AI Leaves the Cloud

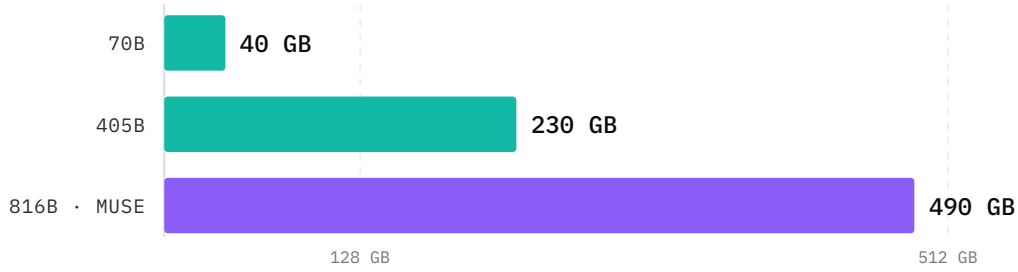
James E. Coleman

STRATEGIC DESIGN · WORKFORCE ACTIVATION · 2026

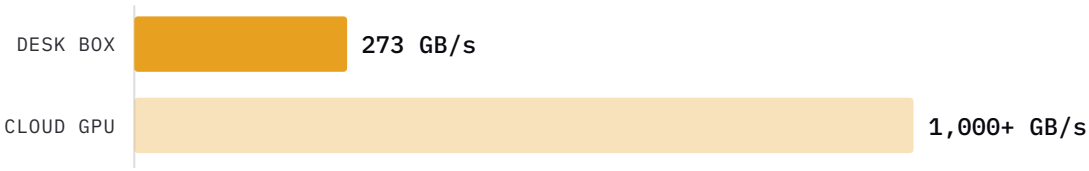
ABSTRACT *The goal is total privacy: artificial intelligence that runs on hardware you control, where your prompts, your files, and your working context never leave for a shared cloud model — not ChatGPT, not Claude, not anyone's. What makes that newly possible is not the GPU but the memory chip. When the shared, cloud-pooled memory that made modern AI affordable gives way to private silicon, the data-for-intelligence bargain built on top of it ends — and the data stays yours.*

BY THE NUMBERS **FIG. 1**

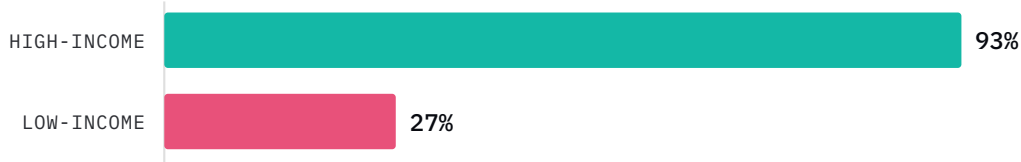
MEMORY CAPACITY — MODEL WEIGHTS AT 4-BIT VS ON-DEVICE MEMORY



MEMORY BANDWIDTH — UNIFIED MEMORY VS A DISCRETE GPU



WHO THE CLOUD REACHED — INTERNET USE BY COUNTRY INCOME



Model weights at 4-bit ≈ 0.5 byte/parameter (approx.); the 816B Muse Spark Open runs across a pair of 512 GB M5 Ultra Mac Studios. DGX Spark: 128 GB capacity,

273 GB/s bandwidth. Internet use 27% in low-income vs 93% in high-income countries; 2.6 billion offline (≈32% of humanity). Sources: refs 1–3, 7.

---

§ 01 Why AI Lived in the Cloud

It was never mainly about *compute.*

The common story is that large models live in the cloud because they need enormous compute — racks of GPUs doing trillions of operations per second. That is true, but it is the second-order constraint. The first-order constraint is memory. A model's weights, and the growing context it reasons over, have to physically sit in fast memory that the processor can reach at full speed. Frontier models do not fit in the memory of any single consumer device, so the industry pooled memory at hyperscale — tens of thousands of accelerators sharing one vast, high-bandwidth memory fabric — and rented slices of it out.

That pooling is the quiet engine of the whole cloud-AI economy. Shared memory is what let providers amortize the most expensive component across millions of users, and it is what made the exchange underneath feel free: you hand over your prompts, your documents, your telemetry, and in return you draw on a memory pool no individual could afford. *The data-for-intelligence bargain is a memory-sharing arrangement first and a data arrangement second.* Break the memory dependency and the data dependency breaks with it: a model running on memory you own has no reason to send your prompts, your files, or your context anywhere at all.

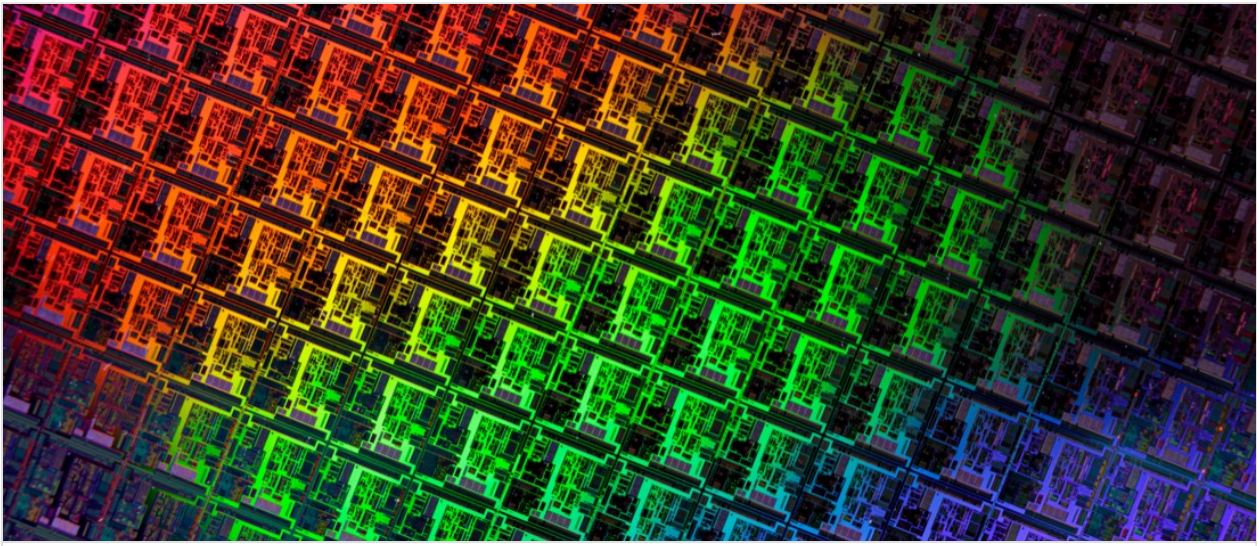
Which is why the interesting question is not "when does compute get cheap enough to run AI locally" — compute has been getting cheap for years. The question is when does enough fast memory, at enough bandwidth, land on private hardware that the pooling stops being necessary. That threshold is now visibly in motion, and it moves the bargain with it.

> *"The GPU gets the headlines. The memory chip writes the rules. Capacity decides whether a model fits on your hardware at all; bandwidth decides whether it runs fast enough to be worth using. Everything about where intelligence is allowed to live follows from those two numbers."*
>
> — JAMES E. COLEMAN

§ 02 The Personal AI Supercomputer

A data-center capability *on a desk — with a catch.*

NVIDIA's DGX Spark is the clearest signal. It ships a GB10 superchip with 128 GB of unified LPDDR5x memory shared between CPU and GPU, rated at roughly 1 PFLOP of low-precision performance, in a box that sits on a desk.[1] For the first time, the *capacity* half of the memory constraint is solved at the individual level: models that used to demand a rented cloud instance now fit in private memory you own outright.



A finished silicon wafer — rows of identical microchip dies, refracting light into a spectrum.

Photo: Rob Bulmahn · CC BY 2.0

The catch is the other half. That unified memory runs at about 273 GB/s of bandwidth — a fraction of a high-end discrete GPU, where a current flagship moves memory at well over 1 TB/s.[2] Capacity lets the model load; bandwidth decides how fast it thinks. So the desk-side machine can hold frontier-scale models the cloud used to monopolize, but serves them more slowly. That trade — *own the capacity, rent the speed* — is exactly the seam the next few years get decided on.

And capacity, unlike bandwidth, clusters. Link two of these units and their memory pools add: 128 GB plus 128 GB behaves, for the purpose of fitting a model, like 256 GB. Pooling — the exact trick that justified the hyperscale cloud — turns out to work on a shelf in an office. The cloud's structural advantage was never magic. It was memory you couldn't yet buy. Now, increasingly, you can.

§ 03 Private AI on Local Hardware

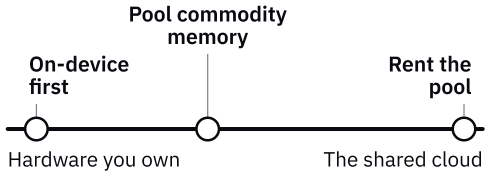
Two camps are already building *the private version.*

Apple has architected for this from the device outward. Most Apple Intelligence requests are designed to run entirely on-device on Apple's own Foundation Models, so the data never leaves the hardware; only the harder requests escalate to Private Cloud Compute, a hardened Apple-silicon server tier built specifically for inference with verifiable privacy guarantees.[4] Apple ships those on-device models to developers through its Foundation Models framework and the MLX runtime, and its newer lineup splits deliberately across on-device and server tiers — some models now served, tellingly, from third-party GPUs in Google Cloud.[5]

The sharpest signal came in October 2026, when Apple and Meta — long rivals — aligned to run Meta's Muse Spark Open, an 816-billion-parameter open-weight model, *locally* on Apple hardware. Meta demonstrated agentic workflows executing entirely on-device across a pair of M5 Ultra Mac Studios (up to 512 GB of memory each) through Apple's MLX framework.[3] This is private AI in the literal sense: the weights are downloaded and yours, the hardware is yours, and not one prompt, file, or token of the work leaves for OpenAI or Anthropic. It is not a consumer convenience — it demands serious, costly machines, and the Muse Model License still governs commercial use — but the capability is real and shipping.

What makes it possible is memory pooling on hardware you own. The open-source project EXO pioneered stitching the unified memory of several Apple-silicon Macs into one address space so a frontier-scale model fits on machines no single box could hold; Apple has since added operating-system support for high-speed Thunderbolt 5 clustering, making the pooling a first-class capability rather than a hack.[6] The exact trick that justified the hyperscale cloud — pool the memory, fit the model — now runs on a shelf in an office.

FIG. 1 — TWO ROUTES OFF THE CLOUD STATE 1 OF 3



APPLE'S APPROACH

On-device first

Apple's approach. Run what fits locally; escalate only the hard requests to a privacy-hardened server tier. Optimizes for governance and data residency. Accepts a smaller model as the price of keeping data on the device.

EXO AND OPEN WEIGHTS

Pool commodity memory

EXO & open weights. Cluster the unified memory of several machines so a frontier-scale model fits on hardware you own. Optimizes for capability and independence from any single provider. Accepts slower, hand-built infrastructure.

THE INCUMBENT CLOUD

Rent the pool

The incumbent cloud. Keep memory shared at hyperscale and meter access. Still wins decisively on bandwidth and convenience — and on the data it collects in exchange. The default, for now.

A diagram of the relation, not a chart of data. Positions are ordinal, from the descriptions: where the model's memory sits, from a machine you own to a metered cloud.

§ 04 OpenAI's Other Bet

Not everyone is betting on *local hardware.*

The move off the cloud is not unanimous, and the dissent is instructive. OpenAI is working with Broadcom and TSMC on a custom *inference* chip — silicon purpose-built to run already-trained models, reported as a roughly $10 billion program first targeted for 2026 and built for OpenAI's own data centers rather than for sale.[8] That is a bet in the opposite direction: instead of pushing intelligence outward to private hardware, drive the *cost of centralized inference* down far enough that the cloud stays the cheapest place to run.

Both bets can be right at once, and that is the point. Local hardware wins where governance, latency, and data residency dominate — regulated industries, sovereign data, personal context that should never leave a device. The cloud wins where raw capability and bandwidth dominate — the largest models, the heaviest context, the users who want the frontier and will trade data for it. The market is not choosing one memory architecture. It is splitting along the seam between them.

§ 05 The Scoreboard

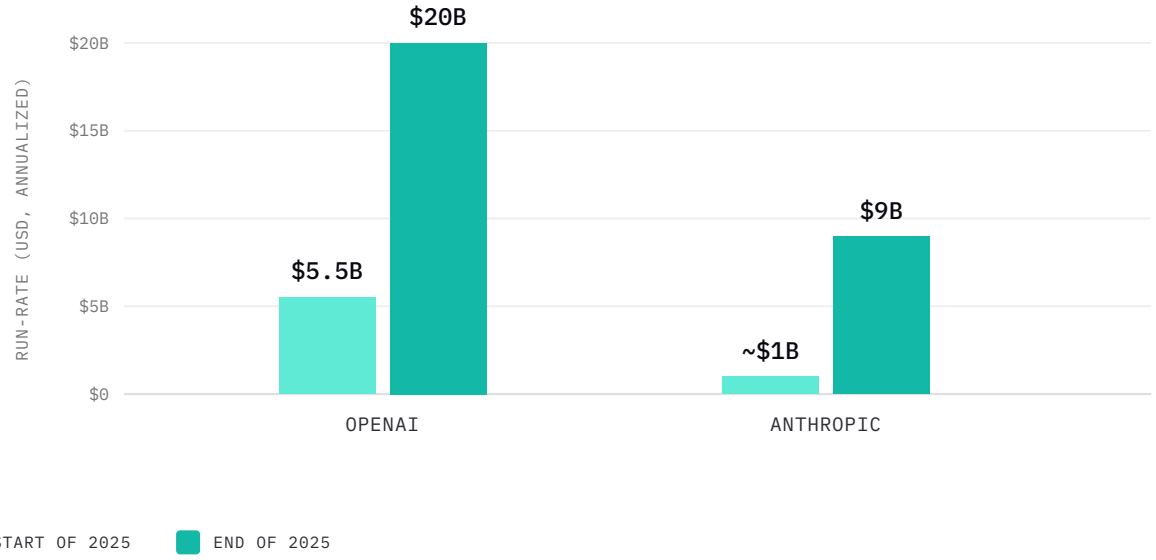
Follow the money — *and read it honestly.*

The capital says the cloud bargain is, for now, enormously valuable. OpenAI closed 2025 with an annualized revenue run-rate above $20 billion (against roughly $13 billion of revenue actually booked for the year); Anthropic closed near a $9 billion run-rate (against about $4.6 billion booked).[9][10] The incumbent dwarfs both: Alphabet booked $402.8 billion in total 2025 revenue, with Google Cloud alone growing 48% to $17.7 billion in a single quarter.[11]

Honesty about those numbers matters, because the industry rarely offers it. A *run-rate* is the most recent month annualized — a projection, not money in the bank — and for the frontier labs it currently

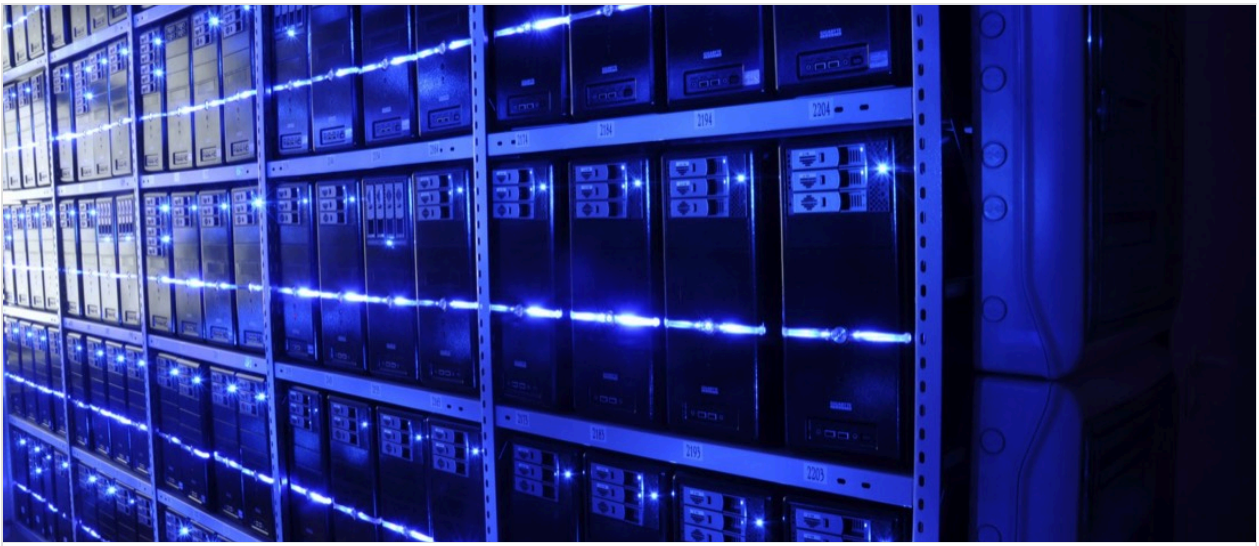
runs roughly double booked revenue. The direction is still staggering, which the chart makes plain: both labs multiplied their run-rate several times over in a single year.

ANNUALIZED REVENUE RUN-RATE · 2025 **FIG. 2**



Run-rates are annualized projections, not booked revenue; figures are rounded. OpenAI ≈ $5.5B → $20B; Anthropic ≈ $1B → $9B over 2025. Alphabet's $402.8B is total company revenue (FY2025), a different measure shown only for scale. Sources: refs 9–11.

Now read the other side of the ledger, because the industry rarely volunteers it. The revenue pouring into cloud AI is dwarfed by the capital pouring out to build it. The four largest hyperscalers spent a record ≈$388 billion on capital expenditure in 2025, are tracking past $800 billion across 2026, and are projected to top $1 trillion by 2027.[12] The balance sheets already show the strain: Alphabet, Amazon, and Meta have each posted negative free cash flow in recent quarters — Amazon swung to roughly −$42 billion on a trailing-twelve-month basis, from about +$15 billion a year earlier — leaving Microsoft the lone hyperscaler still cash-flow positive, and even its cash generation fell 27%.[12] The group's full-year free cash flow is tracking to its lowest level since 2014.

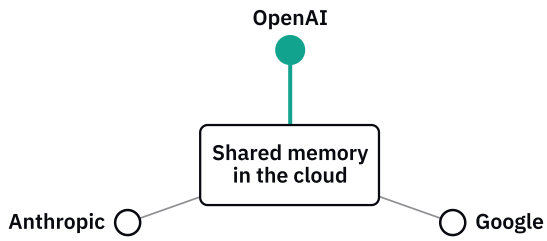


A data-center hall — rows of rack-mounted servers under blue indicator light.

Photo: BalticServers.com · CC BY-SA 3.0

That is the squeeze, and it bears directly on where intelligence lives. The shared-memory cloud is the most capital-intensive way to deliver intelligence ever built, and it is now draining the cash flows of the richest companies on earth. *Every query that runs on memory you own is a query no one had to build a data center to serve.* Privacy is why an organization wants local AI; economics is why the providers may eventually be relieved to let the inference go.

FIG. 3 — WHERE THE THREE ARE BETTING STATE 1 OF 3



MAKES THE CLOUD CHEAPER

OpenAI

Doubling down on **centralized inference** — custom Broadcom/TSMC silicon to make the cloud the cheapest place to run. The data bargain stays intact, just cheaper to operate.

OWNS THE WHOLE STACK

Google

Owns the **full stack** — its own TPUs, its own cloud, its own distribution — and the scale to run every option at once, including hosting a rival's model tiers on its infrastructure.

CLOUD-NATIVE, SOLD THROUGH PARTNERS

Anthropic

Fastest run-rate growth of the pure-play labs, still **cloud-native** and partner-distributed — leveraged to the same shared-memory economics as OpenAI.

A diagram of the relation, not a chart of data. Each lab's bet is described by how it relates to the shared cloud. Choose a lab to hold it.

§ 06 What Changes

When memory moves, *governance moves with it.*

Governance. Once a capable model runs on hardware an organization owns, the central questions of AI data governance stop being "what does the provider promise" and become "what does our own infrastructure enforce." Data residency becomes a physical fact rather than a contractual one. For regulated industries, sovereign institutions, and anyone holding data that should never transit a third party, that is the entire game — and it is why the edge wins its segment regardless of who wins on raw capability.

Concentration. The providers are not only cash-squeezed; they are fusing with the state. Alphabet is in early talks with the Pentagon over a data-security role in the $185 billion "Golden Dome" missile-defense program — reportedly its first publicly known involvement — months after agreeing to deploy its AI models for classified defense operations.[14] Set the economics aside for a moment: when the infrastructure you run your intelligence on is *also* becoming critical national-security infrastructure, dependence on it stops being a procurement question and becomes a sovereignty one. Running AI on hardware you control is, increasingly, the only way to keep your data out of a stack that answers to interests far larger than your own.

Equity. The cloud bargain was always conditional on connectivity, and connectivity is not universal. The ITU estimates 2.6 billion people — about 32% of humanity — were offline in 2024, with 1.8 billion of them in rural areas, and internet use of just 27% in low-income countries against 93% in high-income ones.[7] A model that only works when it can reach a data center simply does not reach a third of the world. Intelligence that runs locally, on modest hardware, without a reliable link, is not a convenience feature for these populations — it is the *only* version of the technology that includes them.



A remote rural village — mud-brick, thatched-roof homes on open ground.

Photo: Frankincense Diala · CC BY-SA 4.0

Honesty about limits. Local hardware is not a free lunch, and pretending otherwise is how this conversation usually goes wrong. On-device and clustered models are smaller or slower than the cloud frontier; private infrastructure is hand-built and unglamorous; bandwidth remains the cloud's durable moat. The useful posture is not triumphalism but calibration: *name what local hardware genuinely delivers — governance, residency, independence, access — and name what it still cannot, so the trade is made with open eyes.*

§ 07 The Shape of What's Coming

A split, not a *migration.*

AI does not leave the cloud in a single move; it migrates workload by workload, and the direction of travel is decisively toward local. Memory capacity landing on private hardware breaks the pooling monopoly that made the data-for-intelligence bargain the only deal on offer. As on-device memory grows and its bandwidth climbs, the set of tasks that genuinely require a data center keeps shrinking — and two forces push every borderline case the same way: privacy pulls data onto hardware you own, and the capex squeeze makes centralized inference the expensive option to defend. The plausible end state is not a tidy equilibrium but a steady tilt — *local models as the default for the overwhelming majority of real work,* with the shared cloud reserved for the genuine frontier until the memory on your desk catches up with that too.

The strategic implication is concrete. Organizations should stop treating "AI" as a single sourcing decision and start treating it as a portfolio question: which workloads belong on shared-memory in-

frastructure because capability dominates, and which belong on owned memory because governance, residency, or connectivity dominates. The firms that answer deliberately — rather than defaulting everything to whichever provider is loudest — will be the ones that keep both their capability and their data.

> *"Watch the memory chip, not the model announcement. Capacity decides what you can own; bandwidth decides what you still have to rent. The map of where intelligence is allowed to live is drawn in those two numbers — and it is being redrawn right now."*
>
> — JAMES E. COLEMAN, AI LEAVES THE CLOUD

§ R References — Primary Sources

Sources *& notes.*

Every quantitative claim is cited to a primary or well-established secondary source below. Revenue run-rates are annualized projections, not audited results, and are labeled as such. Product names are given as the vendors and established outlets report them; where the press has floated unconfirmed internal codenames, this dispatch omits them rather than assert them as fact.

1. NVIDIA. “DGX Spark — Hardware Overview.” 128 GB unified LPDDR5x memory, 273 GB/s bandwidth, GB10 superchip, ~1 PFLOP FP4. docs.nvidia.com. See also StorageReview, “NVIDIA DGX Spark Review.” storagereview.com.

2. Implicator.ai. “NVIDIA ships DGX Spark … a bandwidth trade-off.” On the 273 GB/s figure versus discrete GPUs exceeding ~1 TB/s. implicator.ai.

3. Meta Muse Spark Open — an ~816-billion-parameter open-weight model (released October 2026), demonstrated running entirely on-device on a pair of Apple M5 Ultra Mac Studios via Apple's MLX framework, in an Apple–Meta alignment. MacDailyNews, “Apple and Meta partner to run massive Muse AI model locally on Mac Studios.” macdailynews.com; YourStory, “Meta Muse Spark Open brings 816B-parameter AI to Mac Studios.” yourstory.com.
   Released under the Muse Model License (limited commercial terms); requires high-memory Apple silicon — not a typical Mac.

4. Apple / The Register. “Apple built its own AI cloud servers” — Private Cloud Compute runs larger server-based models on hardened Apple silicon while most requests run on-device. theregister.com; 9to5Mac, “Apple Private Cloud Compute.” 9to5mac.com.

5. Apple’s foundation-model lineup spans on-device and server tiers (some served from Google Cloud infrastructure), per WWDC 2026 coverage. lavx.hu.

6. EXO Labs pioneered pooling Apple-silicon unified memory across Macs to run frontier-scale models; Apple added OS-level Thunderbolt 5 clustering in macOS Tahoe 26.2. Engadget, “You can turn a cluster of Macs into an AI supercomputer.” engadget.com; CO/AI, “Run powerful open-source LLMs on a Mac.” getcoai.com.

7. ITU. *Facts and Figures 2024.* An estimated 2.6 billion people offline (≈32% of the world; 1.8 billion in rural areas); internet use 27% in low-income vs 93% in high-income countries. itu.int.

8. OpenAI’s custom inference chip with Broadcom and TSMC — reported ~$10 billion program, inference-focused, for internal use, first targeted for 2026. SDxCentral. sdxcentral.com; Tom’s Hardware. tomshardware.com.

   Some outlets have reported an internal codename; it is unconfirmed and omitted here.

9. OpenAI 2025 revenue: annualized run-rate above $20 billion at year-end (The Information / Reuters), against ~$13.07 billion of revenue booked for 2025 (reported from leaked audited figures, Fortune). Summary coverage: Yahoo Finance / Bloomberg, “OpenAI revenue run rate tops $40 billion.” finance.yahoo.com.

10. Anthropic 2025 revenue: run-rate of roughly $9 billion at year-end (Anthropic), against ~$4.6 billion booked for 2025 (IPO prospectus, via Reuters). Run-rate rose past $5 billion by August 2025. VentureBeat. venturebeat.com.

11. Alphabet. Q4 & FY2025 earnings — total 2025 revenue $402.836 billion (+15%); Google Cloud +48% to $17.7 billion in Q4. Coverage: DQ India, “Alphabet surpasses US$400 billion annual revenue.” dqindia.com.

12. Hyperscaler capital expenditure & free cash flow: a record ≈$388 billion for the big four in 2025, tracking past $800 billion across 2026, with Big Tech capex projected to top $1 trillion by 2027; Alphabet, Amazon, and Meta posted negative free cash flow in recent quarters (Amazon ≈ −$42.4 billion trailing twelve months, from +$14.8 billion a year earlier), leaving Microsoft the lone positive. CNBC, “Tech giants … pour trillions into the AI buildout.” cnbc.com; Yahoo Finance, “Hyperscalers’ Free Cash Flow Dips.” finance.yahoo.com.

13. Meta's open-weight strategy — releasing Llama (and successors) to commoditize the model layer and avoid dependence on a closed provider such as OpenAI; see Zuckerberg, “Open Source AI Is the Path Forward” (2024), and analysis in Interconnects, “Llama 3.1 405B, Meta's AI strategy.” interconnects.ai.

    “Open weights” is the precise term — the weights are released; the training data is not.

14. Alphabet (Google) in early talks with the Pentagon over a data-security and cybersecurity role in the ~$185 billion “Golden Dome” missile-defense program — reportedly its first publicly known involvement — following an April agreement to deploy its AI models for classified defense operations; talks are not final. Bloomberg, via Yahoo Finance, “Alphabet Is in Talks to Join Trump's Golden Dome Defense Program.” finance.yahoo.com.

    Program cost estimates vary widely — White House ≈ $175B; CBO up to $1.2T over 20 years.

CLOSING STATEMENT

"The cloud did not win because intelligence belongs there. It won because memory was the one thing you couldn't yet buy for yourself. That is changing — capacity first, bandwidth in time — and as it changes, the bargain changes with it. The job now is to decide, workload by workload, what you keep on your own memory and what you still rent on someone else's."

— JAMES E. COLEMAN — FIELD DISPATCH · 2026

James E. Coleman

STRATEGIC DESIGN · AI DATA GOVERNANCE, LOCAL & PRIVATE AI INFRASTRUCTURE

James works with organizations on AI data governance, local and private AI deployment, and the equitable use of public information — with a bias toward plain, cited analysis over hype. This dispatch is part of an ongoing series tracking where intelligence is actually allowed to live, and who the architecture includes or leaves out.