WHITE PAPER

STRATEGY · JULY 2026

THE AI RENT STACK

Forward-Deployed Engineering and the Contest for the Enterprise's Last Mile

Every major lab and cloud now wants to embed its engineers inside your business. The move looks like conviction. Read the economics underneath and it looks more like each layer of the industry defending its ground, and handing you both an opportunity and a liability you have not yet priced.

AUTHOR

Joshua Davis

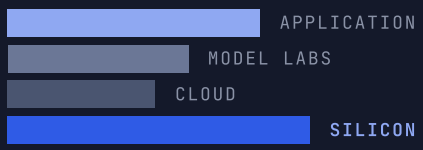
Lead AI GTM Strategist · GitHub

PUBLISHED

July 17, 2026

READING

~24 minutes



Soli Deo gloria

CONTENTS

EXHIBITS

EXECUTIVE SUMMARY

Between December 2025 and July 2026, the largest companies in artificial intelligence did the same thing at almost the same time. Anthropic helped stand up a services firm with Wall Street money. OpenAI spun out a majority-owned deployment subsidiary. Microsoft committed billions to an organization of embedded engineers. Google funded a network of partners to do the same, Amazon expanded a center that gives the work away, and the global consultancies branded matching practices to keep pace. The role at the center of all of it—the forward-deployed engineer, sent to live inside a customer's operation and build against its data—was invented at Palantir more than a decade ago. Its sudden universality is the signal this paper examines.

The argument here is that the deployment rush is not, at root, a services trend. It is what a value chain does when its headline product commoditizes faster than the revenue it generates can cover the capital behind it. The price of raw intelligence is collapsing—the cost to perform a fixed task has been falling at roughly fifty-fold a year, while the infrastructure required to produce it now consumes hundreds of billions of dollars in annual capital expenditure. Caught between a cheapening product and an expensive supply chain, every layer of the industry is racing toward the one place margin and durability still live: the customer's last mile, where a deployment either sticks and consumes or stalls and does not.

To see who is defending what, it helps to stop sorting these companies as "AI-first" versus "incumbent" and instead sort them by where they sit in what we will call the *rent stack*. At the bottom are the chipmakers and memory suppliers, who capture the scarcity premium at gross margins near seventy-five percent. Above them sit the cloud landlords, who own the compute and rent it out. Above the landlords sit the model-maker tenants, who rent that compute and sell tokens. And nearest the enterprise sits the application and deployment layer, which sells outcomes and is largely insulated from the capital below it. Forward-deployed engineering means something different at each altitude: for the landlords it is cheap insurance on a capital bet; for the tenants it is a ladder out of a margin squeeze; for the application layer it is the business itself.

That difference is the entire decision for an enterprise buyer. A forward-deployed engagement can leave you more capable and self-sufficient, or it can leave you dependent on a subsidy that can be withdrawn and a vendor whose own balance sheet you have quietly taken onto yours. This paper maps the stack layer by layer, weighs the long-term viability of each, traces the circular financing that couples them into a single fragile system, and closes with a practical framework for underwriting any deployment offer that lands on your desk. The short version: the vendors have already conceded that the model is no longer the moat. What they are selling in its place is either a partnership or a leash, and telling the two apart is now a core procurement competence.

01 Everyone Rushed the Same Door

On July 15, Anthropic and a consortium of financial institutions introduced Ode with Anthropic, a standalone enterprise AI services firm built from Anthropic's models, an acquired engineering startup called Fractional AI, and roughly $1.5 billion in backing from Blackstone, Hellman & Friedman, Goldman Sachs, and a longer roster that includes General Atlantic, Apollo, GIC, and Sequoia.[1] The company runs about a hundred engineers on a "Claude-first" but multi-model basis and aims at the middle of the market, where organizations have finished their pilots and now need someone to build the production system. Its chief executive, Chris Taylor, reached immediately for the ceiling: "It's pretty easy to imagine this as a trillion-dollar company someday if we execute well."[2]

The number is not the interesting part. The concession is. A frontier lab whose entire market thesis has been that better models win just helped capitalize a company premised on the idea that the model is no longer the hard part. Ode's chief technologist, Eddie Siegel, put it in terms a CFO can price: model selection matters, "but it's not where the majority of calories are spent," a choice, he said, more like the choice of a programming language than a source of durable advantage.[3] When the people who make the models tell you the models are close to interchangeable, that is worth sitting with.

Ode arrived late to a pattern already well established. In May, OpenAI launched a majority-owned subsidiary bluntly named The OpenAI Deployment Company, capitalized with $4 billion from nineteen outside investors at a reported $14 billion valuation, its backers promised a guaranteed minimum return of 17.5 percent—the financing architecture of an infrastructure fund rather than a software product.[4] In July, Microsoft unveiled Frontier, a $2.5 billion organization of roughly six thousand engineers built to embed inside customer operations, though its leadership pointedly declined to use the industry's own term for the role.[5] Google committed $750 million in April to fund partners deploying its agentic tools.[6] Amazon's Generative AI Innovation Center, seeded with $100 million in 2023 and doubled in 2025, offers its embedded engineering to customers at no charge at all.[7] Deloitte branded a Forward Deployed Engineering practice in December; Accenture launched a Microsoft-specific one in March.[8]

The role they are all chasing is not new. Palantir invented it in the early 2010s, sending engineers to work inside customer sites and build against live data rather than selling software over the transom. The model produced numbers most software companies would envy—first-quarter 2026 revenue of $1.63 billion, up 85 percent year over year, an adjusted operating margin near 60 percent, and commercial customers, once a rounding error, accounting for close to half the business.[9] That is the template every latecomer is now copying. The open question, and the one this paper turns on, is whether they are copying the results or only the costume.

> **THE PATTERN IS THE ARGUMENT**
>
> When six competitors who agree on almost nothing simultaneously conclude that the money has moved from the model to the deployment, the conclusion is worth more than any single company's spin on it. The land grab is a collective admission that the product layer has commoditized.

What none of the announcements make legible is the economics underneath—who funds the engineers, what they are really being paid to protect, and whether the demand they manufacture can outlast the subsidy that seeds it. To answer that, we need a map.

02 The Rent Stack: Where the Money Actually Sits

The reflex is to sort these companies into "AI-first" upstarts and lumbering incumbents. That taxonomy obscures more than it reveals, because it files Palantir next to OpenAI when the two could hardly be more different as businesses, and it treats Microsoft and Amazon as a single category when their AI exposures are structurally distinct. A more useful cut sorts the field by altitude in a stack of rents, where each layer pays the layer below it for a scarce input and sells to the layer above.

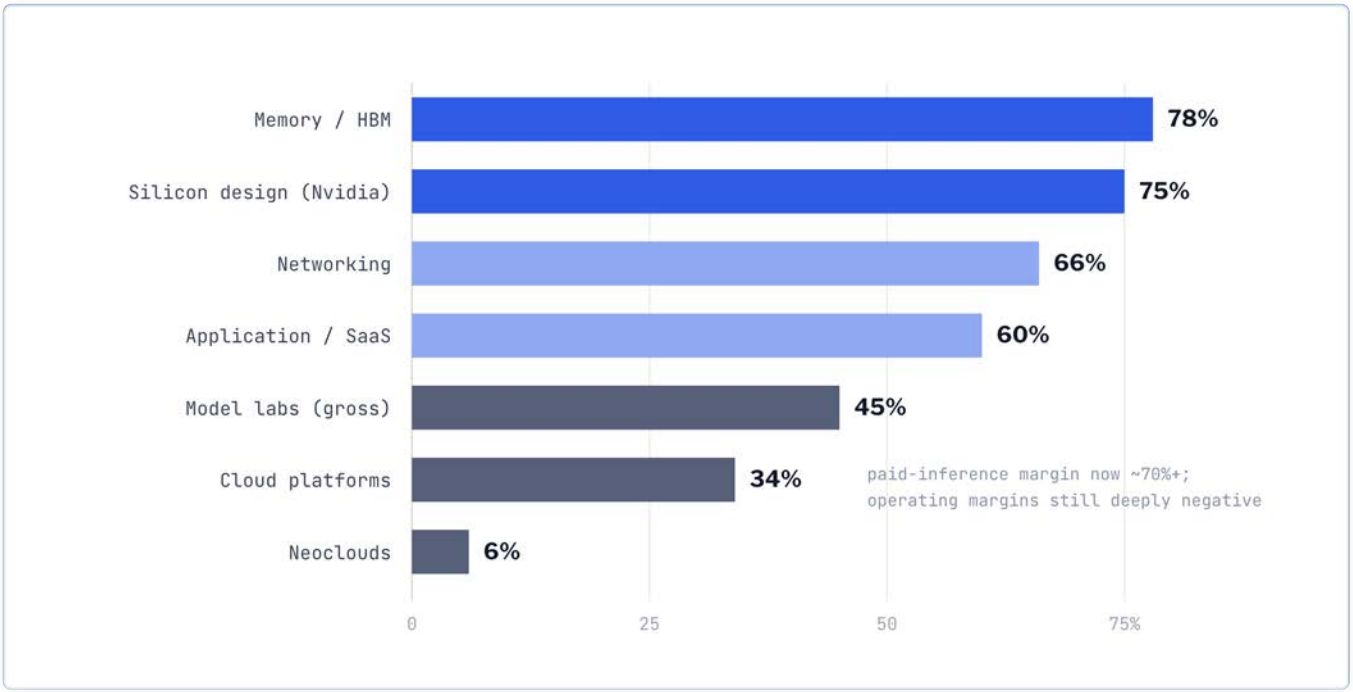
At the base sits silicon and memory. Nvidia designs the accelerators nearly everyone trains and serves on, and in its fiscal year ended January 2026 it booked $215.9 billion in revenue—$193.7 billion of it from the data-center segment—at a full-year gross margin of 71 percent, rising toward 75 percent in the most recent quarter.[10] Alongside it, the memory makers that supply high-bandwidth memory run gross margins in the seventy-to-eighty-five-percent range as an AI-driven shortage has sent DRAM contract prices up several-fold.[11] This is where the scarcity premium is captured most completely. These are the true landlords, and everyone above them pays rent.

One layer up sit the cloud providers—Microsoft, Google, and Amazon—who buy the silicon, assemble it into data centers, and rent capacity to everyone else. They are middle landlords, running cloud operating margins in the low-to-high thirties, and their position is genuinely double-edged: they collect the premium when they rent GPUs out, but they pay it when they buy chips in, and they carry the depreciating hardware on their own balance sheets. Above them sit the model makers—OpenAI, Anthropic, and their peers—who rent that compute and sell intelligence by the token. And nearest the enterprise sits the application and deployment layer—Palantir, the systems integrators, the new services ventures—which sells outcomes, stays largely model-agnostic, and is insulated from the capital intensity of everything below it.

~75%	30–38%	~50×/yr	~$1.5T
Nvidia quarterly gross margin (Q1 FY2027, reported)	Hyperscaler cloud operating margins (2026, reported)	Median decline in cost per fixed AI task	Annual revenue needed to justify one year of 2026 AI capex (Sequoia)

FIG. 1—THE AI RENT STACK

The scarcity premium sits at the bottom of the stack



Approximate margin by layer of the AI stack. The scarcity premium concentrates at silicon and memory; the capital-intensive middle—cloud platforms, neoclouds, and the model labs' blended economics—runs thin. Analyst estimates (Data Gravity, SemiAnalysis 2026); cloud and neocloud shown as operating margin.

Two forces move through this stack in the same direction, and together they explain the deployment rush. The first is the scarcity premium, which accrues downward: because compute is constrained—by chips, and increasingly by the power and transformers needed to energize them—the layers closest to the metal capture the richest margins, and everyone above pays up for access.[12] The second is commoditization, which also pushes downward: the price of the model layer's output is collapsing, with the cost to perform a fixed task falling at a median rate near fifty-fold per year.[13] Put the two together and the model makers are squeezed from both sides at once—paying a premium for

scarce compute below while the price of the tokens they sell above erodes underneath them. That squeeze is the single most important fact in the industry's income statements, and it is why the tenants are climbing toward the enterprise as fast as they can.

Read through this map, forward-deployed engineering is not one strategy. It is four different moves that happen to share a job title. For the landlords, embedding engineers is inexpensive insurance on a capital bet made one layer down: a few billion dollars, funded from operating cash flow, spent to make sure the compute they have already committed to gets consumed. For the tenants, it is an escape route—a way to climb out of the margin squeeze toward the customer relationship, the services fee, and the switching cost. For the application layer, it is simply the business, and always has been. And for the consultancies, it is a rebranding of work they have billed for decades. The rest of this paper takes the stack apart layer by layer, because the viability of each—and the risk each transfers onto an enterprise customer—is different.

> **THE FRAMEWORK IN ONE LINE**
>
> The scarcity premium falls downward and the commoditization pressure falls downward, so the model layer is squeezed in the middle, and forward-deployed engineering is the ladder it is using to climb toward the enterprise, while the landlords use the same ladder as a cheap lock on demand.

03 Why the Whole Field Moved at Once

Simultaneity like this is rarely a coincidence. Six sets of executives reading the same market arrived at the same conclusion in the same two quarters because the same two numbers were staring back at all of them: the price of what they sell is falling, and the demand that would justify what they have spent has not yet shown up in the returns that matter.

Take the price collapse first, because it is the more counterintuitive of the two. Falling prices are usually a sign of health—cheaper intelligence means broader adoption, and broader adoption is the whole thesis. But for the companies producing the intelligence, a fifty-fold annual decline in the cost of a fixed task is a solvent working on their margins.[13] They cannot simply ride it down, because the newest reasoning and agentic systems consume dramatically more tokens per task than their predecessors—so much more that the same analysis measuring the price decline excludes them from the comparison entirely. The result is a paradox that lands directly on enterprise budgets: the price per token keeps falling while the token count per job keeps climbing, and the bill often rises even as the unit cost drops. Cheaper intelligence, more of it consumed, and a meter that spins faster than the price falls.

FIG. 2—THE CONSUMPTION PARADOX

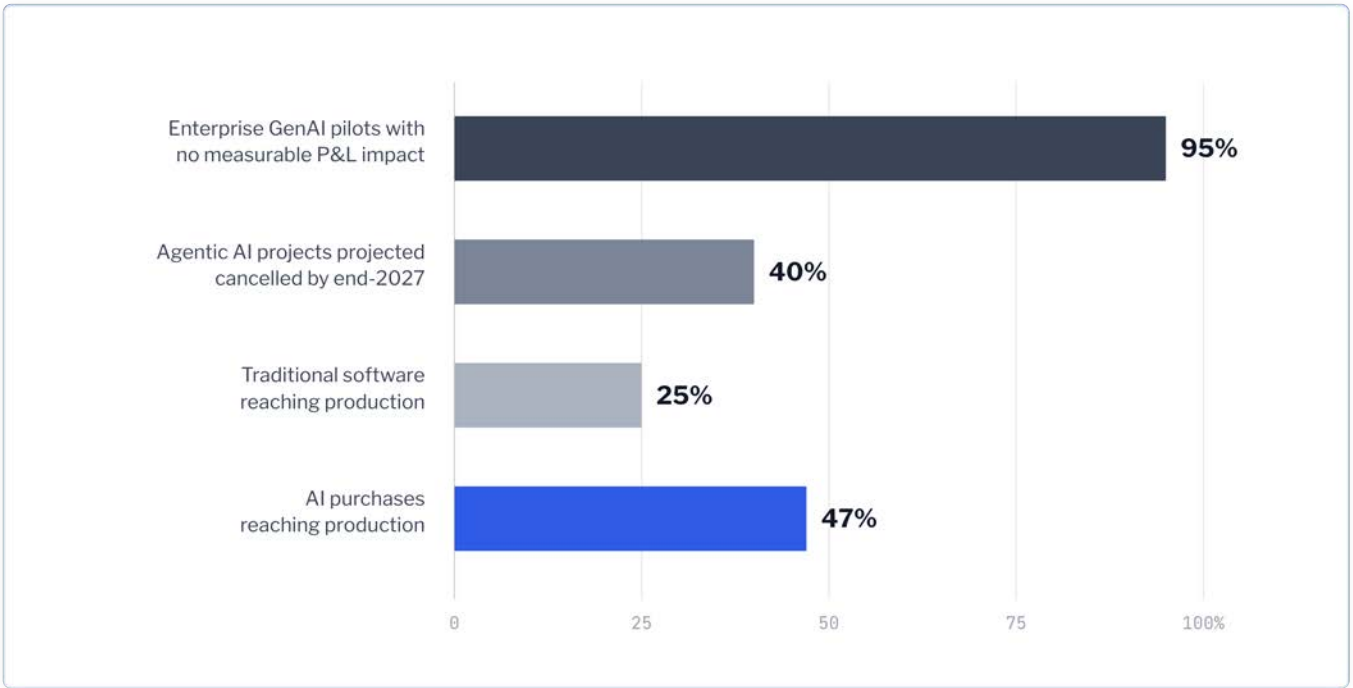
Cheaper per task, yet the meter spins faster

The price to perform a fixed task has fallen roughly fifty-fold a year, even as reasoning and agentic workloads multiply the tokens each task consumes, so the bill often rises anyway. Illustrative of the ~50×/year median price decline per task (Epoch AI, 2026); reasoning models excluded.

The demand is real, and this is worth stating plainly because the skeptical case sometimes overstates its absence. Enterprises spent roughly $37 billion on generative AI in 2025, more than three times the prior year's figure.[14] But spending is not the same as return, and the return is where the gap opens. A widely cited study from MIT last August found that about 95 percent of enterprise generative-AI pilots produced no measurable impact on profit and loss.[15] Gartner projects that more than 40 percent of agentic-AI projects will be scrapped by the end of 2027, undone by cost, unclear value, and weak controls.[16] Even the purchases that do convert convert slowly: by Menlo's count only about 47 percent of enterprise AI deals reach production, better than the roughly 25 percent typical of traditional software but still a market where more than half of what is bought never ships.[14] Both figures deserve a caveat—the MIT number measures the visibility of return, not its total absence, and Gartner's is a forecast—but the shape they describe is consistent with what buyers report: the technology works in the demo and stalls in production.

FIG. 3—SPENDING IS REAL; RETURN IS NOT—YET

The adoption gap deployment is meant to close



Enterprise AI spending is real, but roughly 95% of pilots show no measurable P&L impact and more than 40% of agentic projects are projected to be cancelled by 2027. MIT NANDA (Aug 2025); Gartner (Jun 2025, projected); Menlo Ventures (2025). MIT measures P&L visibility.

That stall is the problem forward-deployed engineering exists to solve, and it is why the timing is not accidental. A pilot that never reaches production consumes almost no tokens, generates no services fee, and justifies none of the capital committed upstream. If the enterprise cannot cross the last mile on its own, the vendor's rational move is to carry it across, because the alternative is a data center running below the utilization its financing assumed. The deployment engineer is the industry's answer to its own adoption gap, and the gap grew visible enough, in the same two quarters, for everyone to reach for the same answer at once.

> THE SIGNAL BENEATH THE SIGNAL
>
> The rush into deployment is a response to a specific and worsening mismatch: intelligence is getting cheaper to produce while proving stubbornly hard to turn into enterprise profit. Embedding engineers is how the industry tries to close that gap before the capital markets lose patience with it.

04 Two Engines Wearing One Name

The single most important distinction in this entire market is invisible from the outside, because it hides behind an identical job title. Two very different businesses both deploy "forward-deployed engineers," and they are not doing the same thing.

For a consulting firm, an embedded engineer is revenue. The client pays a day rate, the engagement carries a margin, and the margin is the point. This is an old, well-understood, cash-generative model, and it is what Deloitte, Accenture, and the new Ode venture are fundamentally running, whatever the branding. For a model lab or a cloud

provider, the same engineer is something else entirely: customer-acquisition spend, a cost absorbed—sometimes given away outright—to drive the thing that actually monetizes, which is consumption. Amazon gives its Innovation Center engineering away precisely because the work was never the product; the Bedrock invoice that follows is.[7] One analyst framing that has circulated among enterprise buyers holds that a dollar of forward-deployed services can pull several dollars of downstream platform usage. No vendor has published that ratio, and it should be read as directional rather than measured, but it captures the logic of the subsidized model exactly.[17]

The distinction matters because the two engines age very differently. The services engine is durable but bounded: it grows with headcount, carries consulting-grade margins, and does not pretend to be anything other than what it is. The subsidy engine is the one to watch, because its health depends on a question the vendor would prefer you not ask: what happens to your economics when the subsidy normalizes to market rates? A deployment that only pencils out while the engineers are free is not a deployment that has proven its value; it is one whose value is still being underwritten by someone else's balance sheet.

The reliable way to hold the distinction is to ask, of any engagement, who pays for the engineer and what that engineer is really being paid to grow.

DEPLOYMENT MODEL	WHO FUNDS THE ENGINEER	WHAT IT ACTUALLY SELLS	HOW IT AGES
Application / deployment (Palantir, the new services ventures)	Customer pays fees	Outcomes, plus product adoption	**Durable; margin-bearing**
Consultancies (Deloitte, Accenture, McKinsey)	Customer pays day rates	Billable expertise	**Durable, but headcount-bound**
Model labs (OpenAI, Anthropic in-house FDE)	**Vendor subsidizes**	Token consumption	**Contingent on the subsidy**
Cloud landlords (Microsoft, Google, Amazon)	**Vendor subsidizes from cash flow**	Cloud + platform consumption	Cheap to sustain; demand-driven

The consultancies occupy a revealing middle position. They bill their embedded pods as ordinary margin-bearing services, so they carry none of the subsidy risk, but they own none of the platform either, which is why industry analysts describe the forward-deployed engineer as "the last mile of the AI value chain"—a stretch of ground the labs and clouds are now racing to occupy before the integrators lock it up.[17] The buyer's advantage in all this is that the different engines leave different fingerprints, and once you know to look for them they are not hard to read.

Palantir is the existence proof that the model can be durable rather than a crutch. Everest Group calls it, not unfairly, "a category of one," because replicating its results requires product depth, embedded execution, and operational credibility at the same time—a combination most imitators lack.[18] But the skeptical reading has equally credible backing. Constellation Research, watching the same trend the labs are celebrating, is blunt that forward-deployed engineers are "often used as a crutch to smooth over product immaturity," and judges the pattern "transitory, not durable."[19] The chief executive of Anaplan, assessing Palantir's own approach, offered the sharpest version of the concern: it is "a good initial deployment model. It's not a great model for running the software. Not at all."[20]

Both readings are correct, because they are describing different engines. The task for an enterprise buyer is not to decide whether forward-deployed engineering is good or bad in the abstract. It is to determine, for the specific offer on the table, which engine is running underneath it, and the reliable tell is trajectory. In a durable engagement,

revenue per embedded engineer rises over time while the number of engineers a given customer needs falls, as five become one. In a crutch, the ratio never improves, because the gap between what the product ships and what the customer needs is not closing. That is a measurable quantity, and it is the one to measure.

05 The Landlords Up Close

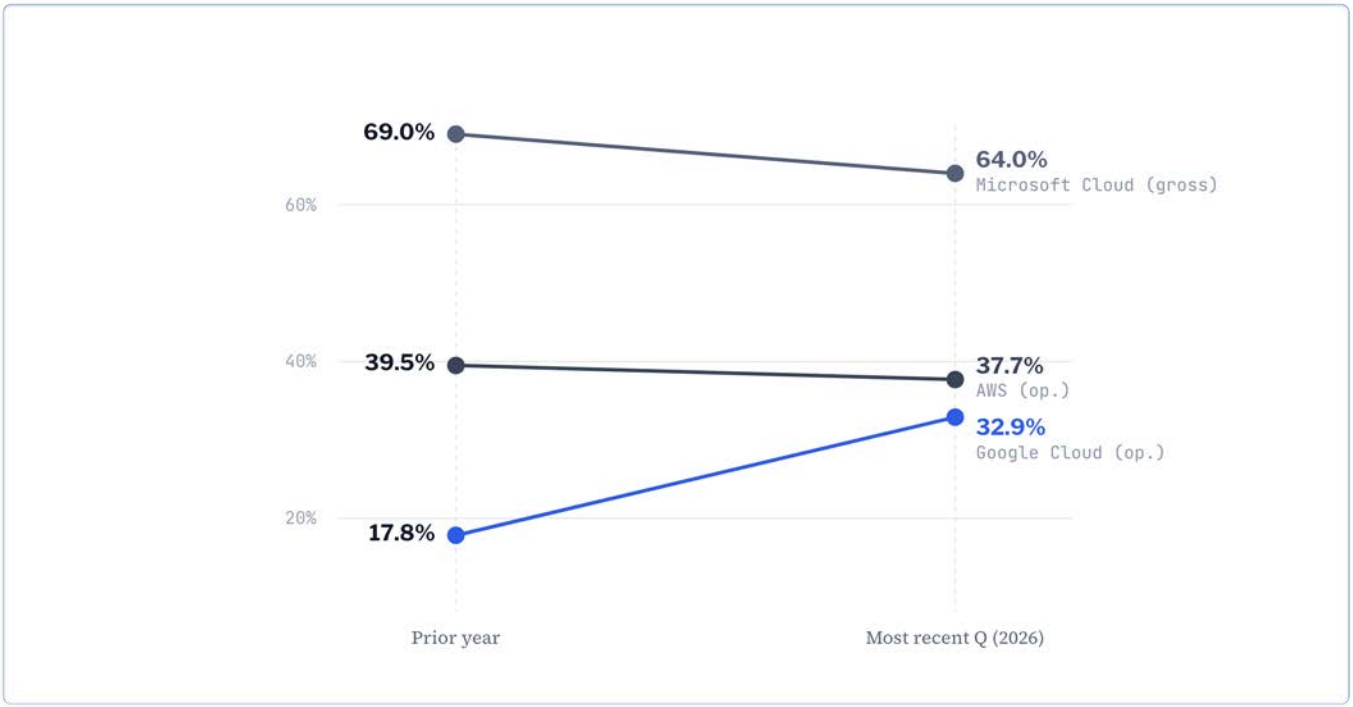
Return to the stack and look hardest at the middle landlords, because they are the layer most enterprises are actually betting on, and because their long-term viability is more varied than the "big cloud incumbent" label suggests. Microsoft, Google, and Amazon are running three materially different AI businesses, and the differences show up precisely where it counts—in whether the AI build-out is expanding their margins or eroding them.

Google looks, on the current evidence, the most durable of the three. In the first quarter of 2026, Google Cloud grew revenue 63 percent year over year and expanded its operating margin from 17.8 percent to 32.9 percent—the only one of the three demonstrating AI-era margin expansion rather than compression—against a backlog north of $460 billion that nearly doubled quarter over quarter, with management noting that revenue would have been higher still with more capacity.[21] Part of that strength is structural: Google's tensor processing units are its own silicon, so the vertical integration that everyone else is scrambling to build already captures margin internally rather than leaking it to Nvidia. Amazon occupies the middle. AWS reaccelerated to 28 percent growth, its fastest in fifteen quarters, and its custom-silicon business crossed a $20 billion annual run rate at triple-digit growth—a genuine profit engine—but its cloud operating margin declined year over year, from 39.5 percent to 37.7 percent, as capital and depreciation weighed on it, and Bedrock, while surging, sits atop that compressing base.[22]

Microsoft is the most exposed, and the exposure is instructive. Its AI business reached a roughly $37 billion annualized run rate, up 123 percent year over year, but its Microsoft Cloud gross margin slipped from 69 percent to 66 percent and was guided toward 64 percent, with the chief financial officer attributing the compression explicitly to AI infrastructure investment and growing Copilot usage.[23] That is a rare thing: a management team naming its own flagship product as a margin drag. Underneath the run-rate figure sits a concentration risk the others do not carry to the same degree. Analysts reading Microsoft's filings estimate that as much as $27 to $30 billion of that $37 billion AI run rate derives from a single customer—OpenAI's own consumption of Azure—a figure Microsoft does not disclose and one that turns a headline growth number into a bet on one counterparty's solvency.[24]

FIG. 4—DIVERGENT TRAJECTORIES

Only one landlord's margin is expanding



Cloud margin, year over year. Google Cloud's operating margin expanded from 17.8% to 32.9% while AWS and Microsoft compressed—one build-out, three trajectories. Alphabet, Amazon, Microsoft earnings, 2026 (reported); Google/AWS operating margin, Microsoft Cloud gross margin.

DIMENSION	GOOGLE CLOUD	AMAZON / AWS	MICROSOFT / AZURE
Cloud margin trajectory (YoY)	17.8% → 32.9% (expanding)	39.5% → 37.7% (compressing)	69% → ~64% gross (compressing)
Own AI silicon at scale	TPU (mature, margin-accretive)	Trainium (>$20B run rate)	Maia (no disclosed run rate)
Anchor-tenant concentration	Anthropic (shared) + first-party Gemini	Anthropic + OpenAI (compute)	OpenAI (~70–80% of AI run rate, est.)
Enterprise distribution moat	Workspace, search-adjacent	AWS install base	M365, GitHub, Teams

Two structural facts sit behind these diverging trajectories. The first is the sheer scale of the capital in motion: the four largest hyperscalers spent on the order of $410 billion in 2025 and have guided toward roughly $700 billion in 2026, an increase of about three-quarters, the overwhelming majority of it aimed at AI.[12] Individually the figures are staggering—Microsoft guiding near $190 billion, Alphabet in the same range, Amazon larger still—each company committing in a single year what a decade ago would have funded a national infrastructure program. That is the number the embedded engineers are ultimately hired to justify. The second is a quieter race running beneath the model wars—the scramble to build proprietary silicon and shrink the tax paid to the bottom of the stack. Amazon's Trainium and Google's TPU are already at scale, capturing internally the margin that Microsoft, lacking a comparable in-house accelerator at volume, still pays out to Nvidia.[22] In the rent stack's own terms, the landlords are trying to become their own chip suppliers even as their tenants, as we will see, try to become their own landlords. Everyone is climbing toward someone else's margin.

FIG. 5—THE CAPITAL BEHIND THE COMPUTE

Combined hyperscaler AI capex, up ~77% in a year

Combined AI capital expenditure at the four largest hyperscalers rose from roughly $410B in 2025 to about $700B guided for 2026, an increase of about three-quarters. Compiled from Q1 2026 earnings guidance (Financial Times; company IR). 2025 actual; 2026 projected.

The point of the comparison is not to rank the three as investments. It is to show that "the landlords" are not one bet. For the enterprise buyer, forward-deployed engineering from these companies is the safest part of their AI strategy, not the fragile part—it is funded from operating cash flow measured in tens of billions, and it drives consumption across products they largely own. The fragility does not live in the deployment motion. It lives one layer up, in the question of whether the capital behind the compute ever earns out, and in what happens to an anchored cloud business if its anchor tenant stumbles or, as we will see, simply decides to stop renting.

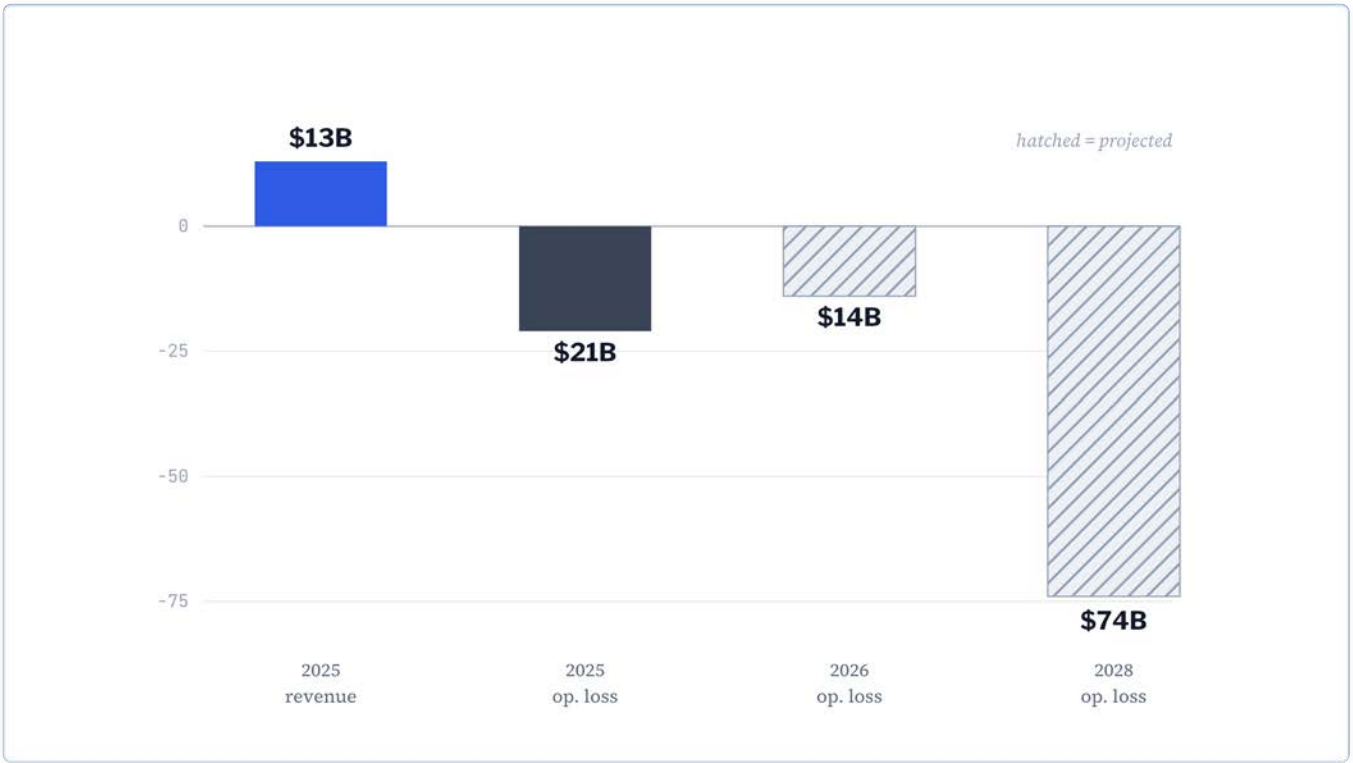
06 The Squeeze in the Middle

The tenants are the most interesting layer, because they are the ones the whole structure is squeezing, and because their response to that squeeze is the deployment rush in its purest form. A model maker pays a premium for scarce compute below and sells a commoditizing product above. That is a hard place to stand, and the income statements show it.

OpenAI is the starkest illustration. Financial statements that leaked in June showed the company generating $13.07 billion in revenue in 2025 against an operating loss of $20.92 billion, and a net loss of $38.53 billion once a roughly $41.5 billion non-cash charge tied to its corporate restructuring is included.[25] Its own internal projections, as reported, show losses continuing near $14 billion in 2026 and swelling toward $74 billion in 2028 before any turn to positive cash flow late in the decade.[26] A business losing money at that scale cannot wait patiently for enterprises to discover adoption at their own pace; it needs the meter running now, which is exactly what a deployment subsidiary is built to ensure.

FIG. 6—LOSING MONEY FASTER THAN IT MAKES IT

OpenAI: revenue against its operating losses



OpenAI booked $13B of revenue against a $21B operating loss in 2025, with internal projections showing losses near $14B in 2026 and $74B in 2028. 2025 from leaked financials (Fortune, 2026); 2026 and 2028 are internal projections. Excludes non-cash charge.

Not every tenant is bleeding equally, and the difference is the most hopeful fact in the sector. Anthropic, whose run rate climbed from roughly $30 billion in April to a reported $47 billion by late May, is projected to post its first operating profit in the second quarter of 2026, with gross margins its backers expect to approach 77 percent by 2028—a profile built on an enterprise and developer mix rather than consumer chat.[27] More telling for the squeeze thesis, analysts at SemiAnalysis estimate that Anthropic's inference gross margin rose from 38 percent to over 70 percent in about a year, as the economics of paid serving inflected sharply positive even while training and free-tier subsidy kept the blended figure under pressure.[28] If that inflection holds across the tenant layer, the squeeze is not permanent—the model makers may be climbing out of the middle under their own power, which would loosen the landlords' leverage over them rather than tighten it. The divergence between the two labs is not luck; it is mix. Anthropic draws the large majority of its revenue from enterprise and developer APIs, where usage is dense and willingness to pay is high, while OpenAI's revenue leans heavily on consumer subscriptions, where the marginal user is costly to serve and slow to monetize—the same structural reason a landlord's warehouse tenant is more profitable than its month-to-month storage renter.[27]

The other route out of the squeeze is to stop being a tenant, and OpenAI is running it in plain view. In roughly a year it went from Azure-exclusive to aggressively multi-cloud—a reported $300 billion, five-year compute arrangement with Oracle, a $38 billion deal with AWS expanding toward $100 billion, and its own $500 billion Stargate build-out, while an October 2025 renegotiation stripped Microsoft of its right of first refusal to supply OpenAI's compute.[29] Anthropic, for its part, deliberately multi-homes across AWS, Google, and Broadcom silicon, arbitrating its two largest cloud backers against each other for capacity and price.[30] The tenants, in other words, are not only climbing toward the enterprise through services; they are also trying to disintermediate the landlords beneath them by owning or diversifying their own compute. The most revealing detail in the entire tangle sits at Microsoft, which now ships Anthropic's Claude inside Microsoft 365 Copilot and runs it on Amazon's cloud—paying a rival landlord to serve a rival model inside its own flagship product.[31]

THE TENANT'S DILEMMA

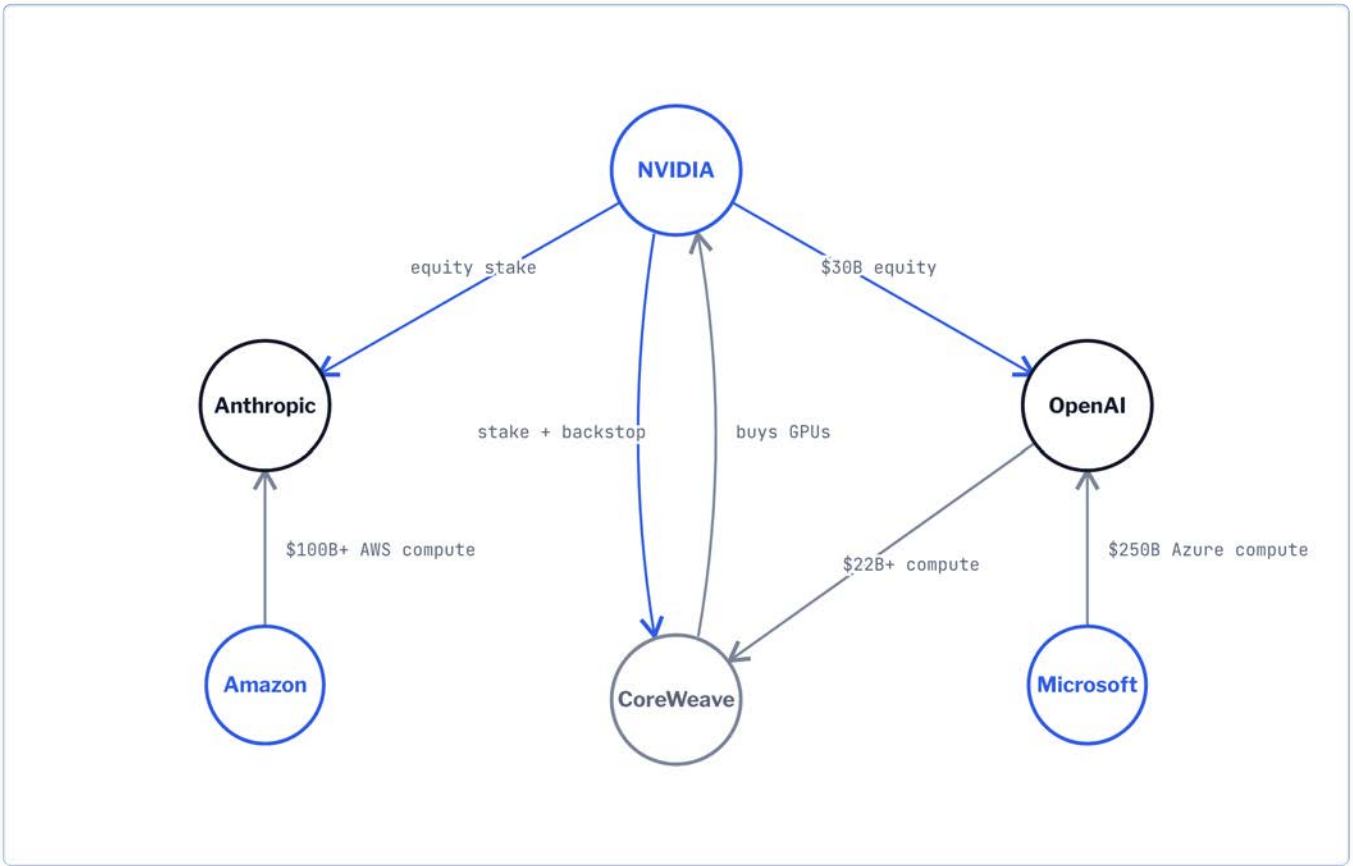
Squeezed between premium compute and commoditizing tokens, the model makers are climbing two ladders at once—toward the enterprise through deployment services, and away from their landlords through their own infrastructure. Both moves relieve the squeeze. Both also raise the capital intensity of an already unprofitable business.

07 Follow the Money in a Circle

The reason a stumble at any single layer matters to every other is that the layers are not merely stacked—they are wired together by an unusually dense web of mutual financing, one that analysts have taken to calling circular. Money invested at one layer reappears as revenue at another, and the same dollar can be counted as demand in more than one place. The clearest thread runs from chips to models and back. Nvidia announced in September 2025 what was widely reported as an investment of up to $100 billion in OpenAI; that figure was a letter of intent, never signed at that scale, and by early 2026 it had been restructured into a roughly $30 billion direct equity stake, with Nvidia's own chief executive conceding the full $100 billion was "probably not in the cards."[32] The pattern repeats across the industry. Amazon has committed up to $33 billion to Anthropic, which has in turn pledged to spend more than $100 billion on AWS and its Trainium silicon; Google holds an equity stake in the same company alongside a commitment of up to a million of its TPUs.[33] The structure is consistent: a landlord invests equity into a tenant, and the tenant pledges the cash back as compute spend, which the landlord books as revenue while also carrying the tenant's rising paper valuation as an investment gain.

FIG. 7—FOLLOW THE MONEY IN A CIRCLE

Equity flows down; the dollars flow back as compute



Equity flows down the stack and returns as compute spend: a dense web of mutual financing that couples chipmaker, model lab, cloud, and neocloud into a single system. Company disclosures and reporting, 2025–2026. Several arrangements announced or contracted, not finalized.

Beneath the hyperscalers sits a further, more precarious layer of leverage in the neoclouds—specialist GPU landlords like CoreWeave, which carried roughly $25 billion in debt secured largely against the chips themselves, with a majority of its 2025 revenue from Microsoft and a large share of its contracted future revenue from OpenAI.[34] The loop tightens here: Nvidia takes an equity stake in CoreWeave and backstops a slice of its capacity, CoreWeave borrows against Nvidia GPUs to buy more Nvidia GPUs, and then leases them to OpenAI, in which Nvidia also holds equity. What makes this more than a curiosity is where the risk has come to rest. The newest tranches of GPU-collateralized debt have been rated investment grade and sold to pension funds and insurers, which means a downturn in GPU economics no longer stays inside the technology sector—it reaches balance sheets that have nothing to do with AI and never chose to bet on it.[34]

None of this is fraud, and it is important not to overstate it. But it is fragile in a specific way: it concentrates the entire structure's stability on the continued solvency and continued spending of a handful of deeply unprofitable model makers. The scale of that mismatch is hard to hold in the mind: OpenAI alone has signed multi-year compute commitments reported to exceed a trillion dollars, spread across Oracle, Microsoft, and Amazon, against a business that booked $13 billion of revenue in 2025.[29] If one large tenant cannot pay, the damage does not stay contained—it propagates to the landlord that booked the revenue, the chipmaker that took the equity, the neocloud that borrowed against the hardware, and the lenders now holding GPU-collateralized debt that has, in the newest tranches, been rated investment grade and sold to pension funds and insurers. There is a live accounting dispute layered on top of this. The investor Michael Burry has alleged that hyperscalers understate the depreciation of AI hardware by extending its useful life on their books, thereby overstating earnings—a contested claim, and one he is positioned to profit from, but not one to dismiss on those grounds alone. The strongest rebuttal is empirical and comes partly

from the accused: older accelerators retain real economic value in inference and fine-tuning long after they leave the training frontier, and Amazon itself recently shortened a portion of its server schedule and took a charge, which is not the behavior of a company uniformly inflating.[35] The useful-life changes at the center of the dispute are not allegations; they are disclosed accounting decisions, each with a stated earnings effect.

PROVIDER	CHANGE TO SERVER USEFUL LIFE	EFFECTIVE	DISCLOSED EARNINGS EFFECT
Microsoft	4 → 6 years	FY2023	+$1.1B operating income (first quarter)
Google	4 → 6 years	Jan 2023	~$3.4B lower 2023 depreciation
Amazon	4 → 5 years (servers)	Jan 2022	+$2.2B operating results
Amazon (reversal)	6 → 5 years (subset)	Jan 2025	−$0.7B operating income; $0.92B charge
Meta	to ~5.5 years	Jan 2025	~$2.9B benefit

Whether those schedules are prudent or aggressive is the question Burry and his critics are actually fighting over, and it is not a trivial one—several billion dollars of reported operating income across these companies rests on the assumption that an AI server earns for five or six years rather than three. Amazon's own partial reversal, taken with a charge, is the sharpest evidence that the assumption is contested even inside the companies making it.[35]

Scarcity is not merely a line on a supplier's slide; it shows up in what compute costs to rent. Contract rates for Nvidia's H100 rose roughly 40 percent over the six months into early 2026, even as an older generation of chips was supposed to be getting cheaper, because the binding constraint had migrated from the processors themselves to the memory packaged around them and the power delivered beneath them.[36] There is also a harder physical limit closing in that no financing can paper over. The binding constraint on the build-out is increasingly not chips but power: of the roughly twelve gigawatts of US data-center capacity announced for 2026, a large share is delayed or cancelled, grid interconnection queues run years long, and transformer lead times stretch to half a decade.[36] All of which sharpens the question the venture investor David Cahn has been escalating for two years—the revenue the industry must eventually generate to justify its capital. His running tally rose from $200 billion to $600 billion and, by mid-2026, to roughly $1.5 trillion, the annual revenue required to justify a single year of the sector's capital expenditure.[37] That is the number the deployment rush is ultimately racing against.

08 The Loop That Eats Its Own Demand

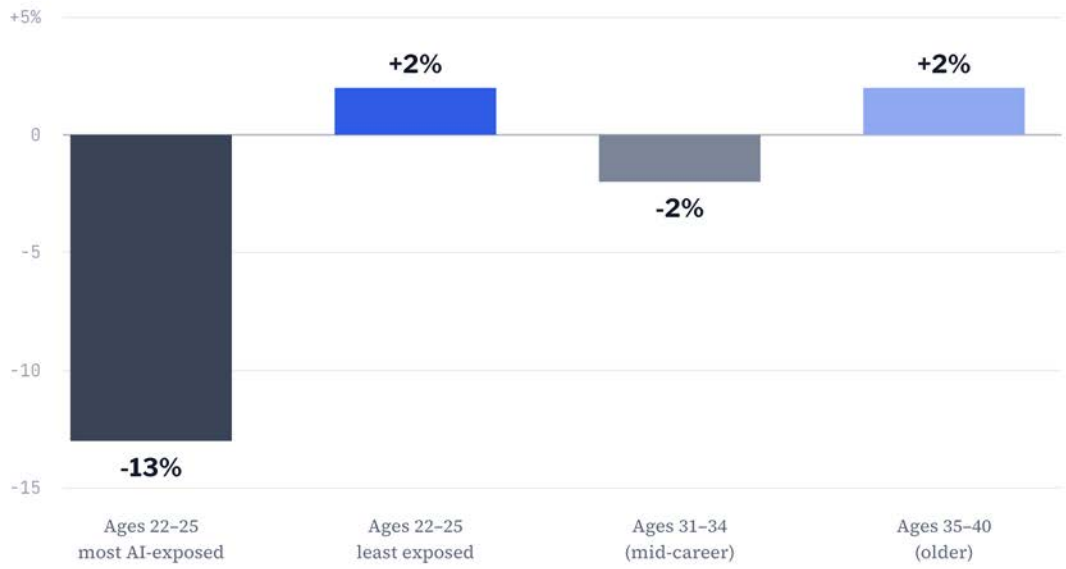
There is a longer-horizon risk that a careful executive should not wave away, because it runs underneath the consumption model that the entire stack depends on. Much of the demand justifying the build-out is premised, explicitly, on AI substituting for labor. Follow that premise far enough and it turns on itself: if automation removes enough wage income across the economy, it removes the consumer spending that enterprise revenue ultimately rests on, and softer enterprise revenue means smaller budgets for the very AI that set the cycle in motion.

This is no longer a speculative worry confined to opinion pages. The economists Brett Hemenway Falk and Gerry Tsoukalas formalized it this year in a model they call the AI layoff trap, in which firms rationally automate to cut costs while the aggregate effect erases the demand they sell into—"at the limit," they write, firms "automate their way to boundless productivity and zero demand."[38] The early labor data is at least consistent with the setup: researchers at Stanford and ADP found that workers aged 22 to 25 in the occupations most exposed to AI, software engineering among them, have seen employment contract sharply since 2022 even as less-exposed peers kept gaining—the entry-

level on-ramp thinning before aggregate employment moves.[39] And Daron Acemoglu, among the most rigorous of the skeptics, estimates AI's ten-year contribution to total factor productivity at well under one percent, a fraction of what the capital expenditure implies.[40]

FIG. 8—THE ON-RAMP THINS FIRST

Employment change by age and AI exposure



Employment for workers aged 22–25 in the most AI-exposed occupations has fallen sharply since 2022, even as less-exposed and older cohorts held or gained. Stanford Digital Economy Lab / ADP, 'Canaries in the Coal Mine?', 2026 (reported). Cohort estimates approximate.

The honest rejoinder is that this is a modeled risk rather than an observed collapse, and serious economists read the same facts the other way. Apollo's Torsten Slok invokes the Jevons paradox: as AI makes knowledge work cheaper, demand for that work may expand rather than shrink, pulling employment up with it, and he points to falling youth unemployment and record business formation as early evidence.[41] History supplies both patterns—the automated teller famously multiplied bank branches even as it changed the teller's job. Which pattern dominates this time is genuinely unknown. But an enterprise betting its cost structure on AI-driven labor substitution should at least notice that one of the demand pillars holding up the industry's forecast is the same force being sold to it as the reason to buy.

09 Read the Capital, Not the Keynote

The most reliable information about what these companies believe is not in their keynotes. It is in their capital allocation, which cannot afford to be aspirational. Read the industry's spending as a set of revealed preferences and a consistent message emerges, one the marketing rarely states outright: nobody at the frontier is betting that the model alone will hold the customer.

Start with the deployment ventures themselves. A company confident that its product sells itself does not stand up an army of engineers to install it, and it certainly does not give that engineering away, as Amazon does through its Innovation Center.[7] The act of embedding is an admission that the last mile is hard enough to require a subsidy.

OpenAI's financing makes the same admission in a different dialect: capitalizing its deployment subsidiary with outside money at a guaranteed minimum investor return of 17.5 percent is the structure of an entity whose sponsors wanted their downside protected, not the posture of a business certain the demand will arrive on its own.[4]

Then read the hedges each layer is placing against its own position. Microsoft, whose entire AI narrative once rested on its privileged access to OpenAI, now ships a rival's model inside Microsoft 365 Copilot and runs it on a rival's cloud—an admission, priced in real dollars, that model-agnosticism beats model-ownership once the models converge.[31] The model makers, for their part, are trying to stop being tenants: OpenAI's Stargate build-out is an attempt to own its compute rather than rent it, a multi-hundred-billion-dollar vote against depending on any single landlord.[29] And the landlords are racing to build their own silicon, Trainium and TPU, to stop paying full freight to the bottom of the stack.[22] Each move is a company spending heavily to reduce a dependence it once accepted. In aggregate they describe an industry whose most informed participants are all quietly diversifying away from the very thing they publicly celebrate.

Even the confident bulls concede the premise on their way to a bullish conclusion. Ben Thompson, who argues persuasively that this is a durable platform shift rather than a bubble, still grants that base models are commoditizing, which is exactly why he expects the value to migrate to integration and the harness around the model rather than the model itself.[42] That is the same migration Ode, Frontier, and the Deployment Company embody. The optimistic and the skeptical readings converge on a single point: the defensible ground has moved from the model to the deployment.

> THE REVEALED PREFERENCE
>
> If the companies with the most information and the most at stake are all spending to reduce their dependence on the model layer, an enterprise building its strategy on a single model is taking a position its own vendors have declined to take.

For an enterprise leader, that convergence is the most actionable signal in the market. You do not have to resolve the bubble debate or forecast which lab wins to act on it. You need only notice that the people selling you model-dependence are themselves buying model-independence, and structure your own commitments accordingly.

10 Underwriting the Last Mile

None of this argues against adopting AI, and none of it argues against forward-deployed engagements. The individual productivity gains are real, the deployment help is often genuinely useful, and refusing it on principle would be its own kind of malpractice. The argument is narrower and more actionable: an enterprise leader should underwrite a deployment offer as the financial instrument it is, not accept it as the free favor it is dressed up to be. A handful of disciplines separate the buyers who will benefit from the ones who will find themselves holding a dependency.

01 Price the withdrawal, not the arrival.

The seductive part of a deployment offer is the arrival—engineers on site, momentum, a system taking shape. The part that determines your outcome is the withdrawal. Before signing, establish what month eighteen looks like. A durable partner will describe a path to your self-sufficiency and rising internal capability; a subsidy will describe an ever-deepening presence. Ask directly who funds the engineers today and what your unit economics become when that funding normalizes to market rates. If no one can answer, you have found the answer.

02 **Measure the ratio that reveals the engine.**

Track revenue per embedded engineer and engineers required per outcome across the life of the engagement. If the vendor genuinely needs fewer people to deliver more value as time passes, you are buying into a maturing product. If the headcount never falls, you are renting a capability you were told you were building. This is the single most diagnostic number in the relationship, and most buyers never compute it.

03 **Underwrite counterparty risk explicitly.**

Concentrating your AI strategy on a single vendor means taking that vendor's balance sheet, compute commitments, and position in the financing web onto your own risk register, whether or not you have written it down. The landlords with diversified, cash-generative businesses are one kind of counterparty; a deeply unprofitable model maker whose survival depends on the next funding round is another. Neither is disqualifying. Both should be priced, and the price should show up in how much of your operation you are willing to make dependent on them.

04 **Keep the models substitutable.**

The most protective architectural decision available to you is also the one the vendors reserve for themselves—even Ode operates multi-model, and Microsoft ships a rival's model inside its own product. Favor contracts and system designs that let you swap the model layer with manageable effort, so that a shock to one provider is a migration rather than a crisis. Substitutability is the practical expression of the paper's central finding: if the model is no longer the moat, you should not build your moat on any single one.

05 **Separate the productivity case from the demand case.**

Justify your AI spending on efficiency and quality you can observe inside your own operations—cycle times, defect rates, throughput per team—not on a macro story about ever-expanding demand that may be partly self-undermining. The gains that show up in your own telemetry are yours to keep regardless of how the broader cycle resolves. The gains that depend on the industry's forecast being right are someone else's bet, and you should not fund it with your capital budget.

> THE UNDERWRITING PRINCIPLE
>
> Treat every deployment offer as a loan you are being extended, not a gift you are being given. Read the terms, price the withdrawal, and keep the collateral—your model choice, your internal capability, your data—in your own hands.

11 How This Resolves From Here

Three broad paths are visible from mid-2026, and an enterprise decision made now should be legible under all of them. In the first, the consumption catches the capital: enterprise returns finally materialize, inference margins across the tenant layer continue the inflection Anthropic is already showing, and forward-deployed engineering matures into a genuine bridge that carries customers to self-sufficiency. In that world the durable engines win, the crutches are quietly retired, and the buyers who insisted on substitutability and measured the ratio simply have a head start. In the second, the capital outruns the consumption: the revenue gap stays open, the financing web transmits a stumble from one large tenant outward, and the subsidy that made many deployments pencil out is

withdrawn to protect margins. In that world, the buyers who priced the withdrawal and underwrote their counterparties absorb a manageable disruption while their less careful peers discover how much of their operation had been resting on someone else's balance sheet. In the third and likeliest, the resolution is uneven—Google-like layers prove durable while thinly capitalized ones do not, some deployments stick and some evaporate, and the outcome for any given enterprise depends less on the macro verdict than on which specific engine it happened to buy.

That is why the framework matters more than the forecast. No one reading this can know which path arrives, but every path rewards the same behavior: seeing the stack clearly, knowing which layer a given vendor is defending, and refusing to take on a dependency you have not priced. The vendors have already told you what they believe, not in their keynotes but in their capital allocation—that the model is no longer the moat, and that the ground worth holding is the last mile into your business. They are racing to occupy it. The only question left for you is whether the party arriving at your door is bringing you a partnership or fitting you with a leash, and by now you have the means to tell the difference before you sign.

ABOUT THE AUTHOR

Joshua Davis is a Lead AI GTM Strategist for Americas—Enterprise at GitHub. Across four decades, he has built software, architected enterprise cloud, and led teams of all sizes. Today, he helps organizations do far more than sell AI: he helps them implement it, manage it, and govern it at enterprise scale. He advises leaders who need AI strategy, adoption, and governance to move as one.

To contact or request advisory options: jdav.is/advisory

REFERENCES

42 sources · company filings, earnings releases, analyst research, and reporting, 2024–2026.

1 "Anthropic, Blackstone, and Hellman & Friedman Introduce Ode with Anthropic, an Enterprise AI Services Firm," *Business Wire*, July 15, 2026; "Building a New Enterprise AI Services Company," *Anthropic*, May 4, 2026. The $1.5B figure and ~100-engineer headcount are as reported by *TechCrunch* and *Fortune*.

2 Julie Bort, "Anthropic, Blackstone Bet the Next Trillion-Dollar AI Business Is Implementation, Not Models," *TechCrunch*, July 15, 2026.

3 Ibid.

4 "OpenAI Launches The Deployment Company," *OpenAI*, May 11, 2026; "OpenAI Launches Professional Services Business with $4B Investment," *SiliconANGLE*, May 11, 2026. The guaranteed 17.5% minimum return is per *SiliconANGLE* and *Axios*.

5 "Introducing Microsoft Frontier," *Microsoft*, July 2, 2026; Todd Bishop, "Microsoft Announces $2.5B Frontier Company to Embed AI Engineers Inside Customers," *GeekWire*, July 2, 2026.

6 "Google Cloud Commits $750 Million to Accelerate Partners' Agentic AI Development," *Google Cloud Press Corner*, April 22, 2026.

7 "AWS Announces Generative AI Innovation Center," *Amazon*, June 22, 2023; "AWS Doubles Investment in AWS Generative AI Innovation Center," *AWS Machine Learning Blog*, July 15, 2025.

8 "Announcing Deloitte Forward Deployed Engineering," *Deloitte*, December 1, 2025; "Accenture Launches Microsoft Forward Deployed Engineering Practice," *Accenture Newsroom*, March 18, 2026.

9 "Palantir Reports Q1 2026 Results," *Palantir Technologies / U.S. SEC*, May 4, 2026; Steve Banker, "Palantir and Forward Deployed Engineering," *Forbes*, July 10, 2026.

10 "NVIDIA Announces Financial Results for Fourth Quarter and Fiscal 2026," and "...First Quarter Fiscal 2027," *NVIDIA Newsroom*, 2026.

11 Chris Zeoli, "Who Captures Value in AI Infrastructure?" *Data Gravity*, July 2026; "AI Value Capture: The Shift to the Model Layer," *SemiAnalysis*, May 1, 2026.

12 "Big Tech's AI Spending and the Power Bottleneck," compiled from Q1 2026 earnings and Sightline Climate / Bloomberg reporting, 2026; "Gartner Predicts 40% of AI Data Centers Will Be Power-Constrained by 2027," *Gartner*, 2026.

13 "LLM Inference Price Trends," *Epoch AI*, 2026 (median decline in cost per fixed task; reasoning models excluded owing to higher token consumption).

14 "2025: The State of Generative AI in the Enterprise," *Menlo Ventures*, 2025.

15 Aditya Challapally et al., "The GenAI Divide: State of AI in Business 2025," *MIT NANDA*, August 18, 2025, as reported in *Fortune*.

16 "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027," *Gartner*, June 25, 2025.

17 Analyst framing; see *Constellation Research* and *TSIA* commentary, 2026; "Forward Deployed Engineers: The Last Mile of the AI Value Chain," *TBR*, July 10, 2026.

18 Abhishek Singh, "Palantir: Inside the Category of One," *Everest Group*, November 5, 2025.

19 "Forward Deployed Engineers: Promise and Peril in AI Deployments," *Constellation Research*, 2026.

20 Charlie Gottdiener, CEO of Anaplan, quoted in Banker, "Palantir and Forward Deployed Engineering," *Forbes*, July 10, 2026.

21 "Alphabet Q1 2026 Earnings Release," *Alphabet Inc.*, April 2026; "Alphabet Q1 FY2026: AI Demand Surges as Cloud Capacity Caps Growth," *Futurum Group*, 2026.

22 "Amazon Q1 2026 Earnings Release," *Amazon.com, Inc.*, April 2026; "Amazon Q1 FY2026: AWS Momentum Builds," *Futurum Group*, 2026.

23 "Microsoft Q3 FY2026 Results," *Microsoft / AlphaStreet*, April 2026; "Microsoft FY26 Q3 Results Explained," *LicenseQ*, 2026.

24 Om Malik, "What Microsoft's 10-Q Says About OpenAI," *om.co*, May 1, 2026 (estimate; not disclosed by Microsoft).

25 "OpenAI's Financials, Leaked," *Fortune*, June 16, 2026; Ed Zitron, "OpenAI's Financials," *Where's Your Ed At*, June 2026.

26 "OpenAI Cash Burn and Loss Projections," *Fortune*, November 12, 2025 (internal projections; forward estimates).

27 Paulo Carvão, "Anthropic, OpenAI, and the Road to Enterprise AI Profitability," *Forbes*, May 21, 2026; "Anthropic Says It Hit a $30 Billion Revenue Run Rate," *VentureBeat*, April 2026; "Anthropic Raises Series H," *Anthropic*, May 28, 2026.

28 "AI Value Capture: The Shift to the Model Layer," *SemiAnalysis*, May 1, 2026 (Anthropic inference gross margin 38% → 70%+; analyst estimate).

29 "The Next Chapter of the Microsoft–OpenAI Partnership," *Microsoft*, October 28, 2025; "OpenAI Signs $300bn Cloud Deal with Oracle," *Data Center Dynamics*, September 2025; "OpenAI and Amazon Announce AWS Cloud Deal," *CNBC*, November 3, 2025; "Five New Stargate Sites," *OpenAI*, 2025.

30 "Expanding Our Use of Google Cloud TPUs and Services," *Anthropic*, October 23, 2025; "Anthropic, Google, and Broadcom Compute Partnership," *Anthropic*, 2026.

31 "Expanding Model Choice in Microsoft 365 Copilot," *Microsoft 365 Blog*, September 24, 2025; "M365 Copilot Adds Choice and Risk with Anthropic's Claude," *Directions on Microsoft*, 2025.

32 "OpenAI and NVIDIA Announce Strategic Partnership to Deploy 10GW of NVIDIA Systems," *NVIDIA Newsroom*, September 22, 2025; "Nvidia–OpenAI Deal Not Signed Yet," *Fortune*, December 2, 2025.

33 "Anthropic and Amazon Expand Compute Partnership," *Anthropic*, 2026; "Amazon to Invest Up to $25 Billion More in Anthropic," *CNBC*, April 20, 2026; "Expanding Our Use of Google Cloud TPUs," *Anthropic*, October 23, 2025.

34 "CoreWeave Q1 2026 Results," *CoreWeave, Inc. / U.S. SEC*, 2026; Phoebe Liu, "CoreWeave Depends on OpenAI—What If the ChatGPT Maker Can't Pay Up?" *Forbes*, April 29, 2026.

35 "Big Short Investor Michael Burry Accuses AI Hyperscalers of Artificially Boosting Earnings," *CNBC*, November 11, 2025; hyperscaler 10-K depreciation footnotes, 2022–2026; "The Big Short Investor, Depreciation, and the AI Bubble, Explained," *Fortune*, November 13, 2025.

36 "US AI Data-Center Delays and Cancellations," compiled from Bloomberg and Sightline Climate reporting, May 2026; "H100 Rental Prices Surge Nearly 40% in Six Months," *SemiAnalysis GPU Cloud Index*, 2026.

37 David Cahn, "AI's $1.5T Question," *Sequoia Capital / dcahn.substack.com*, July 2026 (escalated from earlier "$200B" and "$600B" formulations; projected).

38 Brett Hemenway Falk and Gerry Tsoukalas, "The AI Layoff Trap," working paper, 2026 (arXiv:2603.20617); coverage in *The London Economic*, 2026.

39 Erik Brynjolfsson et al., "Canaries in the Coal Mine? Six Facts About the Recent Employment Effects of Artificial Intelligence," *Stanford Digital Economy Lab / ADP*, 2026, as reported in *Fortune*, June 27, 2026.

40 Daron Acemoglu, "The Simple Macroeconomics of AI," *NBER Working Paper 32487*, 2024.

41 Torsten Slok, "Will AI Kill Jobs? Why Not—Jevons Paradox," as reported in *Fortune*, April 28, 2026.

42 Ben Thompson, "Agents Over Bubbles," *Stratechery*, 2026.

