/MW-day cap when results post on July 14, 2026.

74%

BY JULY 14, 2026

TRIGGER: PJM BRA results publication (after 4 p.m. ET, Jul 14).

PREDICTION 03 SOFTWARE

GPT-5.6 reaches broad GA with the Terra tier priced at or below $2.50/$15 per MTok — half of GPT-5.5's rate — confirming a closed-lab repricing cycle rather than a one-off Sonnet 5 cut, by August 31, 2026.

62%

BY AUGUST 31, 2026

TRIGGER: OpenAI pricing page / API changelog at GPT-5.6 general availability.

PREDICTION 04 HARDWARE

Samsung's HBM4 supply to NVIDIA is publicly confirmed — via earnings call, company statement, or multi-source supply-chain reporting — by August 31, 2026.

66%

BY AUGUST 31, 2026

TRIGGER: Samsung Q2 earnings call (late July), NVIDIA disclosure, or corroborated supply-chain reporting.

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 7-9
WATCH 01

PJM bids close (Jul 7), Fable 5 usage-credit pricing begins (Jul 7), FERC intervenor deadline (Jul 9)

Three hard dates in 72 hours: the auction sets up the Jul 14 print, Fable 5's credit transition is a live test of post-restoration Claude demand, and the FERC intervenor list will show who intends to fight over large-load tariffs.

~JUL 10
WATCH 02

SK hynix SKHY ADS begins trading on Nasdaq

Pricing versus the ~$166/ADS, ~$29.4B expectation is a public-market referendum on the memory supercycle — the week's cleanest test of the scarcity thesis with real money.

JUL 14
WATCH 03

PJM 2028/29 capacity auction results (after 4 p.m. ET)

An at-cap clearing confirms power as the binding AI-factory constraint through 2028; a materially sub-cap print would be the first crack in the pattern and would reprice siting strategy.

MID-LATE JUL
WATCH 04

GPT-5.6 broad GA, DeepSeek V4 official (with peak/off-peak pricing), Gemini 3.5 Pro July target

Three release events that will decide whether the Sonnet 5 price cut was a one-off or the start of a closed-lab repricing cycle — the single biggest input to H2 inference budgeting.

JUL 28
WATCH 05

Alphabet Q2 earnings — first top-4 hyperscaler print of the season

First test of the free-cash-flow-crossover narrative against actuals; capex guidance and AI-revenue disclosure will recalibrate the capex-to-revenue lever for the whole category.

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 11

Week 27 of 2026 · July 4, 2026

WEB

brianletort.ai/industry

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