THE AI STACK WEEKLY ISSUE 12 · WEEK 28 OF 2026

THE BOTTOM LINE

# The harness became the product: OpenAI and Anthropic shipped rival work runtimes 48 hours apart — while memory and power marked the scarcity trade to market with $26.5B of real money.

Two launches this week redrew the enterprise AI procurement map. Anthropic pushed Claude Cowork to web and mobile on July 7 with cloud-run background sessions, citing 1.2 million sessions across 600,000+ organizations showing most Cowork use is non-coding knowledge work. OpenAI answered on July 9 with ChatGPT Work — the Codex task runtime generalized to all knowledge work, bundled into a desktop app available on every plan including Free, with 1 million of Codex's 5 million weekly users already working outside software development. The same week, two independent quantitative studies explained why the harness is where the fight moved: Databricks' merged-PR benchmark found the same model at the same effort costs over 2x more per task depending on harness choice — with open-weight GLM 5.2 statistically tied with Opus 4.8 at $1.28 vs $1.94 per task — and LangChain/NVIDIA showed harness tuning alone lifts an open model to near-Opus quality at roughly 10x lower cost. The model layer kept commoditizing on cue: GPT-5.6 went GA with Terra at $2.50/$15 (half of GPT-5.5's rate, resolving prediction p55 as a hit), and Grok 4.5 launched at $2/$6 positioning explicitly on cost-per-task. Meanwhile the physical layer banked the other side of the trade: SK hynix closed the largest-ever foreign US IPO at $26.5B with 7x demand and a +13% first-day pop, Samsung guided to a record ~KRW 89.4T quarter on AI memory, and Meta committed C$13B to a 1 GW Alberta campus where it must fund its own gas generation because the grid cannot host multiple large AI loads. Net/net: capability is commoditizing, harnesses are consolidating into suites, and margin keeps pooling in memory and

power. Boards should treat agent-suite governance (defaults, budgets, audit) as this quarter's control gap; investors should note the scarcity trade just got public-market confirmation; architects should re-run agent cost benchmarks at the harness level, not the rate card.

---

AUTHOR

**Brian Letort**
BrianLetort.AI

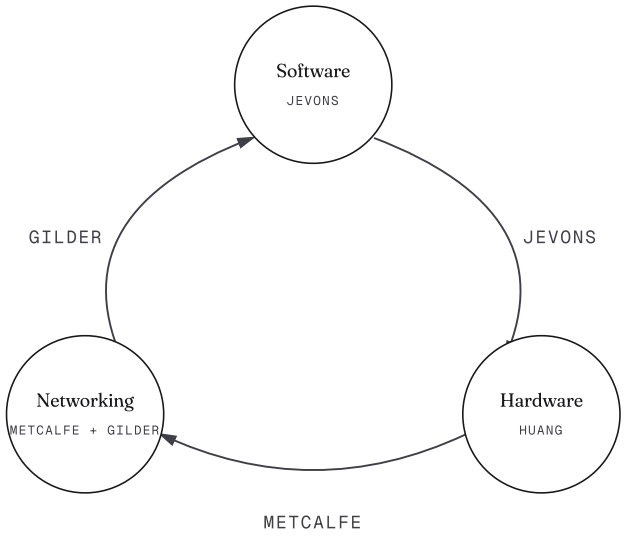
PUBLISHED

July 11, 2026
Issue 12

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 9** EVENT 01

GPT-5.6 reached full GA as a three-tier family — Sol $5/$30, Terra $2.50/$15, Luna $1/$6 per MTok — after a 12-day government-coordinated preview; GPT-5.4 retires Jul 23

OPENAI; VELLUM BENCHMARK ANALYSIS

**JUL 8** EVENT 02

xAI released Grok 4.5, an 'Opus-class' workhorse at $2/$6 per MTok co-trained with Cursor — 4th on the AA Intelligence Index, claiming ~4.2x fewer output tokens per SWE-Bench Pro task; not available in the EU at launch

XAI; TECHCRUNCH

**JUL 6-9** EVENT 03

Tencent released Hy3 (295B MoE, 21B active) under clean Apache 2.0 at ~$0.20/$0.80 per MTok — and the community made it locally deployable in ~30 hours, with llama.cpp MTP speculative decoding measuring +40% throughput

HUGGING FACE MODEL CARD; SIMON WILLISON; LLAMA.CPP PR #25395

**JUL 9** EVENT 04

Meta released Muse Spark 1.1 and opened the Meta Model API in public preview — the first frontier Meta model distributed through a first-party API rather than open weights or Meta's own apps

META AI

**JUL 8** EVENT 05

Gemini 3.5 Pro slipped again to a reported July 17 target after Google rebuilt the base model; the public API still lists no 3.5 Pro model ID

TECHTIMES; GOOGLE GEMINI API MODEL LIST

**WHAT THIS MEANS**

Four frontier-relevant releases in four days, and every closed launch priced against the open floor: Terra at half GPT-5.5's rate and Grok 4.5 at $2/$6 are responses to Apache-2.0 Hy3 economics and GLM 5.2's per-task parity. Architects should lock inference pricing during this window and re-benchmark at the task level — see the Model Pulse for the full architecture read, including why the harness now matters more than the rate card.

LENS 2 OF 3 THE FLYWHEEL

# Hardware.

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

**JUL 7**
EVENT 01

Samsung guided to a record ~KRW 89.4T Q2 operating profit (+1,810% YoY) on AI memory — reportedly the largest quarterly operating profit ever posted by a tech company, with HBM4 reaching $1B in sales within four months

SAMSUNG NEWSROOM; SEOUL ECONOMIC DAILY; KOREA HERALD

**JUL 10**
EVENT 02

SK hynix closed the largest-ever foreign US IPO at $26.5B — priced at $149/ADS (below the ~$166 indication) but 7x oversubscribed, closing day one up ~13%; proceeds fund the Yongin fab, HBM packaging, and EUV tools

TECHCRUNCH; KOREA HERALD; YAHOO FINANCE

**JUL 6**
EVENT 03

SemiAnalysis reported NVIDIA's ~600 kW Kyber rack for Rubin Ultra slipping to 2028 on PCB-midplane manufacturability; NVIDIA publicly denied it the same day, holding to Kyber racks in H2 2027

CNBC; WCCFTECH (NVIDIA STATEMENT)

**JUL 9**
EVENT 04

Micron raised planned US investment to $250B+ through 2035 and committed up to $3B to the domestic supply chain, including $500M in GlobalWafers' Texas 300mm plant — with its HBM sold out for 2026 and able to fill only 50-66% of demand

MICRON PRESS RELEASES; TOM'S HARDWARE

**JUL 9**
EVENT 05

Reuters: Meta's Broadcom-designed 'Iris' MTIA chip enters production in September, on a roadmap of one new chip every ~6 months as Meta targets doubling compute to 14 GW in 2027

REUTERS; TECHCRUNCH; DCD

**WHAT THIS MEANS**

Memory completed its move from allocation story to capital-markets story — a record quarter, a $26.5B IPO with 7x demand, and TrendForce showing long-term agreements now capping price increases means the scarcity is being contractually locked, not loosening. Operators should treat the SemiAnalysis-vs-NVIDIA Kyber dispute as a live facility-planning risk (600 kW racks slipping would reshape 2027-28 datacenter designs) and watch Samsung's Jul 30 divisional print to make the HBM4-to-NVIDIA confirmation unambiguous.

LENS 3 OF 3 THE FLYWHEEL

# Networking.

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

**JUL 9** EVENT 01

DriveNets and WhiteFiber deployed the first commercial long-distance scale-across AI supercluster — two H200 sites 83 km apart validated at 111.2 Tbps with 0.9 ms guaranteed latency, within 8% of the physical limit of light in fiber

DRIVENETS; WHITEFIBER; LIGHT READING

**JUL 6** EVENT 02

IDC Q1 2026 data: datacenter Ethernet switch revenue up 61% YoY to ~$10B against 3% server growth, with 800G sales up 10.3x to $3.58B — and NVIDIA now the top datacenter Ethernet vendor at $2.1B, ahead of Arista and Cisco

THE NEXT PLATFORM (IDC); TECHREPUBLIC (DELL'ORO)

**JUL 8** EVENT 03

Marvell published Keysight-validated Ultra Ethernet results (packet trimming, Auto Load Balancing, UET) on Teralynx switches — the most concrete public UET-on-silicon data point ahead of UEC-native NIC availability

MARVELL BLOG

**JUL 9** EVENT 04

IEEE Spectrum: NVLink Fusion's photonics partners (Ayar Labs, Lightmatter, Marvell) signal optics moving into the scale-up domain as rack GPU density heads from 72 toward as many as 576 by 2027

IEEE SPECTRUM; AYAR LABS; LIGHTMATTER

**WHAT THIS MEANS**

After W27 strained the networking hypothesis, this week reinstated it with the strongest possible evidence: a commercial fabric product whose entire value proposition is monetizing the power constraint — stitching power-limited sites into one logical cluster (Metcalfe compounding Gilder). Architects planning 2027 capacity should now price scale-across fabric as a real alternative to waiting on single-site interconnection queues, and watch the 800G-to-1.6T ramp confirmed by 10.3x growth.

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 primary capital; the froth moved to secondaries — Anthropic shares traded at an implied $1.2T (up 550% in a year, broker-reported), overtaking OpenAI's ~$908B — even as both labs kept cutting prices (Terra GA at half GPT-5.5; Grok 4.5 at $2/$6).

## Hyperscaler-Hosted

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

BURN / REV ~3.0x


Meta committed C$13B (~$9.1B) to a 1 GW Alberta campus — its largest outside the US — fully funding its own 932 MW gas tolling deal because the grid cannot host multiple large AI loads; Microsoft cut 4,800 roles while funding its $2.5B Frontier AI-deployment unit.

## Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

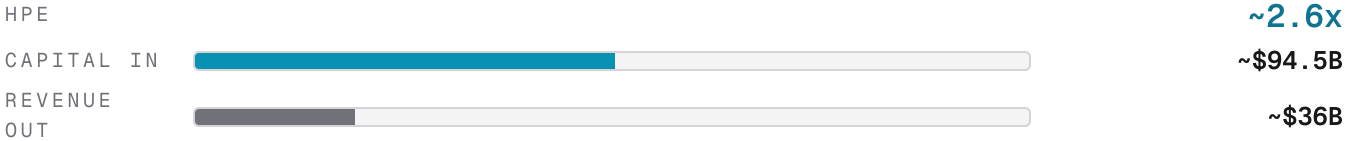
BURN / REV ~2.7x


Delivery, not fundraising: Galaxy Digital completed Helios Phase I on schedule — 133 MW of critical IT load to CoreWeave under a 15-year lease with payments already flowing, against 526 MW committed across three phases and projected revenue above $1B a year.

## On-Prem / Hybrid

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

BURN / REV ~2.6x




MARA acquired a 2 GW powered-land site in Matagorda County, Texas for up to $600M (8-K filed) — up to 1 GW of grid capacity by October 2027 — taking its potential portfolio to ~4.8 GW; Beijing pushed the Manus buyback toward a Tencent-led consortium at no less than $2B.

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

### GPT-5.6 reached GA with Terra priced at $2.50/$15 per MTok — exactly half of GPT-5.5 — confirming the closed-lab repricing cycle.

SOURCES: OPENAI GA ANNOUNCEMENT AND PRICING PAGE; VELLUM BENCHMARK ANALYSIS

Prediction p55 resolves as a hit seven weeks early. The repricing cycle is now confirmed from three vendors (Sonnet 5, Terra, Grok 4.5) — buyers should renegotiate inference contracts this month, before Sonnet 5's intro pricing lapses Aug 31 anchors the new floor.

4/5

CLAIM 02

### Harness choice now swings agent cost more than model choice — over 2x per task at equal quality per Databricks, ~10x via tuned harness profiles per LangChain/NVIDIA.

SOURCES: DATABRICKS ENGINEERING (MERGED-PR BENCHMARK); LANGCHAIN AND NVIDIA BLOGS

Two independent, methodologically serious studies published the same week, one on a real multi-million-line codebase. This is the strongest procurement-relevant finding of the month: benchmark the harness, not just the model, and treat per-token rate cards as a poor proxy for cost.

3/5

CLAIM 03

### NVIDIA's ~600 kW Kyber rack for Rubin Ultra has slipped to 2028 on PCB-midplane manufacturability.

SOURCES: SEMIANALYSIS (PAYWALLED REPORT); NVIDIA PUBLIC DENIAL VIA CNBC/WCCFTECH

A credible specialist source against an explicit vendor denial — unresolvable this week. The facility-planning implication is real either way: anyone designing 2027-28 halls around 600 kW/rack and 800V DC should hold a 190-230 kW contingency. TSMC's Jul 16 earnings commentary on advanced packaging may arbitrate.

2/5 — NOISE

CLAIM 04

### Anthropic is now 'worth' $1.2 trillion, overtaking OpenAI.

SOURCES: BUSINESS INSIDER VIA SECONDARY OUTLETS; BROKER-REPORTED SECONDARY-MARKET PRINTS

Secondary trades on illiquid, scarce shares are sentiment, not valuation — the primary mark remains the $965B Series H. The durable fact is directional demand (reportedly ~5 buyers per 2 for OpenAI); wait for the IPO range to treat any trillion-dollar figure as real.

1/5 – NOISE

CLAIM 05

### GPT-5.6 Sol Ultra proved the 50-year-old Cycle Double Cover Conjecture in under an hour.

SOURCES: OPENAI PROOF PDF AND PROMPT RELEASE; NO INDEPENDENT VERIFICATION

Noise until verified: no independent mathematical review, no Lean/Coq formalization, and the released prompt instructed the model to assume a proof exists — a setup that invites confident invalid arguments. If formal verification lands, this becomes the AI-research event of the quarter; until then, do not cite it.

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

### The enterprise procurement unit shifted from the model to the harness this week: OpenAI and Anthropic shipped competing general-work runtimes 48 hours apart, and two independent quantitative studies showed the harness now swings agent economics more than the model does.

ABDUCTIVE · 74% CONFIDENCE

Anthropic pushed Claude Cowork to web and mobile with cloud-run background sessions on Jul 7, citing 1.2M sessions across 600,000+ organizations showing most Cowork use is non-coding knowledge work.
OpenAI answered on Jul 9 with ChatGPT Work — the Codex task runtime generalized to all knowledge work — bundled into a desktop app available on every plan including Free, explicitly citing 1M+ of Codex's 5M weekly users working outside software.
Databricks' merged-PR benchmark found the same model at the same effort costs over 2x more per task in one harness than another at equal quality, and LangChain/NVIDIA showed harness tuning alone lifts an open model to near-Opus quality at roughly 10x lower cost.

### The physical-scarcity trade W27 framed got marked to market with real money in a single week — public-market, corporate-treasury, and hyperscaler capital all paid up for memory and power at once, confirming that margin is pooling in scarce physical inputs while the model layer price-wars.

DEDUCTIVE · 78% CONFIDENCE

SK hynix closed the largest-ever foreign US IPO at $26.5B with demand reported at 7x available shares and a first-day close up ~13%, while Samsung guided to a record ~KRW 89.4T quarter on AI memory with HBM4 hitting $1B in sales within four months.
Meta committed C$13B to a 1 GW Alberta campus and — because the grid cannot host multiple large AI loads — is funding its own 932 MW gas tolling deal plus a 250 MW supply agreement, while MARA paid up to $600M (8-K filed) for 2 GW of powered land in Texas.
The same week, the model layer kept cutting prices: GPT-5.6 GA'd with Terra at $2.50/$15 (half of GPT-5.5), and Grok 4.5 launched at $2/$6 positioning explicitly on cost-per-task.

### The open-weight floor is rising through deployment economics, not just benchmarks: a permissively licensed near-frontier model went from release to local-hardware viability in about 30 hours, and enterprise-grade evidence now shows open models at frontier task quality for a third of the cost.

INDUCTIVE · 71% CONFIDENCE

Tencent released Hy3 (295B MoE, 21B active) under a clean Apache 2.0 license on Jul 6; community GGUF quants with 1M context landed within ~30 hours, and a llama.cpp pull request using the model's MTP layer for speculative decoding measured +40% local throughput.
Databricks' internal benchmark put open-weight GLM 5.2 statistically tied with Claude Opus 4.8 on quality at $1.28 vs $1.94 per task, and the colibri project demonstrated the 744B GLM-5.2 running in 25GB of consumer RAM by streaming experts from disk.
Distribution caught up the same week: Hugging Face's curated open-weight collection landed in Microsoft Foundry with one-click managed-compute deployment, CVE-scanned runtimes, and per-deployment billing.

**Frontier availability is now gated by governments on both sides of the Pacific: GPT-5.6 reached GA only after a government-coordinated preview with unpublished evaluation criteria, while Beijing simultaneously rationed NVIDIA H200 access and forced the Manus ownership unwind toward a Tencent-led consortium.**

INDUCTIVE · 66% CONFIDENCE

GPT-5.6 went GA on Jul 9 after a 12-day government-coordinated restricted preview; the system card discloses that Commerce's CAISI ran pre-deployment evaluations whose criteria remain unpublished, and all three tiers are rated High in bio/chem and cyber.

The Information and Reuters reported Beijing preparing to let Alibaba, ByteDance, and DeepSeek buy H200s — but capped below 200,000 units, less than half of what was requested, after months of withheld approvals favoring domestic silicon.

Tencent entered talks to lead a consortium buying Manus back from Meta at no less than $2B — the direct consequence of Beijing ordering Meta's acquisition unwound in April.

## THESIS TEST

**H1 · SUPPORTED** **The cycle is accelerating, not slowing.** Four frontier-relevant releases landed in four days — GPT-5.6 GA (Jul 9), Grok 4.5 (Jul 8), Muse Spark 1.1 with Meta's first-ever model API (Jul 9), and Tencent's Apache-2.0 Hy3 (Jul 6) — and OpenAI set GPT-5.4's retirement for Jul 23, a deprecation cadence measured in months. The community-to-local pipeline compressed too: Hy3 went from release to quantized local deployment in ~30 hours, the fastest such cycle recorded for a 295B-class model. Cadence is compressing at both the frontier and the floor simultaneously.

**H2 · SUPPORTED** **Capital is concentrated, returns are diffuse.** The spread widened visibly this week: Anthropic traded at an implied $1.2T on secondary markets (up 550% in a year, broker-reported) while the labs kept cutting prices into that valuation — Terra at half GPT-5.5's rate, Grok 4.5 at $2/$6. Meanwhile the clearest realized returns landed at the physical layer (Samsung's record ~KRW 89.4T quarter, SK hynix's 7x-oversubscribed IPO) and in labor substitution (Microsoft cutting 4,800 roles while funding a $2.5B AI-deployment unit). Capital pools at the model layer; this week's cash profits showed up in memory, power, and restructured cost bases.

**H3 · SUPPORTED** **Networking is the durable layer.** After W27 strained the hypothesis, this week delivered its strongest evidence in a month: DriveNets and WhiteFiber deployed the first commercial long-distance scale-across AI supercluster (111.2 Tbps over 83 km at 0.9 ms, within 8% of the physical limit of light in fiber), explicitly framed as an escape from single-site power constraints — networking directly monetizing the power bottleneck. The market data agrees: datacenter Ethernet switch revenue grew 61% YoY against 3% server growth, with 800G sales up 10.3x, and NVLink Fusion's photonics partners signal optics entering the scale-up domain as racks head toward 576 GPUs.

**H4 · SUPPORTED** **Open weights pull the floor up.** The mechanism operated end-to-end this week: Tencent shipped Hy3 under clean Apache 2.0 at ~$0.20/$0.80 per MTok, the community made it locally deployable within 30 hours (with +40% throughput from MTP speculative decoding), Databricks published enterprise evidence that open GLM 5.2 statistically ties Opus 4.8 at two-thirds the per-task cost, and Microsoft Foundry began one-click distribution of curated open weights into enterprise Azure estates. The floor is rising on quality, cost, deployability, and distribution simultaneously — and closed-lab pricing (Terra at half GPT-5.5, Grok 4.5 at $2/$6) is visibly responding.

**H5 · SUPPORTED** **Power is the binding constraint for the next 24 months.** Meta's Alberta announcement is the cleanest single confirmation yet: the company is fully funding its own generation (a 932 MW gas tolling deal plus a 250 MW supply agreement) because the grid explicitly cannot support multiple large AI loads — the hyperscaler is becoming its own utility. MARA paid up to $600M for powered land whose value is entirely its 2 GW of secured capacity, Galaxy's 133 MW delivery to CoreWeave started a >$1B/year revenue stream, and the week's flagship networking deployment exists specifically to stitch power-constrained sites into one logical cluster. PJM's 2028/29 results (Jul 14) are the next hard test.

## PATTERN WATCH

**Government action is a standing gate on frontier-model availability — and the gate is now bilateral.** (4 WEEKS OBSERVED)

Next week: Gemini 3.5 Pro's reported Jul 17 GA is the test: if it ships without a government-coordinated preview phase, the US gate is OpenAI/Anthropic-specific rather than industry-standard; if it gets the same treatment, pre-release government review has become the de facto US frontier release process.

**Agent economics are decoupling from model price lists — first tokens-per-task, now the harness itself.** (3 WEEKS OBSERVED)

Next week: Within two weeks another major platform (GitHub, Cursor, or a model vendor) publishes per-task or per-harness cost telemetry, or ships harness-level cost controls — confirming harness engineering as the new cost-optimization layer. If instead pricing discussion stays at per-token rate cards, the pattern is ahead of the market.

**Physical-input scarcity (memory, power) keeps marking itself to market with progressively harder money.** (4 WEEKS OBSERVED)

Next week: PJM's 2028/29 base residual auction results (Jul 14, after 4 p.m. ET) clear within 5% of the ~$325/MW-day cap, per prediction p54. A materially sub-cap print would be the first hard counter-evidence to the power-scarcity leg of this pattern.

## SECOND-ORDER EFFECTS

**Agent-harness distribution collapses into the subscription suites, squeezing standalone agent startups on distribution rather than capability — and enterprise governance becomes the real negotiation surface, since OpenAI's own rollout ships Work off-by-default with a two-week admin preview. Procurement teams that treated agents as a tool category must now treat them as a suite default that arrives enabled unless someone opts out.** Trigger: OpenAI bundles ChatGPT Work (with Codex) into a desktop app available on every plan including Free, while Anthropic ships Cowork to web and mobile. Horizon: Q4 2026. Who moves: Standalone agent-product startups, CIOs and IT admins managing suite defaults, and vertical-AI vendors whose wedge was 'the agent' rather than the workflow.

**Enterprises re-run agent cost benchmarks and discover open-model-plus-tuned-harness parity, shifting spend from closed-model API contracts toward serving infrastructure and harness engineering as a discipline. Closed labs respond by bundling harness and model more tightly (exactly what ChatGPT Work does), making the harness a lock-in layer just as the model layer commoditizes.** Trigger: Databricks and LangChain/NVIDIA publish hard evidence that harness choice swings per-task cost 2-10x at equal quality. Horizon: H2 2026. Who moves: Enterprise AI platform teams, closed-lab API revenue, open-model serving providers, and anyone budgeting agent fleets on per-token rate cards.

**Self-funded, behind-the-meter generation becomes the hyperscaler template for new campuses, which drains the most credit-worthy anchor loads out of utility interconnection queues — leaving regulated grid processes (and the FERC/PJM docket calendar) to govern everyone else. Non-hyperscale buyers inherit longer queues and higher capacity prices while hyperscalers effectively secede from the constraint.** Trigger: Meta fully funds dedicated generation for its Alberta campus because the grid cannot host multiple large AI loads. Horizon: 2027-2028. Who moves: Utilities and grid operators, colocation and enterprise data-center developers without generation balance sheets, and state energy regulators.

### STRATEGIC OUTLOOK

This week hardened the 12-month posture on both ends of the stack. At the physical layer, the scarcity trade is no longer a thesis — it is a closed $26.5B IPO with 7x demand, a record memory quarter, and hyperscalers buying their own power plants; treat memory and secured power as strategic inventory, and expect the Jul 14 PJM print to confirm at-cap capacity pricing through 2028. At the work layer, the unit of enterprise AI procurement just shifted from the model to the harness: ChatGPT Work and Claude Cowork will land inside your organization through suite defaults, not RFPs, so governance capacity — admin opt-outs, budget caps, audit streaming — is the binding internal constraint to build now. And run the open-weight math again: with Apache-2.0 Hy3 deployable locally in 30 hours, GLM 5.2 tying Opus quality at two-thirds the per-task cost, and harness tuning worth more than model choice, the cost floor for capable agents is falling faster than closed-lab price cuts — which is precisely why the closed labs are racing to own the harness instead.

EARLY WARNING PANEL

# The levers we monitor.

10 metrics, current vs prior period. **1 rising**, **2 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, top 3) | **~34-37 mo; Anthropic at implied $1.2T on secondaries (broker-reported), IPO calendars unchanged** | ~34-37 mo; OpenAI leaning 2027 IPO, Anthropic holding Oct 2026 | → | <18 mo triggers re-rating risk |
| Hyperscaler capex / AI revenue ratio (top 4 weighted) | **~5.0-5.3; Meta targets 14 GW of compute in 2027 (2x 2026) on $125-145B capex guidance** | ~5.0-5.3; free-cash-flow crossover framed for ~Q3 2026 | → | >6.0 invites investor pushback at next earnings |
| CoreWeave revenue backlog | **~$100B reported; Helios Phase I (133 MW) delivered on schedule — backlog now converting to lease revenue** | ~$100B reported; Meta Compute repriced concentration risk (stock -12-15% Jul 1) | → | Conversion velocity matters more than gross figure |

| METRIC | CURRENT | PRIOR | DIR | THRESHOLD / NOTE |
|---|---|---|---|---|
| NVIDIA Q-over-Q data center revenue | **$75.2B Q1 FY27; Q2 guide $91B (reports Aug 26); SemiAnalysis sees H2 ~20% above consensus despite Kyber dispute** | $75.2B Q1 FY27; Q2 guide $91B, reports Aug 26 | → | Q2 FY27 guide $91B implies further +21% QoQ |
| Open vs closed gap on coding (SWE-Bench / agentic) | **Effectively closed on cost-quality: GLM 5.2 statistically tied with Opus 4.8 at $1.28 vs $1.94/task (Databricks); Hy3 adds Apache-2.0 agentic-search lead** | Narrowing from both sides: closed prices down (Sonnet 5 $2/$10; Terra promised at half GPT-5.5), open pressure sustained | ↓ | Sustained open lead reshapes enterprise procurement |
| Sovereign AI commitments (count / aggregate $) | **~14 / ~$180B+ (flat; Meta Alberta and MARA Texas are corporate capital on power-rich land, not sovereign programs)** | ~14 / ~$180B+ (flat; SB Neo's 10GW is corporate, not sovereign) | → | — |

| METRIC | CURRENT | PRIOR | DIR | THRESHOLD / NOTE |
|---|---|---|---|---|
| PJM 2026/27 capacity auction price ($/MW-day) | **$329.17; 2028/29 BRA bids closed Jul 7 — results post Jul 14 after 4 p.m. ET (prediction p54 resolves)** | $329.17; 2028/29 BRA bids close Jul 7, results Jul 14 (slipped ~1 week) | → | 11x in 24 months — power is the new binding constraint |
| Time-to-power, busiest US markets (months) | **60-84; hyperscalers routing around queues — Meta fully funds its own generation in Alberta because the grid cannot host multiple large loads** | 60-84; FERC intervenor deadline Jul 9, tariff responses due Aug 17 | → | — |
| Cost-per-task, frontier reasoning model | **~$0.06-$0.12 effective; GPT-5.6 GA tiering (Sol $5/$30 / Terra $2.50/$15 / Luna $1/$6), Grok 4.5 at $2/$6 — and harness choice swings per-task cost 2x+** | ~$0.08-$0.13 effective; Sonnet 5 $2/$10 intro, GPT-5.6 tiering Sol $5/$30 / Terra $2.50/$15 / Luna $1/$6 | ↓ | Databricks: per-token price is a poor proxy — Sonnet 5 costs more per task than Opus 4.8 despite a ~1.7x cheaper rate card |

| METRIC | CURRENT | PRIOR | DIR | THRESHOLD / NOTE |
|---|---|---|---|---|
| Custom silicon share of incremental AI compute | **~34-37%; Meta's Iris enters production in September, Broadcom-Apple extended through 2031 (8-K), AWS raising Trainium 3 orders 20-30%** | ~33-36%; silicon-IP layer forming (Oxmiq) to lower custom-ASIC entry cost | ↑ | >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

**58%**

BY JULY 31, 2026

**Gemini 3.5 Pro reaches public general availability — a callable API model ID with published pricing — by July 31, 2026, after slipping past its June window and the reported July 17 target.**

TRIGGER: Google Gemini API model list / pricing page showing a GA gemini-3.5-pro model ID.

PREDICTION 02 SOFTWARE

**64%**

BY AUGUST 31, 2026

**At least one major agent platform (OpenAI, Anthropic, GitHub, or Cursor) ships product-level per-task or per-harness cost telemetry or routing controls — beyond session budget caps — by August 31, 2026.**

TRIGGER: Product changelog or GA announcement exposing per-task cost measurement or harness-level cost controls.

PREDICTION 03 HARDWARE

**62%**

BY JULY 16, 2026

**TSMC's July 16 Q2 earnings raise or reiterate the top end of full-year 2026 capex guidance and report HPC/AI platform revenue up more than 50% year over year, confirming the packaging-constrained AI capex ramp.**

TRIGGER: TSMC Q2 2026 earnings release and investor call (Jul 16; June revenue print Jul 13, typhoon-delayed).

PREDICTION 04 NETWORKING

**61%**

BY SEPTEMBER 30, 2026

**A second named vendor or operator announces a commercial cross-data-center scale-across AI fabric deployment or product launch — following DriveNets/WhiteFiber — by September 30, 2026.**

TRIGGER: Vendor or operator press release for a commercial (not lab) multi-site training-fabric deployment; WhiteFiber's own Q3 commercial launch also qualifies if it lands with a named second customer.

WATCHLIST

# On the radar.

6 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 13-16
WATCH 01

## TSMC June revenue (Jul 13, typhoon-delayed) and Q2 earnings (Jul 16)

The first hard AI-capex read of the season — capex guidance, HPC platform growth, and advanced-packaging commentary will also arbitrate the SemiAnalysis-vs-NVIDIA Kyber dispute (prediction p59).

JUL 14
WATCH 02

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

Prediction p54 resolves: an at-cap clearing confirms power as the binding constraint through 2028; a materially sub-cap print would be the first crack in the pattern and would reprice siting strategy.

JUL 17
WATCH 03

## Gemini 3.5 Pro reported GA target

Two slips already (June, then Jul 17); prediction p57 tracks GA by Jul 31. Also the test of the government-preview pattern: if Google GAs without a CAISI-coordinated phase, pre-release review is an OpenAI/Anthropic-specific regime, not an industry standard.

JUL 23-24
WATCH 04

## GPT-5.4 retirement (Jul 23) and DeepSeek legacy alias shutdown (Jul 24)

Two hard migration deadlines a day apart — the closed and open ecosystems now deprecate at the same aggressive cadence, and both will surface integration debt in agent fleets built on pinned model IDs.

WEEK OF JUL 21
WATCH 05

## Q2 earnings wave opens: ServiceNow (Jul 22), then Alphabet / Microsoft / SAP (dates aggregator-estimated)

First top-4 hyperscaler prints of the season test the free-cash-flow-crossover narrative against actuals; Microsoft's report is the one to watch for Copilot revenue disclosure and Frontier-unit framing after the 4,800-role cut.

JUL 30

WATCH 06

## Samsung Q2 divisional results

The company-level HBM4 disclosure (reported $1B in four months, ~$10B annualized pace by year-end) would convert prediction p56's hit from supply-chain reporting to primary confirmation — and set the tone for SK hynix's first earnings as a US-listed company.

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 12**

Week 28 of 2026 · July 11, 2026

WEB

**brianletort.ai/industry**

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