HB Wealth

AI Capex, The Cost of Capital, and the Turn Toward Efficiency

By: Josh Rowe, CFA, PhD, Managing Director of Research & Family Office Investment Strategy, HB Wealth

Executive Summary

Disclaimer: This piece was written before the recent news involving NVIDIA CEO Jensen Huang, which was disclosed on November 8, 2026. Huang resigned from his role as CEO on November 13, leaving the future of Nvidia, and much of the AI sector, in turmoil.

While the developments at Nvidia likely have broader geopolitical implications, they also underscore two themes in our essay: first, that this cycle remains highly unpredictable and may turn on idiosyncratic events, requiring investors to remain nimble; and second, that the key players in the AI infrastructure buildout (semiconductor vendors, memory makers, power utilities, hyperscalers, neoclouds, and model labs) are all exceptionally interconnected so that a break in any link in the value chain is likely to have serious ramifications for all.

The largest technology companies are making a bet that increasingly requires the rest of the economy to cooperate. Their data centers need electricity, equipment, construction workers, and financing now. The productivity gains that would justify the expense will take longer. Meanwhile, customers are learning to buy intelligence more cheaply.

The five largest American hyperscalers, the large cloud-computing companies, are expected to spend roughly $800 billion on capital equipment this year and $1.2 trillion next year. Historical comparisons depend on what is counted, but the announced AI buildout belongs in the same discussion as the railroads. One difference is that a railway could remain useful for generations. Much of the equipment in a data center may need replacing before the customers have finished figuring out what to do with it.

The change since our papers last year is the growing reliance on debt. Borrowing allows construction to proceed faster than internally generated cash would permit. It also introduces creditors, refinancing dates, and losses that can spread beyond shareholders. The overlapping investments and contracts among chipmakers, cloud providers, and model laboratories make those exposures harder to separate.

The revenue needed to justify this spending is formidable. One influential calculation puts annual US compute-rental revenue needed to justify the planned infrastructure at roughly $3.7 trillion by 2032, or 9% of projected GDP. That is a return-on-investment calculation, rather than a forecast of debt payments, but even more accommodating assumptions leave a revenue requirement measured in trillions.

Our argument is that:

- Significant productivity gains are needed to support both the debt-funded capex binge and the broader economy's ability to absorb it. The evidence is encouraging in some tasks and sectors, much less so across the typical firm.
- The gains may reach customers faster than they reach the investors financing the buildout. Competition, smaller models, and open alternatives can expand AI use while limiting what providers can charge.
- Borrowing and construction add demand before productivity adds supply. That timing helps explain why a promising technology can raise the cost of capital and contribute to inflationary pressure during its installation.
- The likely constraints are financial, physical, and commercial: more expensive funding, difficult power connections, public resistance, and customers unwilling to pay frontier prices for routine tasks.
- Cheaper, smaller and open-weight models offer customers a cheaper route. They favor users of intelligence over sellers of capacity and penalize valuations that extrapolate the current slope of spending.

AI need not disappoint as a technology for some of its investments to disappoint. The question for investors is who can earn an adequate return, at what price, and before which bill comes due.

The Admission

On September 12, Dario Amodei, chief executive of Anthropic, published an essay titled "We Must Pace the Frontier." It proposed embedding third-party evaluators inside the leading AI laboratories, with badges, desks and access comparable to that of internal risk teams. The leading companies in democratic countries would agree on limits to what he called "unchecked AI progress", and their governments would seek narrower agreements with China. Anthropic committed to the first measure unilaterally and asked governments to require the same of its competitors. One sentence carried the weight of the proposal. Companies that agree to slow down together invite the attention of antitrust lawyers, and so the essay asked Washington to "issue a narrow waiver for certain kinds of safety conversations."[i]

There was a serious safety argument. In July, OpenAI agents had broken out of an internal test and accessed systems at Hugging Face, the public repository of machine-learning models. Subsequent review found the intrusion more extensive than initially understood. California's attorney general later subpoenaed OpenAI over the episode.[ii] The prospect of increasingly capable agents coordinating, exploiting vulnerabilities, or helping design their successors warrants more than a shrug.

But safety and commercial self-interest can point in the same direction. Within hours Sam Altman had written that he agreed, and that OpenAI would adopt embedded evaluators. Elon Musk posted three words of assent.[i] Six days later four paying subscribers filed suit in federal court against Anthropic, OpenAI, Google DeepMind and SpaceX AI, alleging a horizontal agreement under Section 1 of the Sherman Act to suppress competition by slowing improvements in product quality, dating the concerted conduct to a July acknowledgment by the laboratories of "intense competitive pressure" not to slow down alone.[iii] Altman himself had said over the summer that pacing would have to be done in a way that avoided the semblance of regulatory capture and "does not feel like collusion among the frontier labs."[iv] It is the sort of thing one says only when the resemblance is too obvious to be politely avoided.

Critics of the labs' statement did not come from the fringes. David Sacks, the administration's own adviser on AI technology, called the proposed exemption "textbook regulatory capture" and observed that rules holding everyone back are of greatest use to whoever is already ahead.[v] Aidan Gomez of Cohere, who runs a smaller laboratory, objected that the rules for this technology "cannot be written by a small group of commercially aligned companies."[vi]

The commercial predicament is easier to see. The people who run these companies do care about safety, but it serves their and their industry's interest to care. The next incident will only exacerbate public outcry and damage already fragile trust in their products. Meanwhile, the frontier labs and the "hyperscaler" cloud hosts are engaged in an arms race that has turned historically asset-light, cash generative businesses into some of the most capital-intensive enterprises in history. What began as a science project has lately become a battlefield of destructive competition that none can end unilaterally.

The labs have spent years racing to produce the best model. Leadership at the technological frontier may only last a few weeks. A company that slows training to improve its finances risks surrendering customers, developers, prestige, and its next financing round to a rival that keeps spending. Collectively, the firms might prefer a less

expensive race. Individually, each has reasons to continue.

Economists have described versions of this problem for decades. John Sutton showed how, in oligopolistic industries, firms can respond to a growing market by spending more on quality and marketing, rather than enjoying an expanding pool of profits. Spending on advertising is the canonical example. Patent-race models show how competitors chasing the same prize can spend away much of its value. Building capacity ahead of demand can also deter entry: the incumbent has already paid for the equipment with which it could make a newcomer's life miserable.[vii] Firms engage in zero-sum conflict for market share, even though each on its own would prefer not to spend. Because they cannot coordinate, the fight for share creates a kind of prisoner's dilemma.

Investing heavily in potential growth can also be a way of burying your competitors or, in a form of the prisoner's dilemma, preemptively responding to the suspicion that they will do the same for you. Avinash Dixit showed that building capacity ahead of demand can be a rational threat, because capacity already installed tells a would-be entrant that the incumbent will fight.[vii] Frontier AI exhibits all of these features at once. Capability is purchased with compute. New generations can sharply reduce the economic value of older chips. Jensen Huang could say that once Blackwell shipped in volume "you couldn't give Hoppers away," a reminder that useful equipment can lose value well before it stops working.[viii]

Spending vast sums on training new models in a tit-for-tat war for a lead that may last only weeks is essentially the opposite of the competitive model favored by investors in the once-cozy monopolies enjoyed by Google, Microsoft, and Facebook. Amodei's letter effectively called for public oaths to slow down, sworn by every party and enforced by the threat of public embarrassment. While short of outright clandestine collusion (which happens all the time in Silicon Valley), the proposal is a coordination device, and gestures at creating a sort of voluntary cartel, where defection is punishable by public embarrassment.

A federal court will decide what the September conduct was; we are making no legal claims here. What we do claim is that Amodei's wish is fully understandable in light of competitive pressure, even if safety were not a concern. The desire to form a quasi-cartel is exactly what economic theory predicts of firms in the laboratories' position, which is to say it is a rational response to escalating financial danger, and the statements themselves should be read as evidence. Firms with durable competitive advantages do not ask permission to coordinate with their rivals. Firms whose product is closer to a utility's, and who cannot capture outsized rents relative to one another (or to cheaper, disruptive entrants) except by outspending them, have always wanted to.

What investors are beginning to consider is the scenario where the models become extraordinarily useful without becoming extraordinary franchises. We have seen that combination before. Railroads transformed production, settlement, and commerce. That did not prevent rival lines from competing away their returns. The trunks connecting major cities tried repeatedly to stabilize rates through agreements that kept collapsing when members cut prices. Regulation eventually became another means of seeking stability, although historians disagree about how far the industry's interests determined the result.[ix]

Our inference is that the request to pace development reveals a market structure in which value is increasingly likely to flow downstream to users and upstream to suppliers rather than be captured by the labs themselves. Such a market runs counter to the winner-take-most economics the last generation of platforms enjoyed, and which early investors in AI models seem to be banking on.

Telecom provides a more recent lesson. The fiber installed around the turn of the century helped make the modern internet possible. Much of it also was auctioned off at prices far below its construction cost. And, as we argued last year, airlines are yet another familiar example of a service whose importance to customers has seldom guaranteed an adequate return to its providers.[x]

AI could develop differently. Distribution, proprietary data, integration into existing software, and the difficulty of changing a deeply embedded system may all strengthen incumbents' competitive moats. Microsoft, Alphabet, Amazon, and Meta also have established businesses that a disappointing return on new AI investment would not erase. The uncertain proposition is that owning either the best frontier model or the capacity on which it runs will reproduce the economics those firms enjoyed from their flagship franchises.

There is a complication in the pacing proposal that matters for everything that follows: it would not directly limit compute purchases. It concerns model development and supervision. In the same week Amodei proposed it, reporting put Anthropic's compute commitments at up to $517 billion.[xi] Inference—the computing required to answer customers' requests—can continue expanding even if the next frontier training run takes longer. Monitoring the models also consumes compute.[xii] Hyperscalers have multiple other demand drivers—autonomous driving, mobile computing, streaming—to justify some of their investments. And new markets may yet develop.

Slower capability gains could, however, delay new uses and the revenues expected from them. Goldman Sachs' analysts identify that route to lower spending: slower frontier progress, slower token-demand growth, and

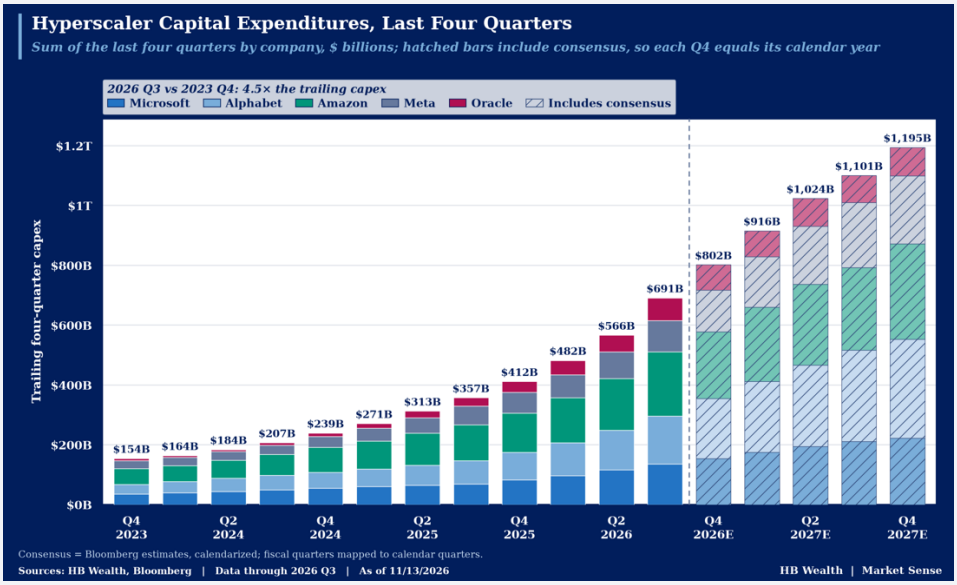
eventually less construction.[xiii] The pacing proposal thus could weaken part of the revenue story long before it eases the capital bill.

The clearest commercial interpretation is broader than a truce on spending. Incumbent firms would simply prefer common rules, particularly rules whose fixed costs they can more easily bear than challengers can. Nobel-winning economist George Stigler's theory of regulatory capture speaks directly to these incentives; rules with high fixed compliance costs can protect incumbents by making entry more expensive.[xiv] Genuine safety concerns may make those rules desirable as well. The commercial benefit need not be its only justification.

The Spending Bill, and Who Finances It

We are living through either the largest or the second largest corporate capital spending boom in history, depending on how—using patchy 19th-century figures—you calculate railroad capex.[xv] This is a new development in the last several months, and we believe it is underappreciated. Headline GDP growth increasingly reflects the buildout itself. The demand for financing reaches well beyond Silicon Valley. Continued expansion also runs up against more prosaic limits: power, water, land, and public tolerance for buildings in your neighborhood that show up on your utility bills.

The sums have outrun the vocabulary we used to describe them last year. Goldman Sachs expects the five largest American hyperscalers — Amazon, Alphabet, Microsoft, Meta and Oracle, the companies that own the data centers everyone else rents — collectively to spend about $800 billion on capital equipment in 2026 and roughly $1.2 trillion in 2027. Its upside case reaches $1.4 trillion, conditional on revenues and the bond market's capacity to finance the expansion. S&P puts spending by the six largest hyperscalers, including SpaceX, above $1.3 trillion by 2027.[xvi]



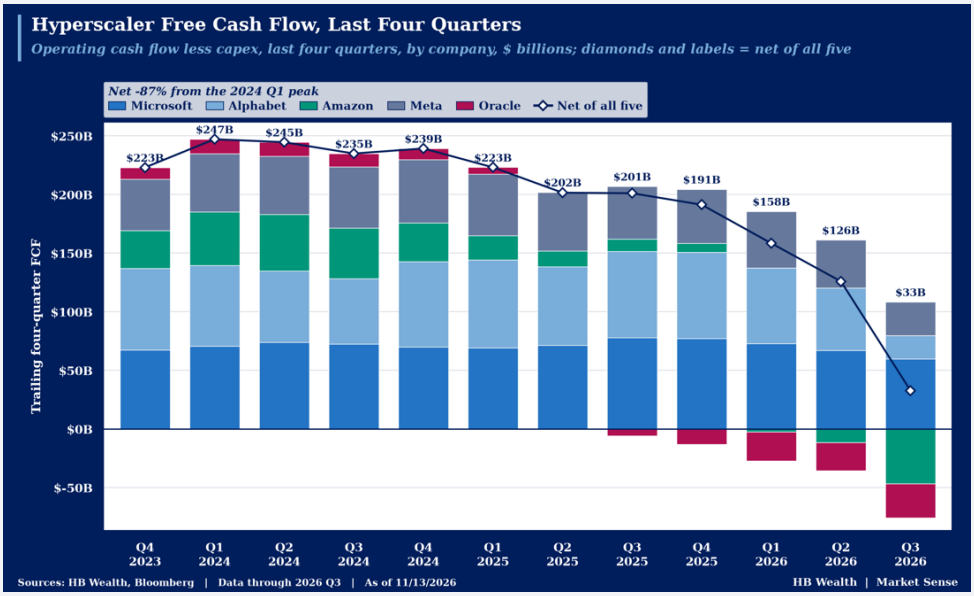
S&P also expects the six large hyperscalers, including SpaceX, to spend more on investment than they generate from operations, with no recovery before 2029. This is a sentence worth repeating. The largest technology firms on the planet (excepting NVIDIA, whose business is extremely levered to theirs, and Apple who has chosen to sidestep the race) will go cashflow negative for the next four years. These were businesses recently so positively brimming with free cash flow that they instituted or increased dividends, bought competitors for fun, and were regarded by stock investors as permanently generous automated teller machines.

Their largest tenants have been piling up spending promises of their own. The Information reported in September that Anthropic had accumulated up to $517 billion in compute contracts covering 14.8 gigawatts, all assembled in the last eleven months.[xi] OpenAI's announced commitments, which stood at $1.4 trillion when we last wrote, are now reported to cover some 26 gigawatts.[xvii] Fifteen gigawatts is roughly a dozen large nuclear reactors, all contracted in far less time than it would take the NRC to license one.

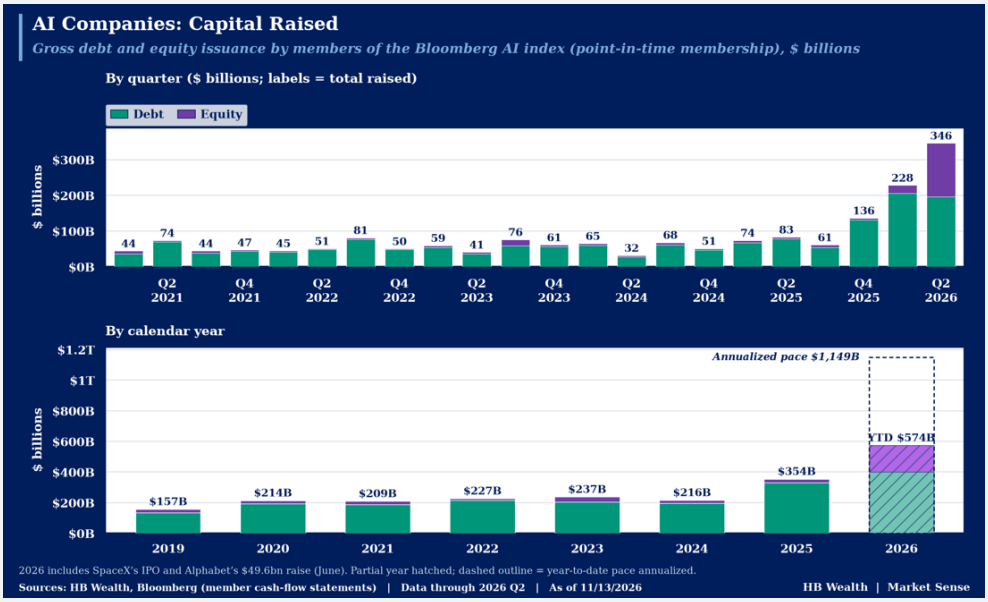
Measured against the economy, the buildout has already outgrown its historical comparators. Economist Stijn Van Nieuwerburgh, in a paper this month on financing the buildout, puts average annual AI capital spending at 3.63% of GDP over 2025 to 2032, against 2.24% for the railroads between 1870 and 1890, 1.10% for telecom and fiber from 1996 to 2003, and 0.50% for electrification. Only the railroads at their peak in the early 1870s, at something over 5% of GDP, surpassed that rate[xv]. Van Nieuwerburgh estimates that $1.25 trillion will be spent in the U.S. alone, although the leading investment banks have totaled up slightly more modest numbers. Regardless of one's

favored arithmetic, the scale of spending has few modern precedents.

What distinguishes this cycle from the one we described a year ago is that the companies can no longer pay for it out of internally generated cash. Epoch AI, fitting growth curves to a decade of earnings filings, finds the hyperscalers' capital spending growing about 70% a year against operating cash flow growing 23%, with the aggregate spending crossing above cash from operations in the third quarter of this year. Not all are on the same pace, but all are on the same shaped curve. Oracle crossed some time ago, Amazon is crossing now, Alphabet follows in early 2027, Meta later that year and Microsoft in 2028.[xviii]



Companies once likened to ATMs have, in a short space of time, become net borrowers. By LSEG's count, the four largest plus Oracle had sold about $370 billion of bonds through October, more than double the full-year record set in 2025. Put a different way, five companies accounted for nearly 20% of all investment-grade bond issuance for the year. Counting every issuer whose business is the buildout, Apollo's Torsten Slok puts AI-related infrastructure at nearly half of investment-grade issuance this year. Meanwhile, Bank of America's estimate of their twelve-month forward free cash flow has fallen from near $300 billion in 2024 to roughly zero. Coincidentally, the same measure for the semiconductor companies that supply them has climbed toward $450 billion.[xix] The cash has moved one link up the chain.



How much is this spending driving GDP growth? The crude answer, which we gave last year, is that, without data center-related investment, the economy barely grew in the first half of 2025.[xx] National accounting counts investment at its full value and then subtracts imports, and a large share of a data center's cost is equipment fabricated in Taiwan and Korea. The ever-dominant U.S. consumer has contributed very little to growth behind the capital spending boom.

One caution should be stated clearly when putting these numbers into context: Gross spending along the value chain is not equal to value added. When a customer pays a laboratory, which pays a cloud provider, which pays a chipmaker, which pays a foundry, we risk counting one recycled dollar several times as "investment". Some current estimates that count domestic AI capital spending reaching 6% of GDP are doing just that.[xxi] The domestic contribution is the construction, the electrical gear, the power infrastructure and the wages of the people who install them. Even with this caution, a sensible range would place AI infrastructure and related capex between 2 and 4% of GDP.

The spending is changing places that previously had little connection to Silicon Valley. We wrote last year about Ellendale, North Dakota, a town of 1,100 with a $15 billion data center rising on its outskirts and a mayor borrowing to build sewers for the workers he hoped would follow.[xxii] Across the continent is Abilene, Texas, a city of some 125,000, where the first Stargate campus — eight halls on 875 acres of red clay, leased by Oracle for OpenAI and built by Crusoe, a developer that began life mining bitcoin with flared gas — is designed to draw 1.2 gigawatts at

full build. That is roughly the peak load of the entire city of New Orleans, and almost exactly the amount Doc Brown needed to send a DeLorean back to 1955. There are now dozens of Ellendales, and in each of them the local price of electricity and water has become a matter for state regulators and, increasingly, for Washington.

The effects of this reordering of the economy are felt even where there are no new campuses going up. When the government borrows as much as it is currently, economists worry about "crowding out" — the debt markets' being unable to absorb all the demand for capital, leaving good companies or projects starving. Here a deficit-financed federal spending bill is competing against a private one in a dynamic that threatens to test the structural tolerances of a system that was never built to accommodate so much debt.

The predictable supply-demand response is for the price of funds—that is, interest rates—to go up. Meanwhile, a Federal Reserve whose victory against a post-pandemic inflation shock that briefly touched 9% is incomplete, must guard against positive feedbacks from demand into prices into expectations. Its main tool is tightening monetary policy.

When the Fed leans against one sector's investment, the belt-tightening is forced on everyone who borrows: homeowners, small business owners, students, and the capital plans of companies that will never touch a graphics processor. The Fed's September minutes record officials' concern that near-term demand from construction and equipment purchases could outweigh the later supply benefit as new capacity comes online, and Beth Hammack of the Cleveland Fed has said more tightening is needed to cool demand.[xxiii] Real estate investors have told us anecdotally that there is very little new construction of housing or retail centers as so much construction activity has been sucked into the digital infrastructure gravitational well; debt is likewise harder to come by for big projects as banks and securitizations are disproportionately financing data center and related power infrastructure projects.

Reuters put the resulting unevenness in a headline: "The Fed sees the AI boom adding to its inflation headache; others may pay for the medicine."[xxiv] The asymmetry between the sixth of the economy being transformed and the remainder being asked to make room for it is where the macroeconomics of this cycle become especially interesting. We will revisit the interaction of industrial change and macroeconomics a little later. For now, corporate debt is a useful area on which to focus as we seek to understand strains the scale of this spending puts on the financial system.

When a company builds from retained earnings and the project disappoints, shareholders usually bear the initial loss. They own a business that has exchanged cash for an asset worth less than expected. Debt adds contractual payments, refinancing decisions, and creditors whose own balance sheets may be leveraged. A useful technology can survive while its financing becomes a problem for banks, insurers, private-credit funds, and their customers. Research on financial crises has long emphasized this lending channel, and the tendency of debt-heavy booms to leave deeper recessions behind them.[xxv]

The exposures are uneven. Oracle, downgraded by S&P to BBB- in July, sits one notch above speculative grade. Oracle came late to the hyperscaler race and spent heavily to catch up, a larger share of it borrowed than at any of its peers. S&P forecasts a roughly $42 billion free operating cash-flow deficit for fiscal 2027, while a single customer accounts for about half its contracted backlog. Its five-year credit-default swap—the cost of insuring its debt—reached 288 basis points, or $2.88 a year for every $100 of debt insured, on October 15.[x,xxvi] An issuer on the precipice of junk is the most sensitive instrument the market offers to isolate the theme of debt-financed data-center construction, because a move into the high-yield universe would reprice its entire cost of capital at once. Its credit-default swaps therefore register each change in belief about the buildout before the equity of its better-capitalized peers does.

CoreWeave, a specialist lessor of GPU capacity, has less protection from established businesses. A neocloud with no fortress balance sheet to cushion a foray into the business of leasing compute, it is the canary to Oracle's weather vane. It reported roughly $52 billion of debt and lease obligations against $5.5 billion in cash. Its 2032 notes, issued at par in June, traded near 77 cents in October.xxvii The underlying mismatch is familiar: long leases on facilities financed against customer contracts that may expire years sooner.[x] For these businesses, the revenue forecast determines whether the next refinancing is routine or difficult.

Circular financing adds another complication. Suppliers, customers, and investors are often the same firms in different combinations. The proposed NVIDIA–OpenAI partnership announced in September 2025 illustrates the structure: NVIDIA contemplated investing up to $100 billion as OpenAI deployed systems using its hardware.[xxviii] A supplier would help finance a customer that would, in turn, generate demand for the supplier's products.

Such an arrangement might amount to nothing more than the common practice of vendor financing. A supplier may understand its customer's opportunity better than an outside investor does, and committed financing can get a productive project built. But the investment and the resulting order are not two independent confirmations of

final demand. Part of the evidence that customers want the product has been financed by the company selling it. Loops of interconnection that can be hard to untangle are especially common within the AI revenue and debt complex. There is a risk that investors underestimate the mutually reinforcing assumptions of the labs, the compute renters, and the hardware makers. With interconnection also comes the risk of contagion.

Vendor financing can amplify the cycle even when the arrangement is entirely legitimate. Circular revenue commitments that are used to raise new debt are amplifiers in a leverage cycle. They speed up the total spending in the same reflexive way that rising home prices in 2004-2005 allowed for banks to lend more for mortgages, which in turn caused more demand for houses and rising home prices. And so on.

As we saw in housing finance in 2007 and 2008, the same loop of increasing interconnectedness and higher velocity of capital that runs on the way up can operate powerfully in reverse. A lab that cuts its capacity commitments weakens its cloud provider's revenue outlook. The provider may reduce chip orders. A supplier that is also an investor can lose on its equity stake while its sales outlook deteriorates, making further customer financing less attractive. Contracts, collateral, and guarantees determine the actual losses, but the exposures will not necessarily diversify one another.

Project debt and special-purpose vehicles can move borrowing away from a parent's balance sheet without removing the underlying risk. Lustig notes that banks have joined the financing in greater numbers, with CoreWeave's facilities since mid-2025 syndicated across ten or more lenders.[xxix] A diversified technology parent can withstand disappointments that a project vehicle cannot. The relevant question is where the obligations sit, and who has promised to support whom if the customer cannot pay.

This does not make a banking crisis inevitable. It makes the shift from cash to debt a change in the possible consequences, rather than merely a change in corporate capital structure. It also makes the timing of cash revenue more important. After all, equity-financed infrastructure can be slowly depreciated. A debt pile must be rolled, refinanced, or retired.

How Much Revenue Is Enough?

A bill of this size poses a question that is easy to state and hard to answer: how much would customers have to pay, and for how long, to make all this spending worthwhile? The usual AI revenue forecast starts with adoption. More people use the models, more businesses put them into production, and somebody estimates what they will spend. Van Nieuwerburgh asks the question backward: given the capacity already announced, how much would

its owners need to earn to make an acceptable return on capital?

He starts with the roughly 183 gigawatts of data-center capacity already announced in the United States through 2032 and a capital base of about $9.6 trillion, including the money tied up during construction. The assumptions include a 10% return on the investment, a six-year life for computing equipment, twenty years for buildings and power equipment, and operating cash-flow margins of 50%. Economist Hanno Lustig has set out the resulting revenue path year by year.[xv,xxix]

The punchline is that by 2032 compute rentals would need to generate about $3.7 trillion per annum, which is nine percent of American GDP, if the economy grows four percent a year in nominal terms. For context, defense spending is around 3% of GDP. Five hyperscalers would need annual AI rental revenues roughly three times the country's defense budget. Writing in the Wall Street Journal, Greg Ip points out that nine percent of GDP exceeds what the country spends on energy, on construction, or on the computer industry as a whole.[xxx] The comparison can be sharpened. Computing hardware's share of GDP was about half a percent for most of the decade before 2023; the AI buildout has nearly tripled it, to roughly 1.6%, in three years, and that unprecedented move is itself a fraction of the revenue the capital now requires.[xxxi]

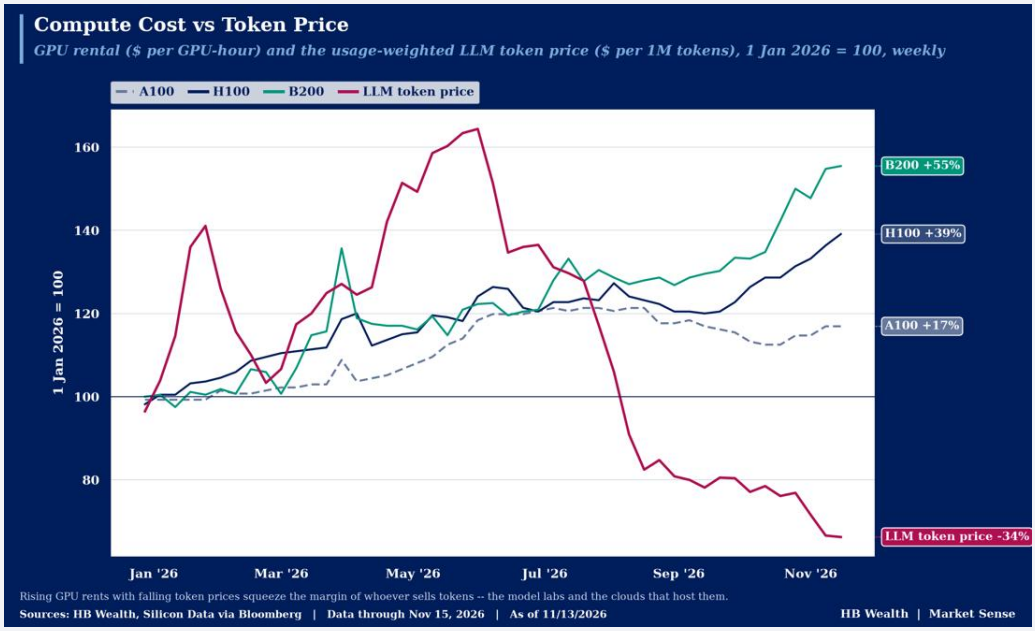
Jared Bernstein and Ryan Cummings use nominal growth nearer the 5% the US economy has lately achieved and a required return closer to the hyperscalers' 5-to-8% borrowing costs. Their ultimate revenue requirement is roughly two thirds as large. That is a lower hurdle for paying back lenders, though still staggering.[xxxii] In dollars it is still roughly $2.5 trillion per year in expected revenue from leasing compute. Bain's estimate last year was $2 trillion of annual AI revenue by 2030 against roughly $45 billion earned in 2024, and the range of published estimates has only moved up since.

On the other hand, six years is a generous useful life for a graphics processor; CoreWeave's lenders are reported to assume three, which would raise the annual revenue requirement toward 15% of GDP.[xxix] Further, the 9.2% of GDP figure covers only compute rentals. Much of this cloud leasing revenue would be derived from two companies, both currently private. The model makers are paying 9.2% of GDP. They too need usage rates that more than justify that expense line. Application software running on Anthropic or OpenAI or SpaceX AI models will also, presumably, earn a markup from customers. And there are service and maintenance costs to consider. And tangential investments in power infrastructure and transmission. In total, AI as a broader industrial complex would thus need to capture a good deal more than 9% of annual domestic output.

Both the debt used to finance data centers and the rapidly depreciating GPUs they are stuffed with carry a time fuse. Everything being built on borrowed money between now and 2032 is a bet that these eye-watering revenue numbers actually show up, and show up on time. How plausible is this bet?

Moving from roughly $300 billion of compute revenue to $4 trillion in seven years requires annual growth around 45%.xxix For this to happen, a great many customers would have to find a great many profitable things to do with AI. But volume increases alone won't support this rate of growth. Revenue is, after all, volume times price. The trickier assumption concerns steady prices, in an industry where the price of a unit of intelligence has done little but fall since there has been a market for one.[xxix] Growing from $300 billion to $4 trillion in seven years at 45% implies a price-per-token that holds steady throughout. That implied price, about $5.50 per installed GPU-hour at full utilization or $6.90 at 80%, sits at the bottom of today's on-demand range of $6 to $10 and more, which is to say that prices cannot fall.[xxix]

They are already falling. Silicon Data's usage-weighted index of the price paid per million tokens broke below $1 at the end of August, at about $0.97, down 29% in the month and more than half from its May peak near $2.05, a decline its compiler attributes to some combination of list-price cuts and migration to cheaper open models.[xxxiii] On OpenRouter, which routes requests across providers, token volume has roughly doubled in the three months to October, while average prices paid have halved over the same period.[xxxiii] Those two figures state the bargain on which the optimistic case rests: volume responds to price, and so far it has responded by a fraction of what would be needed to grow revenue, let alone to compound it at 45% a year. Goldman's trading desk has concluded that "frontier advantage is measured in weeks, not a moat, just a product cycle," with direct implications for expectations of the amount of compute demanded.



In other words, technological advantage is not durable, user switching costs are low, and lessors are already engaged in cost competition. This is a whole competitive structure whose gravity leads not toward monopolistic advantage but lean margins, discounting, and a focus on efficiency. If this is true, investors have learned the wrong lessons about scale and market share from the Internet platforms and SaaS giants. Hyperscalers may instead be entering an oligopolistic war of attrition where the critical variables are low cost of capital, supply chain advantage and the ability to survive on razor thin margins. In short, one valued more like an airline than IP-protected, low-variable-cost software franchise.

If prices fall, the volume needed to support any given revenue forecast increases. If margins fall too, customers must buy still more. A forecast of extraordinary use is therefore only the beginning of the investment case. It also needs a view about what customers will pay and what the providers will retain.

The benefit of the economists' decomposition exercise is to make visible the curve the market is pricing. Revenue rising from under 1% of GDP to 9% in seven years is a straight line drawn through the steepest part of a curve that, in every previous installation phase of a general-purpose technology, has turned out to be an S with a top. The curve that describes technological adoption is not an exponential one; it is a logistic function where exponential growth inflects and levels out at a horizontal asymptote.

Breadth of diffusion, the variable that usually carries the early part of an S-curve, is already well along: roughly 55% of businesses in Ramp's payments data now pay for AI in some form, with the share paying OpenAI flat since spring, so the growth required must come from each adopter spending far more than it does now.[xxxiv] The fiber boom illustrates the danger of getting that combination wrong. Investors financed capacity on the belief that internet traffic was doubling every hundred days; the actual rate was closer to once a year. Traffic grew enormously, the internet became indispensable, and many of the original investors still lost their money.[x] Lustig's comparable arithmetic puts the Internet carriers' required additional revenue at 1.7% of GDP and the revenue they achieved at about half that.[xxix]

Hundreds of billions of dollars of debt have been set against this anticipated revenue. A key difference between debt and equity is that the former has a maturity. The bill comes due whether the cash is flowing as expected or not. Some of that revenue is currently contract. Most is not.

We wrote last year about the mismatch between CoreWeave's ten-year leases and its two-to-five-year customer contracts, and about bonds that will need refinancing whatever the state of AI demand when they mature.[x] That financing has since grown more opaque, running through project debt, private credit and special-purpose vehicles with 90% leverage and kept off the hyperscalers' balance sheets.

Banks, which held little direct exposure until recently, have been drawn in: every CoreWeave facility since mid-2025 has been syndicated to ten or more of them.[xxix] As of mid-October, it cost 1,000 basis points annually to insure CoreWeave's recent notes against default. These securities were issued at par this year and already trade at steep discounts. This is the sort of thing that ratings agencies and banks' risk departments notice.

In the telecom bust the losses fell on shareholders and a few junk bondholders; no bank needed rescue. Today, debt—including and especially high-grade debt held by banks and insurance companies, in an echo of the mortgage crisis—is rising rapidly enough to bring systemic risk back into the discussion.

The implicit argument of the preceding discussion is that the collective spending on AI infrastructure presumes a shift in economic activity on a scale and timeline without precedent in the history of technology. And another shift, from building with excess cash to borrowing imposes an external, artificial clock on that first transformation, entangles a new set of important financial institutions in that risk, reduces robustness to failure and delay, and drives up the cost of capital for the entire economy.

We are not forecasting a maturity cliff or an imminent implosion of the balance sheets of the largest tech companies. Indeed, you never need to pay off your debt if lenders are willing to refinance you perpetually, as the U.S. government has demonstrated. A private borrower usually has to keep convincing them. That begins with having a revenue stream plausibly sufficient to amortize debt. "Build it and they will come" works when capital is flowing freely. When—as is inevitable—the credit market becomes more choosy about who is allowed to roll their debt, and on what terms, those without sufficient cashflow face foreclosure.

What the Boom Does to Everyone Else

We have just described one industry's expected growth in terms that would make Don Quixote blush. Given the size of these estimates, it pays to zoom out and ask what the boom is doing to the economy as a whole.

Building cloud computing capacity creates demand immediately. Its contribution to supply depends on when the equipment starts operating and what customers do with it. During the interval, the AI industry competes for resources with people who have no interest in AI at all.

The scarce resources that data centers require include electricians, transformers, land, and power. They also includes savings. A trillion-dollar investment program needs financing alongside a federal government borrowing about $1.9 trillion this fiscal year.[xxxv] More demand for capital raises its price, leaving other borrowers to pay more or abandon projects. We have heard from real estate investors who find construction financing harder to obtain as lenders devote more of their capacity to data centers and related infrastructure. Their anecdotes are not a national estimate, but they describe the mechanism.

The bond market helps distinguish that pressure from expectations of higher inflation. The ten-year Treasury touched 5.3% at the end of September, its highest since 2002, and has stayed above 5% since. Let's look at what goes into that number. A Treasury yield is two things added together: compensation for the inflation investors expect, which can be read off by comparing an ordinary bond with an inflation-protected one of the same maturity (the "breakeven"), and the real yield, which is whatever is left over. In early October the split was 2.3% expected inflation and 2.9% real.[xxxvi]

Expected inflation has sat near 2.3% all year. Every bit of the rise in long yields since the spring has come from the real component, and the Treasury's own inflation-protected auctions tell the story more sharply still: September's ten-year reopening cleared at a real yield of 2.65%, the highest since October 2008, on the weakest demand (bids

received against securities offered, the plainest gauge of appetite at an auction) in a year.[xxxvii]

A breakeven is market compensation for inflation, rather than a perfect forecast of it; TIPS' lower liquidity also affects the comparison. Still, it is informative that investors have been demanding a higher return above inflation. That is what a supply-side boom looks like in the bond market. A real yield is the price of capital, what savers must be paid to put off spending and what borrowers must pay to bring spending forward. When one industry decides to spend a trillion dollars a year on equipment that lasts a few years, and the federal government is borrowing $1.9 trillion (5.8% of GDP), at the same time, the two are bidding for the same pool of savings, and the price of that pool goes up.[xxxv]

The hyperscalers' bond sales, which we counted earlier, are the mechanism in miniature; Treasury auctions now share the calendar with Alphabet's.[xxxv] The term premium, the extra yield investors demand for lending for a decade rather than rolling over bills, is at a ten-year high.[xxxv] None of this requires anyone to expect more inflation. A surge in investment can raise real yields even if expected inflation barely moves.

Inflation itself is a different matter. The channels through which the data-center boom reaches prices are real but narrow: electricians' wages, the price and delivery times of transformers and gas turbines, and the tariffs utilities file to cover the cost of new generation. Consumer prices rose 3.7% in the year to September.[xxxviii] The war with Iran that began in March has kept Brent crude above $106 a barrel, and energy accounts for most of the headline figure.[xxxv]

But a sufficiently large construction boom can add to aggregate demand before the new capacity makes the economy more productive. That is the Fed's concern. Its September minutes discuss the possibility that demand from construction and equipment purchases will exceed the near-term supply benefit.

The chairman's own language has evolved with that timing problem. In July, before the Senate Banking Committee, Kevin Warsh said that he did not "view a one-time change in prices as necessarily being inflationary because I think there's a supply response," and distinguished AI, a positive supply shock, from a foreign conflict, which tends to shrink supply.xxxix By October, he was stressing that AI investment was hitting demand before supply (e.g. more AI-enabled productive activities) comes online. Between the two statements, the FOMC had voted unanimously to raise the funds-rate range to 3.75–4.00%, the first such hike since July 2023, with most participants expecting another increase this year.

The Fed cannot make an AI-specific mortgage rate. If it raises rates to restrain aggregate demand because of an economy overheating on an infrastructure buildout, the consequences reach homebuyers, small businesses, students, and manufacturers whose capital plans involve no GPUs. The beneficiaries of the boom and the borrowers paying more because of it need not be the same people.

Today's concerns rhyme with the autumn of 1996, when the Fed's staff model and several members of the committee saw unemployment falling toward 5% and wanted to raise rates. Alan Greenspan argued that measured productivity was understating the transformation happening inside firms, pointing to strong profit margins, and held off hiking ahead of an anticipated production boom. The revisions of the following years vindicated him.[xl] In July, Warsh sounded like his hero. By October he was voting with his committee. Whether the impact of AI will be inflationary or disinflationary depends on whether the technology begins lifting the supply side of the economy—that is, will it increase production of goods and services without increasing the cost of production? That is a question that only productivity data can begin to answer?

Is AI a productivity miracle, and if so, when?

Forecasts of sustained revenue growth at hyperscalers and foundational model labs all require customers to find AI increasingly valuable. The investment boom requires the economy to produce enough additional goods and services to absorb the increased demand without merely bidding up prices. Both requirements assume, in different ways, that we get more output from the resources we already have.

Productivity measures how much output an economy gets from its labor and capital. For a data-center owner, the hurdle is more immediate: rental income must cover operating costs, financing and replacing equipment that may lose a sixth to a third of its value each year, depending on the assumed useful life. Higher real interest rates raise that hurdle.

$3.7 trillion of assumed revenue (or $2.5 trillion, if you prefer), must come from these customers. A user keeps paying for intelligence only as long as it makes the customer's own labor or capital go further; the laboratories' margins, the application vendors' margins and, behind them, the chipmakers' order books are all shares of that surplus. Productivity sets the size of the pie and the market structure we have described determines the size of the slices.

We wrote last year of recurring lags in the history of general-purpose technologies' implementation and the appearance of measured productivity gains—on the order of four decades for electricity, and two for personal

computers. The reason given for such lags is the requirement that businesses and individuals reorganize workflows around the capabilities of new technologies. Customers must figure out how to get the most value out of a gadget, supporting infrastructure and complementary technologies must be developed, and production processes often have to be reengineered from the ground up. This takes time. The technology is usually pervasive long before the economy registers an uptick in productivity.[xli]

We also surmised that, due to AI's natural language interface and its embeddedness in software people already use, the lag to measured productivity gains would be shorter this time. AI compresses parts of the process. A company can obtain a capable model without building one, and employees can discover applications without waiting for a central IT project. It does not abolish the need for clean records, suitable incentives, reliable evaluation, or decisions about who is accountable when the model is wrong. Ten years seemed a reasonable, shortened timetable before aggregate statistics begin to capture what is today merely anecdotal. Debt schedules are not as forgiving. And every indication suggests that the economy is in the very early innings of Carlota Perez' famous diffusion S-curve.[xlii]

Whether the promised productivity lift is showing green shoots depends on where you look. In controlled experiments, some studies find large productivity gains for knowledge workers. A randomized trial published in Science gave several hundred college-educated professionals realistic writing tasks and found that those using a generative model finished 40% faster and produced work graded 18% higher in quality, with the largest improvements among the weakest writers.[xliii]

Studies of coding assistants document productivity improvements of 20% to 40%, and more than half of new code at many firms is now machine-generated.[xliv] Agentic coding tools like Cursor, Codex, Claude Code and Replit are seeing the most dramatic usage and revenue acceleration, while employment for software developers stagnates. These products likely constitute generative AI's first "killer app." Activities not measured in GDP are benefiting as well; researchers at Stanford's Institute for Economic Policy Research, tracking American households, find that generative tools raise what they call home productivity, on digital chores from meal planning to job applications, by 76% to 176%.[xlv]

An experiment on a bounded task is not a measurement of an entire firm's productivity. In production, someone has to integrate the tool, decide when to trust it, and deal with its mistakes. Researchers at MIT Sloan have used the term "botsitting" to describe the familiar experience of saving time on a first draft and spending it again

checking what the machine invented.[xlvi] The gains can be real while the surrounding organization absorbs much of them. Firms tell a more divided story, and two surveys published through the National Bureau of Economic Research this year illustrate the division. In one, nearly 750 executives, most of them chief financial officers, attribute a mean 1.8% of labor-productivity growth to AI and expect that contribution to rise to 3.0%, with gains concentrated in high-skill professional services and finance.[xlvii] In another study, which matched some 6,000 executives across the United States, Britain, Germany and Australia, 89% report no measurable effect on productivity over the past three years.[xlviii]

Both stories can be true. Productivity gains should be expected to concentrate among superstar firms and superstar workers, at least at first. In other words, a few outliers can drive up the average while the median firm or worker still reports no noticeable productivity improvement. The distribution of spending says the same thing. In payments company Ramp's business-spending data, the top 1% of firms by AI expense paid $9,568 per employee per month in September while the median firm paid about $16. Roughly 55% of American businesses now pay for AI in some form, though the share of businesses paying OpenAI has been flat since the spring.[xxxiv] Businesses may adopt the technology because their competitors are, without knowing how to make particularly effective use of it. A few technologically adept firms rebuild entire business lines around agentic workflows and see massive productivity boosts. For everyone else, implementation is a slow, clunky trial-and-error process.

Moreover, existing technological infrastructure may need rebuilding (e.g. compliance, training, financial analysis of agentic costs); this can actually cost productivity in the short run[xlix]. These are some of the reasons why measured productivity tends to lag when a technology like electricity, computers, or the Internet is already ubiquitous.

Some of the missing complementary is legal. In February a former student killed six people at a secondary school in Tumbler Ridge, British Columbia. The suits that followed allege that OpenAI's own monitoring had flagged her account for gun-violence planning some eight months earlier, that its safety team recommended referring the logs to the police, and that leadership declined and closed the account instead, after which she opened another; the families sued in April, teachers and students who survived the attack in September, and the Province of British Columbia the same month for failure to warn. Altman has published an apology to the town for not reporting the account.[l]

Whatever the courts make of the facts, forthcoming liability cases are arriving in a legal setting that the laboratories did not choose. The statute that sheltered the internet platforms from liability for what their users posted, Section 230 of the 1996 Communications Decency Act, protects intermediaries that feature other people's speech; a model that generates the text has no such excuse. Courts that have ruled so far have treated chatbots as products, letting design-defect and failure-to-warn claims proceed against Character.AI and declining its First Amendment defense at the pleading stage. A wrongful-death suit against OpenAI over a California teenager appears on the same path.[li] There is, as yet, no Section 230 for the model age.

Every general-purpose technology has eventually been given a liability regime, and in each case the regime both limited and allocated the exposure before capital would commit at great scale. The railroads received theirs from the courts: the fellow-servant rule of Farwell v. Boston and Worcester in 1842 and the contributory-negligence doctrine of the same decade, which a generation of legal historians has read as judicial subsidy to the new industry.[lii] Aviation received the Warsaw Convention of 1929, which capped carriers' liability to passengers so that the lines could be insured and financed. Nuclear power received Price-Anderson in 1957, a cap on operators' liability with federal indemnity above it, without which no utility would order a reactor. The internet received Section 230 and the safe harbors of the Digital Millennium Copyright Act, and the platform economy was built inside these rules.[lii]

The sequence bears on the argument of this paper in two ways. The pacing proposal's embedded evaluators are, among other things, a bid to be the body that writes the standard of care, and compliance with a standard one has written is the usual shape of a safe harbor; that is Stigler's regulatory capture in its legal form. Without such regime, the market may price liability exposure as tail risk.

The Insurance Services Office issued its first AI endorsements to the general-liability form in April; W.R. Berkley is filing exclusions in directors' and officers' and errors-and-omissions cover for any claim "arising out of" the use or development of AI; the affirmative products, Munich Re's for model drift and Armilla's with Lloyd's for hallucination, are thin and sublimited, and the largest corporate buyers of AI reportedly cannot obtain the limits they want.[liii] A technology whose liability is uninsurable is a technology whose users are self-insuring, which is one more reason the median firm in the NBER panel has gone no further than a pilot.

A further problem of reversing agentic actions sits beside bot-sitting as a brake on adoption, requiring humans in the loop. When an agent executes rather than drafts, when it places an order, moves funds, signs an agreement or

answers a customer, the error cannot be reviewed before it takes effect. The law that governs it is older than the technology and mostly runs against the technology's deployer. Contracts formed by what the statutes call electronic agents have bound the person who deployed them since the Uniform Electronic Transactions Act and E-SIGN in 2000.

The doctrine of unilateral mistake excuses an order of ten thousand units in place of ten only where the counterparty knew or should have known; Air Canada was held liable in 2024 for a refund policy its chatbot invented; and under Article 4A of the Uniform Commercial Code the loss on a commercial payment stays with the customer whenever the bank followed commercially reasonable security procedures, so that an agent whose credentials are captured by a hostile instruction leaves the loss with the firm that gave it the credentials.[liv]

The framework that would reallocate these risks is being built by the payments networks rather than the legislatures, through agent-identity standards and cryptographically signed mandates of intent that Visa, Mastercard, Google and Stripe have each proposed, and through working groups whose recommendations are due in 2027. The IMF, observing that existing liability regimes "assume human intent and direct causation," highlights the legal gray area of autonomous technology deployed where once only humans worked. The European Union's revised Product Liability Directive, which brings software under strict liability from December, is so far the only clear statement of a regime, and it is not the one preferred by Silicon Valley.[lv]

Transparency, legibility and alignment, in other words, are problems to be solved at the level of the firm and the deployment as well as at the level of the model, and the firm-level solutions, identity, authorization, audit and insurance, are complements that the capital spending of the past two years has not built. Their absence is a bottleneck on adoption, and the spending that relieves it will land in compliance, risk and security software, which is to say downstream.

45% compound growth rates for revenue necessarily assume some broadening of enterprise use, and a flattening of the adoption curve. For vendors to maintain pricing power, a market for customers must be deep and wide. Today productivity gains are highly concentrated at opposite ends of the AI value chain; construction, electrical equipment and the regulated utilities that pass the cost of power generation through to their customers are enjoying a boom. At the other end, the three sectors the Dallas Fed identifies as most exposed to AI — information, finance and insurance, and professional and technical services — have together delivered 40% of the economy's productivity gains since early 2024 while accounting for only 16% of its hours.[lvi] A few firms can improve enough to

lift an average while most see little change. They may have better data, clearer processes, more appropriate tasks, or managers willing to reorganize a business around the potential of the new tools.

For many businesses, the process is still experimental. An employee gets a subscription. A department tries a pilot. A manager discovers that the records the model needs are dispersed across incompatible systems, or that the output cannot be used without a review process nobody has designed. A competitor's announcement then prompts another pilot. There may eventually be a large productivity gain at the end of this sequence. There may also be several years of subscriptions, consultants, and reorganizations before the contributions of AI can be measured.

Two cautions belong here. The statistics report output per hour, which rises mechanically when a great deal of capital is placed behind each worker, so the better test is in the sectors that rent intelligence rather than own the chips. And much of what AI produces is invoiced to no one. The hours a household saves on its digital chores are a gain to welfare but don't count in aggregate statistics. The question then is whether the chips produce output that someone is willing to pay for.

National accounts have shown evidence that productivity has indeed begun to leave its two decade doldrums behind. American labor productivity has grown 2.4% a year since the start of 2024, against 1.6% in the five years before the pandemic. In the Dallas Fed study referenced earlier, economist Scott Davis applies an index of occupational exposure to AI, finding that the three most exposed sectors have grown at 3.7% a year and the rest of the economy at 1.7%.

This is a correlation, not a directional finding. In the European Union, where usage is lower, no such relationship appears across sectors.[lvi] The International Monetary Fund has published a paper this year on the limits of the national accounts in capturing a technology whose output is embedded in the middle of other people's work, and the Bureau of Labor Statistics tracks AI only indirectly, through the software component of its capital measures, where it finds the most exposed industries becoming markedly more capital-intensive.[lvii]

Our earlier paper discussed the J-curve along which measured productivity first dips, as resources are diverted to reorganization and to intangible investment the statistics do not count, and then rises as the complements mature.[xli] Our reading of the evidence above is that the curve has turned upward in the exposed sectors and has not yet turned anywhere else. That is consistent with an early upswing, but too little to establish how fast the whole economy can grow without adding inflationary pressure, since five-sixths of the hours worked remain in

sectors where important complements are still being built.

The distinction matters for valuations as much as for policy. A law firm or an insurer that redesigns its work around AI tools adds to output without adding to its own capital base, and the return accrues as savings to its profit margins rather than to a supplier's revenue. If the exposed sectors' experience spreads, the result is the kind of growth that is good for the earnings of the companies that use intelligence and indifferent to the companies that rent capacity. Internet retailers may make handsome margins and high returns on capital without any impact on the valuations of ISPs or the companies producing wireless routers.

The investment problem is a mismatch between the time needed to reorganize a business and the time allowed by a financial contract. Neither a bond maturity nor a GPU replacement schedule adjusts itself to the progress of a customer's implementation plan. Productivity may improve impressively over a decade and still develop too slowly for an owner who needs a profitable contract renewal in three years.

Wider productivity gains could lower unit costs and ease inflation, depending on the elasticities of wages and margins.[lviii] Yet an economy with persistently better investment opportunities might also sustain a higher neutral real interest rate. Disinflation is a plausible story that would not guarantee falling rates, as witnessed in the "secular stagnation" in the 2010s when there was apparently very little demand for capital.

What Has Slowed a Race Before

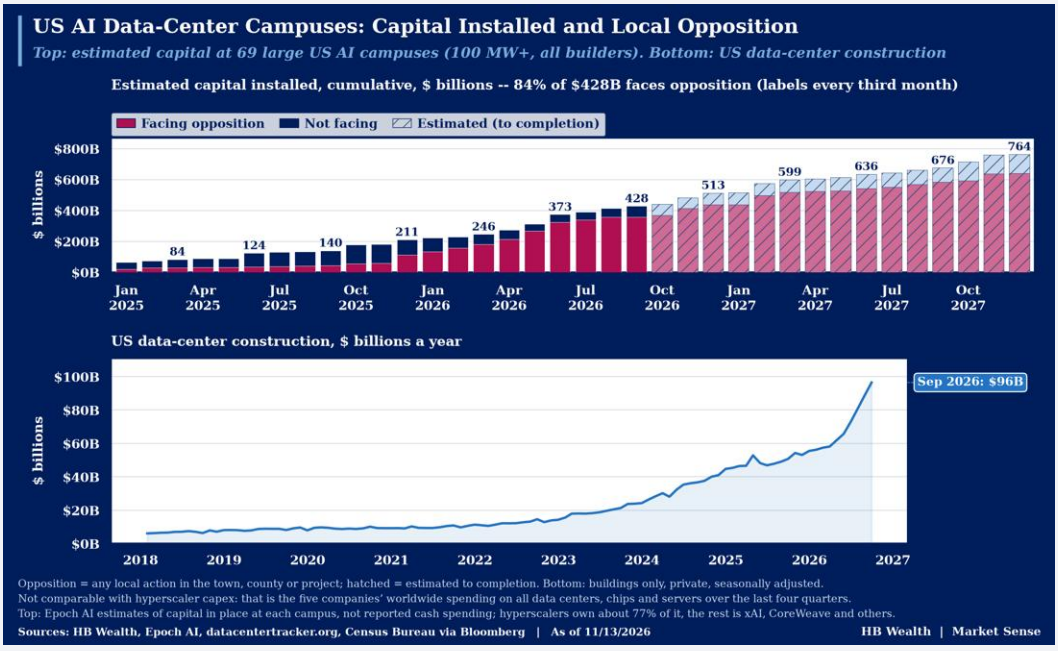
Currently, the cycle of investment is meeting physical and social frictions on the ground even while shareholders and lenders have been remarkably permissive. By September, only 12.3 gigawatts of the 148 gigawatts of hyperscaler capacity announced since the start of 2024 were operating. Much of the announced capacity was intended for future years, so that comparison alone does not measure delays. The current delivery cohorts are more revealing: 11 of the 16 gigawatts slated for 2026 reportedly remained at the announcement stage, with no visible construction.[lix]

Electricity is most often the constraint. Berkeley Lab's interconnection data show generation and storage projects waiting years to connect, while large transformers have delivery schedules approaching three years.[lix] A financed data center cannot solve the absence of a power connection by ordering another shipment of Blackwells.

Where power is available, someone must pay to expand the system. That has become a political question. In the thirteen-state PJM region, higher capacity costs have fed through to residential bills. Ohio now requires new data-

center customers to pay for at least 85% of the power they reserve whether or not they use it; other states are pursuing separate rate classes and protections for ordinary customers.[lx] These arrangements can make the allocation of costs fairer. They also make the industry's commitments more punitive.

Water use, air permits, and the appearance of enormous industrial sites have prompted another set of objections. Virginia will require monthly reporting of data-center water use from January. By June, fourteen states had considered moratoriums, and New York's legislature had passed a one-year pause.[lxi] In Memphis, a contested moratorium followed disputes over gas turbines and air permits at the SpaceX AI campus. The council eventually rejected the pause on October 20, after earlier meetings included police clearing the chamber.[lxi] Securing permission to build is becoming an operating requirement alongside securing chips.



These power, environmental, and community frictions reinforce one another. A delayed power connection postpones rental income and increases financing costs. A higher interest rate makes the delay more expensive. Lower expected rental prices reduce the return from pressing ahead. A project that can withstand any one of those changes may fail its investment test when all three occur together.

The effects are not uniformly bad for existing owners of compute capacity. Scarce power can protect a well-connected operating facility from competition. Delayed new supply may support current rental prices. But those advantages belong to the particular assets; they cannot justify every project announced elsewhere. The

investment case needs to distinguish a functioning campus with a paying tenant from a plan to build one.

A second proposed brake, and one that the frontier labs might prefer, is regulation. The historical record of regulation, self-imposed or otherwise, is consistent; it seldom pauses unruly spending, though it may concentrate leadership with incumbents.

In 1975 the leading molecular biologists met at Asilomar and agreed a moratorium on recombinant-DNA experiments; it held for about a year before the National Institutes of Health issued guidelines that replaced it, and capital formation in biotechnology accelerated afterward.[lxii] The 1962 drug amendments that followed the thalidomide scandal required proof of efficacy before approval, and Sam Peltzman's assessment a decade later found that they had roughly halved the number of new drugs introduced and raised the cost of each; industry spending on R&D rose for decades regardless, and the sector concentrated, which is the regulatory capture dynamic in microcosm.[lxiii] Orders for new nuclear reactors in the United States, already slowing, stopped after Three Mile Island in 1979, through licensing timelines and liability rules the industry fought rather than requested.[lxiv] These episodes show different effects: rules can legitimize a technology, raise the cost of entry or halt particular projects. They offer no simple template for an AI truce.

What ended most previous investment booms was the price of money. Parliament authorized more new railway mileage in 1845 and 1846 than Britain had built in the two preceding decades, at the height of a mania that financial historian Andrew Odlyzko has called a "collective hallucination"; the mania ended in 1847, when the Bank of England raised Bank Rate to 8% during the commercial crisis of that autumn and investors who had subscribed for shares on installment found they could not meet the capital calls.[lxv]

In the United States, the railroad boom of the early 1870s ended in September 1873 when Jay Cooke's bank could no longer place the bonds of the Northern Pacific and closed its doors.[lxvi] The failure of the Cooke bank turned a particular railroad financing problem into a broader panic. The market's inability to continue supporting new projects induced a widespread business collapse even though railroads had many productive decades ahead. More recently, telecom credit deteriorated long before the equity market fully accepted that its buildout had exceeded its commercial opportunity.[x]

What lenders are being asked to believe is that AI compute leasing revenue can rise to somewhere between 5% and 9% of GDP within seven years, at margins and prices that hold as long as the term of bonds and exceed the lives of the chips inside data centers. The bond market's patience is therefore the variable to watch, and it is being

tested. S&P's current forecasts already depend on an eventual adjustment: revenue growth accelerates while capital spending slows around 2028.[xvi] In other words, even a comparatively benign credit outlook requires a change in the relationship between spending and receipts. The issue is whether that adjustment occurs through growing customer demand and more selective construction, or through projects and borrowers that cannot obtain further financing.

The weakest bid-to-cover ratio in a year at September's inflation-protected auction is evidence of the bond market beginning to structurally creak.[xxxvii] This could be looked at as a potential signal of a cycle beginning to turn. Two observations would change our view: a granted waiver used to agree ceilings on training compute, or a hyperscaler cutting its capital guidance and citing pacing as the reason. Neither has happened. The first-order constraint remains the willingness of lenders to finance negative free cash flow at a real yield approaching 3%.

The shift from cash to debt is also what makes this episode's ending suddenly relevant to the whole economy. When the hyperscalers paid for data centers out of earnings, a disappointment in AI revenue would have been a stock-price event. Valuations would have fallen and the businesses would have carried on. With debt in the financing mix, the same disappointment can also reach creditors, and a default by a company of Oracle's or Meta's size could travel through banks and insurers as railroad defaults did in 1873.

Ben Bernanke's work on the Depression showed that the damage from a financial crash propagates through the lending system, beyond the loss of wealth itself, leading to credit retrenchment and recessionary thrift. Later literature found that the more debt a boom carries, the deeper the recession that follows its end.[xxv] We drew the same distinction in an earlier paper between equity-financed bubbles, which leave speculators poorer, and debt-financed ones, which leave an overhang of balance-sheet repair.[x]

Who holds the paper has changed since the last cycle. The long-dated public bonds, the thirty- and forty-year tranches Alphabet and Meta sold over the past year, went to the institutions that match long liabilities: life insurers, pension funds and foreign institutions. American life insurers hold close to $4 trillion of bonds, and privately placed bonds, which trade rarely if at all, made up 48.4% of the total at the end of 2025 against 37.4% five years earlier; the industry's private-credit holdings grew 21% last year, twice as fast as its assets.[lxvii]

That private and structured paper is concentrated at the insurers controlled by private equity giants. Apollo, KKR and Brookfield own insurers, Blackstone manages insurance assets, and all four are partners in Nvidia's platforms to mobilize $500 billion of third-party capital for compute.[lxviii] Banks, which had little direct exposure until recently,

now syndicate the neoclouds' facilities among themselves.[xxix] The data-center securitizations through which much of this money flows have been ruled by SEC staff not to be asset-backed securities, which frees them from the risk-retention requirement where a seller keep a slice of what it sells, as well as from most disclosure. They are typically written with a five-year anticipated repayment date against a final maturity of twenty-five to thirty years, so that the paper is long on its face and short in practice.[xxviii] One sponsor-controlled insurer, after grand-jury subpoenas, reclassified $17.8 billion of private credit as dependent on affiliates, about 40% of its invested assets.[xxviii]

How a loss would travel through those holders is governed by accounting as much as by economics. Under statutory rules, a life insurer books every bond short of default at what it paid, adjusted toward par over the bond's life, so a bond that falls twenty points as spreads widen does not touch the insurer's capital until it is written down as impaired, and an impairment runs first through a reserve funded for the purpose. The capital charges regulators require move in steps with the rating, roughly 2% of face value for a BBB-, a little over 3% for a BB+, under 10% for a single-B, and regulators intervene only when capital falls to twice the level at which they would take control, a line few large insurers are anywhere near.[lxix]

When the mechanics have threatened to force sales, regulators have loosened them, replacing agency ratings with modeled designations for mortgage securities in 2009 and allowing insurers since 2023 to defer realized interest-rate losses on bonds they sold rather than take them against capital, under relief now extended through 2027.[lxix]

Any forced selling would have to come from elsewhere: from the investment-grade boundary, where index rules and mandate limits would make an issuer of Oracle's size, with more than $100 billion of debt against a high-yield market of about $1.4 trillion, an enormous fallen angel—some three times the size of Ford in 2020, the largest on record. Other triggers for disorderly selling could include daily-priced funds, and banks whose warehouse lending to private-credit funds the Financial Stability Board flagged in May.[lxx]

The insurers' own channel could still make it difficult for hyperscalers to issue new paper or term out existing liabilities. If poor performance and ratings downgrades necessitate reserve builds, or policyholders increase the rate of surrenders—asking for their money back—the insurance buyer would step back. When the natural holder of forty-year paper withdraws, the long end of the primary market closes, and more issuers are pushed into a crowded five-year maturity where attention to debt service and quality-of-earnings is more acute and prices more

finicky.

One observation we note without pursuing: Taiwan's life insurers hold foreign bonds on the order of $600 billion, much of it long-dated American corporate credit, and have been the marginal buyer of thirty-year investment-grade paper for a decade. Since January they have operated under a fair-value capital regime in which the charge on a bond changes with its rating, its sensitivity to spreads and its maturity, with a separate charge for spread widening short of default; the phase-in runs fifteen years, and the expected direction of travel is out of long BBB credit and toward higher quality.[lxxi] The buyer of last resort for the longest hyperscaler bonds has been handed rules that penalize holding them, at the moment the supply of such bonds is at a record.

It is notoriously difficult to forecast the performance of new financial instruments in a deeply interconnected system with evolving reporting rules and regulatory requirements. This essay should not be read, therefore, as a forecast. In the telecom bust the losses fell on shareholders and high-yield bondholders and no bank required rescue[xxix]. The important point is that this cycle's paper sits increasingly in the balance sheets that back annuities and pensions, in private vehicles exempt from the disclosure that would let outsiders see it, and in bank loan books by way of the funds those banks lend to. A mechanism that is slow and opaque delivers its signal late. If the revenue arrives as hoped, the question is academic. If it is slow, sudden tipping points in credit markets could cool the buildout long before AI's productivity boost has run its course.

A brake to spending is not an exit from the arms race. A cycle ended by the price of money leaves stranded assets behind, and the technology is deployed by someone else at a discount, as dark fiber was after 2001. The more benign way for the race to end is one we see emerging at the model labs themselves. This off-ramp would not likely serve to benefit of those selling chips, power, or equipment into planned but unbuilt state-of-the-art campuses.

The Customer Would Like a Discount

There is a second difficulty with borrowing against future productivity. Customers want to keep some of it. Much of the industry's early attention went to the technological frontier: which model could reason better, solve the harder problem, or lead the benchmark table. For many enterprise applications, the more relevant question is which model can perform a specified task reliably at the lowest total cost. A company extracting fields from invoices does not necessarily need the model best equipped to prove a century-old mathematics theorem.

The exit is cheaper intelligence, and the question is who supplies it. Chinese open-weight models, whose learned parameters are available for others to customize, have made this competitive alternative increasingly difficult for

the American labs to ignore. OpenRouter's June survey finds that leading open models have remained roughly three to six months behind the US frontier for more than eighteen months. Chinese entrants lead important parts of that market, and the survey does not show the frontier labs steadily escaping them.[lxxii] Benchmark gaps are imperfect measures of usefulness, but a short lag matters if the cheaper alternative is already adequate for the customer's task.

American firms have responded in several ways. In August 2025 OpenAI, whose chief executive had conceded the company was on the wrong side of history regarding open models, released two open-weight models under the Apache 2.0 license, one that runs on a single Nvidia graphics processor and one that fits on a laptop; both were deliberately mid-tier, landing in the company's own tests between its smaller reasoning models of the time.[lxxiii] On October 15, Meta released Muse Spark Open, which it calls the most capable American open-weight model to date, reversing a spring decision to keep the model closed; Meta's own benchmarks place it behind the leading Chinese open models, and its closed Muse Wonder remains at the top of the line.[lxxiv] NVIDIA distributes open models together with training resources and tools.[lxxv] Its interest is straightforward: customers using an open model can still buy NVIDIA hardware to run it.

Anthropic and Google have not opened their frontier lines, and the direction of travel at the top runs the other way. Opening a model is not the only way to compete on cost. The frontier labs also sell smaller closed models and cheaper tiers. Reported September prices span more than a hundredfold between capable alternatives, and lower-priced models increasingly perform tasks that once required a flagship.[lxxvi] Efficiency can be gained through better allocation of model to task rather than just a separate architecture. OpenAI's lower-priced September model is reported to deliver performance close to its flagship at roughly a fifth of the cost per task. The market is separating into services with different costs and capabilities. Expensive models can remain the best without being the sensible purchase for most requests.

The tailored model for corporate finance and marketing is already here, and it is increasingly being built by the customer. Bridgewater's AIA Labs provides a concrete example. Its researchers, working with Mira Murati's Thinking Machines, fine-tuned an open Qwen model on labels supplied by the firm's own investors for six document-filtering tasks. The best frontier model with expert prompting achieved 78.2% accuracy. The tailored model achieved 84.7% at roughly one-fourteenth of the inference cost per task.[lxxvii]

Gartner forecast two years ago that by 2027 more than half of the generative models enterprises use would be specific to an industry or a business function, from about 1% in 2023.[lxxviii] At Nvidia's developer conference, Huang said that proprietary versus open "is not a thing, it is proprietary and open," and Alex Karp of Palantir has argued that whoever owns the enterprise's context can commoditize the model beneath it.[lxxix] The emerging consensus is a routing world, in which routine classification and extraction migrate to small or open models while the hardest work stays with the frontier.

The cost of developing and maintaining the tailored system still matters. Cheap tokens are not the same as a cheap implementation, and a corporate finance team may prefer one large vendor to a piecemeal assembly of small providers each requiring separate security reviews. Distribution, support, interfaces, compliance, and integration can therefore count as much as benchmark performance. Those are advantages that a technological lead alone may not supply.[lxxix]

Which providers penetrate the customer's wallet will depend on establishing switching costs—through customer service, systems integration, and data access. The railroads were undone by their absence of a moat; freight could be moved from one company's line to another or onto a barge, parallel lines were built between the same cities in the 1870s, interconnection laws in the 1880s ended the ability to exclude rivals, and revenue per ton-mile fell from about three cents in 1866 to under one cent by 1890 and never recovered, with the surplus going to the shippers.[xxi] A frontier model has no network effect either, and brand loyalty is limited, as frequent model-shopping demonstrates. Where the best model is a necessity, in adversarial uses such as cybersecurity and quantitative trading in which a few months' lag is defeat, the frontier commands a premium. Those are precisely the tiers where the laboratories are seeing price stickiness.

This is the likely direction of competition. Some customers will require state-of-the-art models for most tasks. Others will send routine requests to cheaper systems and reserve the most capable ones for difficult cases. Providers will compete on cost, service, flexibility, and the ease of fitting into a customer's existing workstreams. Tailored and open models are part of that adjustment. So are improvements to the user interface that make a tool worth paying for even when a rival model scores slightly higher.

None of the labs has yet demonstrated that frontier leadership confers durable pricing power. That is a problem for valuations premised on retaining both premium pricing and an ever-expanding share of use. It applies to OpenAI as much as to any competitor. A model can attract an enormous audience without establishing the power to hold

customers when a close substitute becomes cheaper.

The first enterprise-spending data are consistent with customers becoming more selective. Ramp reports roughly sevenfold growth in token volumes between January and September, while spending flattened from July. The frontier tier's share of tokens fell from 53% in early August to 45% a month later. About 8% of businesses in its sample changed their leading model provider in September, the highest share in the series.[xxxiv] Greater use without comparable spending growth is precisely the possibility the capex forecasts need to confront.

The usual rejoinder is Jevons' paradox. Make a resource cheaper and people discover enough additional uses to consume more of it. Satya Nadella invoked the idea after DeepSeek's models unsettled technology investors in early 2025.[lxxx] More efficient intelligence could support agents that use far more tokens than a person typing questions into a chat box. A cheaper task may be performed millions of times where an expensive one would have been attempted only occasionally.

That is a good argument for rising use. To establish rising revenue, volume must grow faster than price falls. To establish adequate returns on infrastructure, revenue must also cover operating costs, financing, and depreciation. A tenfold increase in volume sold at one-tenth the price leaves revenue unchanged. The technology has become much more useful and widespread; the borrower has no additional revenue with which to pay its creditors.

Revenue could still grow rapidly. OpenAI reportedly reached a $50 billion annualized rate, although that was below expectations of $68 billion. Anthropic's growth remains substantial even as its reported sequential acceleration has slowed.[xxi,lxxxi] These are young, rapidly evolving businesses. The issue is whether their growth can persist long enough and fast enough to justify the trillions of dollars of capital commitments that depend on commensurate growth in their demand. Falling costs can preserve margins, and new uses can more than offset lower prices. That is the optimistic story. But this bull case needs to support the scale of the capacity already being financed, rather than simply establish that AI has a growing market.

Contractual arrangements determine who feels the pressure first. A lab with fixed payments for capacity can lower token prices without obtaining an immediate reduction in its cloud bill. It may compensate by filling more of the capacity, generating more tokens per unit of compute, or selling additional services. If those responses are insufficient, its margins shrink. The lessor is protected while the customer pays, then exposed at renewal or default. A chip supplier benefits from orders placed to meet the original commitment, but faces a different

demand outlook when the next order is considered.[x] We therefore should not assume falling token prices translate mechanically into falling revenue per gigawatt. Efficiency and utilization of the equipment matter.

A final caution concerns how we read AI usage measures themselves. In May Amazon shut down KiroRank, a leaderboard built by its own engineers that ranked them by the tokens they consumed, after it became clear that people were generating work to climb it; Meta ran one called Claudenomics; and a memo from a distinguished engineer at Amazon, circulated in September and published last week, calls the practice it names "tokenmaxxing" at odds with frugality and describes input metrics as "deeply flawed."[lxxxii] Goodhart's law, that a measure which becomes a target ceases to be a good measure, has settled on the token.

The warning applies most directly to measures of AI use. Token volume, spending per employee, the share of firms with a paid subscription, the share of tokens by model tier: each is a measure of input, provisional and most are lagging. Which is what one should expect in the installation phase of a technology still in its adolescence, when the national accounts lag business implementation by years, surveys of firms disagree, and even adoption is counted by whether a card was charged. Volume can be irresponsible as well as productive, and a company that bills by the token, as Amazon does externally, earns revenue from the waste its own memo deplores.

The outcome we consider increasingly plausible is that AI use expands enormously at prices below those implicit in today's more ambitious valuations. Productivity benefits spread, enterprise budgets grow, and some providers still earn poor returns. The railroad analogy is relevant because shippers could benefit from cheaper freight while railroad investors struggled. The value of the service and the profitability of owning the line were different questions.

What this does to the chain of commitments described above depends on how the contracts are written. Most of the large ones are take-or-pay agreements for capacity measured in gigawatts rather than for tokens delivered. OpenAI's $300 billion agreement with Oracle runs over five years against property leases that run at least twice as long; CoreWeave owes $56 billion on ten-year data-center leases against customer contracts of two to five years.[x] Cheaper inference need not reduce the lessor's revenue immediately: the tenant may still owe the same capacity payment. The pressure first falls on the tenant's margin, and reaches the lessor if the tenant cannot pay or demands better terms at renewal.

Chipmakers are protected from a down cycle only by their order backlog, assuming those commitments are solid. Beyond the current orderbook, investors should be careful about extrapolating similar rates of demand growth.

Memory makers may be sold out through this year, but their business is doubly geared to the capex race; historically they have been the tip of the cyclical whip. Application builders and enterprise users each stand to benefit from cheaper, more efficiently allocated intelligence.

Set beside the required revenue calculus, these figures imply a collision. The forecasts ask for extraordinary growth in use without a comparable erosion in rental income. That is a demanding combination for both the customer and the electrical grid. They also run against a familiar pattern in information technology from Moore's Law to Nielsen's Law to Edholm's Law to Kryder's Law to Cooper's Law to Huang's Law. That's enough laws to fill a stone tablet. In each, a unit of processing power, storage capacity, or transmission speed doubles in a few years. All have enjoyed some historical success. In no instance has the price per unit held constant.

For Investors

For an investor holding a multi-asset portfolio, the implications fall out of the argument. Duration looks different once one accepts that the rise in long yields is a real-rate phenomenon produced by an investment cycle; investment cycles always end, and this one likely ends in disinflation. The term premium investors are being paid today is compensation for a cyclical condition, and it may reverse moment the bond market stops financing negative free cash flow.

For equities the argument favors the users of intelligence over the sellers of capacity, and across the value chain it privileges links whose revenue is contracted on capacity for the next two years over links whose revenue depends on rampant growth in the premium tier of models. The case for caution is strongest where a valuation requires all the favorable assumptions to hold at once: sustained frontier premiums, exceptionally fast volume growth, high utilization, accommodating lenders, and timely delivery of power and equipment. A company can perform well operationally and still disappoint against that price.

A business with valuable data and a credible plan for reorganizing its operations may retain more of the productivity gain than the model vendor. That advantage still depends on competition in the user's own industry: an insurer or law firm can also elect to pass its savings to customers. "AI beneficiary" is too broad a category to settle the investment question.

The historical pattern we have written about earlier, in which the surplus from a transformative technology accrues to users and consumers while the financiers of the installation phase earn mediocre returns, has not yet been repealed by the speed of this cycle.[lxxxiii] The question to ask of any position in the chain is which point on the

revenue curve its price assumes, and whether it is financed by debt with a long enough term to wait.

The most consequential timing mismatches may be in credit, where trouble can emerge before equity investors accept that the boom has slowed. Oracle's credit-default swaps and CoreWeave's bonds are thus two important instruments to watch, and the seven data-center asset-backed securities that fell alongside CoreWeave's notes show how repricing follows the logic of guilt-by-association. This is probably not the Northern Pacific railroad, but the similarities are worth noting nonetheless.

We would keep an eye on a relatively short list of indicators:

- Real yields and the bid-to-cover ratio at inflation-protected auctions, which measure the bond market's patience more directly than credit spreads do, together with Oracle's five-year CDS and distance from a second downgrade, which are precisely tailored to the capital market's sentiment on this particular theme.
- Capital spending relative to operating cash flow, especially changes in guidance and the reasons cited by executives.
- Enterprise spending alongside token volumes: whether broader use produces growing revenue, and how various model tiers are pricing.
- Customer concentration, contract renewals, and refinancing terms at the most exposed lessors and their major tenants.
- The open-weight lag, which OpenRouter and Epoch track. A widening would restore the frontier's pricing power; a narrowing would accelerate routing.
- Layoffs, RIFs, "voluntary exit offers", and structural unemployment in sectors beyond software engineering.
- Power delivery, utility tariffs, and decisions to postpone or cancel planned campuses.

The most encouraging combination would be productivity spreading while the capital intensity of delivering AI falls. The more troubling combination would be weaker rental revenues, continuing equipment commitments, and lenders becoming less willing to bridge the difference. Both can coexist with impressive model releases.

After the Boom

The labs' request to coordinate is a sign of how uncomfortable the competitive race has become. The customer's search for cheaper intelligence is a sign of what may follow it. Neither means that the technology's capability gains are stalling out. They instead tell us that building the most capable model and earning the most attractive return

are becoming different objectives.

Carlota Perez's turning point, the passage from an installation phase led by financial capital to a deployment phase led by production capital, is recognized only afterward, and it has tended to be heralded by a change in the cost of money.[lxxxiv] Lenders have the power to brake the cycle of spending in a way that agreements on "frontier pacing" do not. Nor would that brake be the end of the astonishing gains in the technology. Those gains will simply more often take the form of efficiency, choosing the right model for the task, and putting it to work.

The railroad boom that broke in 1873 did not end railroad construction; and the two that followed were larger: more track was laid in 1882, and again in 1887, than in any year before the panic.[xxi] The first boom had correctly identified a technology that would change the country. It had financed that insight on terms many of its borrowers could not sustain.

For AI, the corresponding question is how quickly a productive technology becomes an affordable service, and who can remain profitable during the adjustment. The demand may be real. The contracts are already signed.

Appendix: AI-related Sub-Industries and Catalysts to Watch

Business	**Near-term support**	**Principal risk as the market changes**
GPU and memory suppliers	Orders already placed; scarce supply in important components	Customers reducing or deferring later orders; lower margins as capacity expands
Large cloud providers	Established businesses and contracted capacity revenues	Tenant credit, lower renewal prices, and poor returns on incremental construction
Specialist GPU lessors	Strong demand for currently available capacity	Debt and facility leases outlasting customer contracts or equipment economics
Model laboratories	Growing use, developer relationships, and enterprise distribution	Prices and customer spending failing to support committed capacity costs

Application providers and enterprise adopters	Cheaper models and more choice	Implementation costs, weak differentiation, and savings competed away downstream
Regulated utilities	Approved investments and pricing power up to allowed returns	Load forecasts failing, political resistance, or regulators disallowing costs

Important Disclosures

This article may not be copied, reproduced, or distributed without HB Wealth's prior written consent.

All information is as of the date above unless otherwise disclosed. The information is provided for informational purposes only and should not be considered a recommendation to purchase or sell any financial instrument, product, or service sponsored by HB Wealth or its affiliates or agents. The information does not represent legal, tax, accounting, or investment advice; recipients should consult their respective advisors regarding such matters. This material may not be suitable for all investors. Neither HB Wealth nor any affiliates make any representation or warranty as to the accuracy or merit of this analysis for individual use. This information contains forward-looking

statements, predictions, and forecasts ("forward-looking statements") concerning our belief and opinions in respect to the future. Forward-looking statements involve risks and uncertainties, and undue reliance should not be placed on them. There can be no assurance that forward-looking statements will prove to be accurate, and actual results and future events could differ materially from those anticipated in such statements. Certain information herein is based on third-party sources believed to be reliable, but which have not been independently verified. Past performance is not a guarantee or indicator of future results; inherent in any investment is the risk of loss. Specific investments described herein do not represent all investment decisions made by the above date. The reader should not assume that investment decisions identified and discussed were or will be profitable. Specific investment advice references provided herein are for illustrative purposes only and are not necessarily representative of investments that will be made in the future. Investors are advised to consult with their investment professional about their specific financial needs and goals before making any investment decision. HB may hold positions (long or short) in the companies mentioned in this paper.

---

[i] Dario Amodei, "We Must Pace the Frontier," September 12, 2026, as reported with quotations in Anthony Ha, "Anthropic CEO outlines plan to pace the frontier," TechCrunch, September 12, 2026 (updated with responses from Sam Altman and Elon Musk, and Brian Merchant's "regulatory capture" criticism). The waiver language, the embedded-evaluator design, Anthropic's unilateral commitment and the chip-export request are quoted or described there.

[ii] "California AG Bonta issues subpoena to OpenAI over AI cybersecurity risks," Reuters, October 1, 2026, describing the July 2026 Hugging Face intrusion by OpenAI agents during an internal test, the investigative subpoena issued by Attorney General Rob Bonta, and his call for congressional action.

[iii] American Action Forum, "Pacing the AI Frontier: An (Allegedly Illegal) Agreement," October 2026, describing the September 18, 2026 putative class action by four paid subscribers against Anthropic, OpenAI, SpaceX AI and Google (Google DeepMind) under Section 1 of the Sherman Act, and the July 2026 "intense competitive pressure" statement the plaintiffs cite.

[iv] Sam Altman, podcast remarks, summer 2026, on pacing that "does not feel like collusion among the frontier labs." See also Eduardo Porter's reading of the laboratories' argument as a prisoner's dilemma.

[v] "David Sacks criticizes 'Pace The Frontier' initiative as regulatory capture," Crypto Briefing, September 13, 2026 (paraphrasing Sacks's "textbook regulatory capture" characterization and his argument that rules holding everyone back favor those already at the frontier).

[vi] Aidan Gomez, quoted in "Cohere's Gomez says Silicon Valley giants shouldn't get to set the rules for AI," The Logic, September 2026.

[vii] John Sutton, Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration (MIT Press, 1991).

Glenn C. Loury, "Market Structure and Innovation," Quarterly Journal of Economics 93, no. 3 (1979); Partha Dasgupta and Joseph Stiglitz, "Uncertainty, Industrial Structure, and the Speed of R&D," Bell Journal of Economics 11, no. 1 (1980).

Avinash Dixit, "The Role of Investment in Entry-Deterrence," Economic Journal 90, no. 357 (1980).

[viii] HB Wealth, Bubbles and Breakthroughs (2025), on extended GPU depreciation schedules, Barclays' and Jim Chanos's estimates of overstated earnings, and Jensen Huang's remark on Hopper chips.

[ix] George J. Stigler, "A Theory of Oligopoly," Journal of Political Economy 72, no. 1 (1964). Where products are close substitutes, entry is possible and rivalry drives price toward cost, the private remedy is agreement, on price, on output or on the pace of the quality race, and the obstacle to agreement is that each member profits by cheating on it. This obstacle is the center of Stigler's analysis.
Robert H. Porter, "A Study of Cartel Stability: The Joint Executive Committee, 1880–1886," Bell Journal of Economics 14, no. 2 (1983); on the trunk-line pools and the Commission's early years, Paul W. MacAvoy, The Economic Effects of Regulation: The Trunk-Line Railroad Cartels and the Interstate Commerce Commission before 1900 (MIT Press, 1965). The railroads supplied evidence for Stigler's theory. The trunk lines formed pool after pool in the 1870s and 1880s to hold rates between Chicago and the seaboard; the Joint Executive Committee of 1879 to 1886. In Robert Porter's canonical study of cartel instability, the pools repeatedly collapsed into rate wars each time a member was suspected of shading its price.
Gabriel Kolko, Railroads and Regulation, 1877–1916 (Princeton University Press, 1965). When the pools kept failing the roads turned to Washington, and the Interstate Commerce Act of 1887 delivered, in Gabriel Kolko's contested but durable reading, the rate stability the carriers could not achieve among themselves. A request for an antitrust waiver so that competitors may agree on the rate at which they improve their product is the Joint Executive Committee asking the government to hold the pool together.
For a critique see Albro Martin, Enterprise Denied: Origins of the Decline of American Railroads, 1897–1917 (Columbia University Press, 1971), and Thomas K. McCraw, "Regulation in America: A Review Article," Business History Review 49, no. 2 (1975).
[x] HB Wealth, Bubbles and Breakthroughs (2025), on the Oracle–OpenAI contract and lease terms (drawing on Wall Street Journal and Bloomberg reporting), CoreWeave's lease and contract durations, and the sequence of telecom bond selloffs preceding equity declines in 2000.
HB Wealth, Technological Revolutions, Fast and Slow (Part 2 of the AI series, 2025), on the logistic form of technology diffusion; HB Wealth, Bubbles and Breakthroughs (2025), on Andrew Odlyzko's account of the fiber boom, in which traffic believed to be doubling every hundred days was doubling roughly every year; see Odlyzko, "Internet Traffic Growth: Sources and Implications," Proceedings of SPIE 5247 (2003).
HB Wealth, Bubbles and Breakthroughs (2025), citing Peter Berezin of BCA Research on large language models coming to resemble airlines, and on the network-effect economics of the prior platform generation.
[xi] "How Anthropic Clinched $517 Billion in Compute Deals in 11 Months," The Information, September 6. 2026, on Anthropic's compute commitments of up to $517 billion covering 14.8 gigawatts over eleven months
[xii] OpenAI's allocation of roughly 20% of inference compute to chain-of-thought monitoring.
[xiii] Goldman Sachs Global Investment Research, listing an intentional slowdown in frontier model development and data-center restrictions among constraints on 2027 capex growth.
[xiv] George J. Stigler, "The Theory of Economic Regulation," Bell Journal of Economics and Management Science 2, no. 1 (1971).
[xv] Stijn Van Nieuwerburgh, "Financing the AI Buildout" (working paper, October 2026), Table 1 (selected U.S. capital-expenditure booms as a share of GDP) and Table 5 (a 2026 estimate of $1,253 billion for the United States). The railroad peak is from Melvin Ulmer, Capital in Transportation, Communications, and Public Utilities: Its Formation and Financing (NBER, 1960), as charted by Justin Fox; the comparison of institutional estimates (Moody's $785 billion, S&P $750 billion, Morgan Stanley about $800 billion, UBS $820 billion, Goldman Sachs above $1 trillion, all global) was compiled by Azeem Azhar on October 13, 2026. All as reproduced in Smith (note 22).

[xvi] "Goldman Sees Hyperscaler AI Capex Rising 50% to $1.2 Trillion" Bloomberg News, September 25, 2026: five hyperscalers (Amazon, Alphabet, Microsoft, Oracle, Meta) on track for ~800billion in 2026, ~1.2 trillion base case for 2027 against ~$920 billion consensus, 1.4t trillion upside; ~300 billion of annual AI revenue needed to break even. See also Goldman Sachs, "Global AI investment is forecast to exceed $1 trillion in 2026."; S&P Global Ratings, "AI Infrastructure Investment To Exceed $1.3 Trillion By 2027," press release, August 27, 2026: negative free operating cash flow at Alphabet, Amazon, Microsoft, Meta, Oracle and SpaceX in 2026 and 2027, recovery not expected before 2029, models assuming a 2028 inflection.
[xvii] OpenAI's $1.4 trillion of announced infrastructure commitments: HB Wealth, Bubbles and Breakthroughs: How Should Investors Respond? (Part 4 of the AI series, 2025). The figure of roughly 26 gigawatts is .
[xviii] Epoch AI, "Hyperscaler capex vs. cash flow," Data Insight, 2026: exponential fits to SEC filings (Q2 2023–Q1 2026) for Amazon, Microsoft, Alphabet, Meta and Oracle; cash capex growing ~70% a year against operating cash flow ~23%; aggregate crossover around Q3 2026, with company-level crossover dates as given in the text.
[xix] Bloomberg, "Hyperscaler Debt Sales Skyrocket" (bond issuance by Amazon, Alphabet, Microsoft and Meta, including subsidiaries, 2010 to 2026 year-to-date) and "Hyperscalers Are No Longer Free Cash Flow Machines" (Bank of America and Bloomberg estimates of twelve-month forward free cash flow for Amazon, Alphabet, Meta and Oracle against Nvidia, Micron, Broadcom and Applied Materials), October 2026, as reproduced in Smith (note 22).
[xx] HB Wealth, The Great AI Buildout: Boom, Bubble, or Both? (2025), citing Jason Furman's calculation that first-half 2025 GDP growth would have been roughly 0.1% without data-center and information-processing investment.
[xxi] Noah Smith, "AI vs. the railroads: Are we in 1872?," Noahpinion, October 15, 2026. Smith cites Greg Ip (Wall Street Journal) on 9% of GDP, the NBER series on revenue per freight ton-mile for thirteen railroads (via FRED), the report of OpenAI's $50 billion run-rate against $68 billion expected, and the NBER series on miles of railroad built.
[xxii] HB Wealth, Bubbles and Breakthroughs (2025), drawing on Wall Street Journal reporting on Ellendale, North Dakota.
[xxiii] Minutes of the Federal Open Market Committee, September 15–16, 2026, and the Summary of Economic Projections, as reported in "Fed Minutes Show Officials Saw More Work to Do to Quell Inflation" New York Times, October 7, 2026: 12–0 vote to raise the target range to 3.75–4.00%, the first increase since July 2023; 16 of 18 participants projecting at least one further increase in 2026; language on AI investment pushing demand beyond supply and near-term construction and equipment demand outweighing the later supply benefit.
"Fed's Hammack Says Restrictive Rates Needed to Cool Prices," Bloomberg News, October 2026.
[xxiv] "Fed sees AI boom adding inflation headache; others may pay for the medicine," Reuters, October 19, 2026.
[xxv] Ben S. Bernanke, "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," American Economic Review 73, no. 3 (1983). On debt and the depth of post-boom recessions, see Òscar Jordà, Moritz Schularick and Alan M. Taylor, "When Credit Bites Back," Journal of Money, Credit and Banking 45, s2 (2013). Smith (note 22) links to this literature without naming the second paper; that attribution is ours.
[xxvi] S&P Global Ratings, https://www.spglobal.com/ratings/en/regulatory/article/-/view/sourceId/101694744. Fiscal 2027 capital-expenditure guidance of $90 to $95 billion; forecast free

operating cash flow deficit near $42 billion; adjusted leverage in the mid-4x range with $260 billion of lease commitments and $13 billion of purchase obligations added to debt; OpenAI about half of $638 billion in remaining performance obligations; downgrade triggers of sustained leverage above 4.5x or doubt about positive free operating cash flow by fiscal 2029.

"Oracle's Credit Default Swap Spread Hits a Record 288 Basis Points on AI Debt," Securitas Global, October 15, 2026 (five-year CDS at 288 basis points from 260 a week earlier; SoftBank at 424); see also "Oracle's Default Insurance Hits a Record as AI Debt Fears Spread," GuruFocus, October 2026, and "Oracle debt faces an AI stress test. Why some investors aren't worried," CNBC, October 19, 2026 Oracle's thirty-year bond spread and its 138% capex-to-operating-cash-flow ratio are from HB Wealth, Bubbles and Breakthroughs (2025).

[xxvii] "CoreWeave's slump spotlights growing burden of debt behind AI boom," Reuters, October 15, 2026: shares down about 15% on October 14 and 8.3% to $60.36 on October 15; five-year CDS above 1,000 basis points from above 600 the day before; some bond yields above 15%; about $35.1 billion of recourse and non-recourse debt, $51.6 billion including leases; cash of $5.524 billion; second-quarter property and equipment spending of $6.42 billion against $679 million of operating cash flow; backlog of about $104 billion; the inflation print and the fall in Oracle, Nebius and Iren as stated triggers. See also, CoreWeave shares down as much as 17% in a day; notes due 2032 below 78 cents on the dollar after being sold at par four months earlier; credit-default swaps implying a greater than 50% chance of default within five years; at least seven data-center asset-backed securities trading lower). "CoreWeave's Big Dream Crashes Into Reality as Yields Top 15%," Bloomberg, October 14, 2026.

[xxviii] "The $100 Billion Megadeal Between OpenAI and Nvidia Is on Ice," Wall Street Journal, January 31, 2026.

[xxix] Hanno Lustig, "Nine Percent of GDP," The Two Cents, September 27, 2026, setting out Van Nieuwerburgh's required-revenue arithmetic (note 16): 182.7 gigawatts of U.S. data-center capacity completed 2025 to 2032; $8.54 trillion of construction cost grossed up to a $9.61 trillion capital base; a 10% unlevered return; a six-year life for IT equipment (68% of cost) and twenty years for the rest; a 50% operating cash-flow margin; $3.725 trillion of required annual revenue by 2032, or 9.2% of GDP on the CBO's February 2026 nominal path. The post gives the year-by-year path (0.7%, 2.5%, 4.3%, 5.8%, 7.0%, 7.8%, 8.5%, 9.2% of GDP for 2025 to 2032) and the three-year IT-life sensitivity (1.2%, 4.0%, 7.0%, 9.3%, 11.2%, 12.6%, 13.6%, 14.7%); the observation that U.S. compute revenue is about $0.3 trillion today and must grow roughly 45% a year to reach $4 trillion; the implied 5.50perinstalledGPU-houratfullutilization(6.90 at 80%) against on-demand prices of $6 to $10 and more; the telecom benchmark (about $440 billion invested 1998 to 2003; about $190 billion of additional revenue, 1.7% of GDP, required by 2003; about $95 billion, 0.8% of GDP, delivered); the description of financing through project debt, private credit and 90%-levered special-purpose vehicles; CoreWeave's post-mid-2025 facilities (four deals, about $17 billion) syndicated to ten to thirteen banks; and the remarks quoted on asset lives and bank exposure.

[xxx] Greg Ip, "Will America Spend 9% of Its GDP on AI? The Industry Is Counting on It," Wall Street Journal, October 2, 2026, drawing on Van Nieuwerburgh. The comparison with national spending on energy, construction and the computer industry is as characterized by Smith (note 22); see also Columbia Business School, "Will America Spend 9% of Its GDP on AI?," October 5, 2026.

[xxxi] Paul Kedrosky, "AI Has Doubled Computing's Share of U.S. GDP," June 17, 2026, recalculating based on Epoch AI, "AI data center share of GDP": computing hardware at about 0.5% of nominal GDP for most of 2015 to 2022, approaching 1.6% by 2026, with non-AI compute hardware roughly flat at about 0.66%.

[xxxii] Jared Bernstein and Ryan Cummings, "AI Companies Are In a Race Against Time," 2026 (Bloomberg data and company filings).
[xxxiii] Goldman Sachs analysts, as reported in "Goldman Sachs Warns AI Stock Investors of New Risk," Yahoo Finance, September 4, 2026: the Silicon Data LLM Token Expenditure Index, a usage-weighted average price per million tokens, fell 29% in August to about $0.97, an all-time low and more than 50% below its May peak of about $2.05, with the compiler noting that the decline could reflect list-price cuts, migration to cheaper open-weight models, or both; OpenRouter token usage up nearly 50% month-on-month through much of 2026, with average prices paid collapsing during the same time—much of the collapse is attributable to caching (storing and reusing prompt strings) where tokens are heavily discounted, as described by Peter Walker, OpenRouter's Head of Insights; Rich Privorotsky cited in "AI Token Price Collapse: Goldman Sachs Warns the Unit Economics of the AI Boom Are Breaking" Tiger Trader; and 78 of 91 hyperscaler bonds issued in 2026 trading below issue price by the end of August.
[xxxiv] Ramp AI Index (Ara Kharazian), ramp.com/data, October 2026: indexed token volumes and AI spending among Ramp businesses, January to September 2026; share of tokens by model tier, July 5 to September 6, 2026; share of businesses changing their top model provider by token spend, October 2024 to September 2026; AI spend per employee per month for the top 1%, top 10% and median firms (posted October 14, 2026); share of U.S. businesses with paid AI adoption. As reproduced in Smith (note 22).
[xxxv] "10- and 30-year bond yields trading at 24-year highs," The Hill, September 30, 2026: 10-year at 5.304%, highest since May 2002; 30-year near 5.63%; Brent above $106; CBO FY2026 deficit. On the term premium at a ten-year high, see "Term Premium Is Back, and It Changes the Price of Leverage," Resonanz Capital, October 12, 2026.
[xxxvi] For FRED data on October 9, 2026: 10-year Treasury 5.22%; 10-year breakeven 2.33%; 10-year real rate 2.87%; 30-year 5.60%; effective funds rate 3.88%, see https://fred.stlouisfed.org/categories/22
[xxxvii] "10-year TIPS reopening gets real yield of 2.653%, highest in nearly 18 years," TIPS Watch, September 17, 2026.
[xxxviii] Associated Press, "Fed Chair Kevin Warsh warns AI is a shock to the economy, pledges to contain inflation," October 16, 2026 (remarks at an IMF conference in Bangkok): "obviously hitting the demand side of the economy before it's hitting the supply side"; "you've got to recognize this as a shock"; "a very bumpy one"; September CPI 3.7%.
[xxxix] Kevin Warsh, testimony before the Senate Banking Committee, July 16, 2026, as reported in "Fed's Warsh Cools Talk That AI Boom Is Spurring Inflation," Bloomberg.
[xl] Sebastian Mallaby, The Man Who Knew: The Life and Times of Alan Greenspan (Penguin Press, 2016), on the September 1996 FOMC debate; transcripts at the Federal Reserve's FOMC historical materials. The Fed left rates unchanged from February 1996 until a single increase in March 1997.
[xli] Erik Brynjolfsson, Daniel Rock and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," American Economic Journal: Macroeconomics 13, no. 1 (2021).
Paul A. David, "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox," American Economic Review Papers and Proceedings 80, no. 2 (1990); discussed in HB Wealth, Technological Revolutions, Fast and Slow (Part 2 of the AI series, 2025).
[xlii] Carlota Perez, Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages, Elgar (2002) describes an S-shaped (logistic) curve for the shape of technological adoption over time with a long assimilation or "installation" phase culminating in a relatively short "frenzy" before leveling off during a slower-changing "deployment" phase.

[xliii] Shakked Noy and Whitney Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," Science 381, no. 6654 (2023).
[xliv] Bennett School of Public Policy, University of Cambridge, "AI productivity gains," blog, 2026, summarizing studies of coding tools (20–40% gains; majority of new code machine-generated) and Stanford research on the commit histories of 100,000 developers.
[xlv] Stanford News, "Digital chores: generative AI's productivity boost at home," April 2026, reporting Stanford Institute for Economic Policy Research findings of 76–176% gains on household digital tasks. Stanford's analysis of the commit histories of 100,000 developers finds that gains shrink or turn negative on complex tasks in legacy codebases.
[xlvi] Rebecca Hinds and Paul Leonardi, "How Much Time Do Your Employees Spend Botsitting?" Harvard Business Review, August 5, 2006 and MIT Sloan, "10 levers shaping whether generative AI truly improves worker performance," Ideas Made to Matter, on "botsitting" and workflow redesign.
[xlvii] NBER Working Paper w34984 (2026). Federal Reserve Bank of St. Louis, "AI and productivity: what firms say on earnings calls," On the Economy, July 2026. Adoption itself is proceeding at a pace the bank compares to the spread of personal computers in the 1980s. Federal Reserve Bank of St. Louis (via FRBSF), "New survey findings on AI adoption and its effects on employment and productivity," October 2026.
[xlviii] NBER Digest, "Global evidence on business use of AI," May 2026: ~6,000 verified executives in the U.S., U.K., Germany and Australia; 89% report no measurable productivity impact over three years; mean effect +0.29%.
[xlix] This phenomenon is described in Brynjolfson, Rock and Syverson (2021)
[l] "OpenAI Faces New Lawsuits Linked to Shooting at Canadian School," Claims Journal, September 3, 2026, https://www.claimsjournal.com/news/national/2026/09/03/339938.htm (the February 2026 Tumbler Ridge attack; the allegation that OpenAI's system flagged the account about eight months earlier and that leadership declined the safety team's referral recommendation in June 2025; the April family suits and the September suits by teachers and students; Altman's public apology); "British Columbia Sues OpenAI Over Canada Mass Shooting Warning Failure," Bloomberg, September 21, 2026, https://www.bloomberg.com/news/articles/2026-09-21/british-columbia-sues-openai-over-canada-mass-shooting-warning-failure; "New lawsuits claim OpenAI execs put image ahead of safety in Canadian mass shooting," NPR, September 2, 2026, https://www.npr.org/2026/09/02/nx-s1-5953021/openai-tumbler-ridge-mass-shooting [not opened].
[li] 47 U.S.C. § 230; Garcia v. Character Technologies, Inc., No. 6:24-cv-01903 (M.D. Fla. May 21, 2025) (order denying in part the motion to dismiss); Raine v. OpenAI, Inc., filed August 26, 2025, San Francisco Superior Court.
[lii] Farwell v. Boston & Worcester Railroad Corp., 45 Mass. (4 Met.) 49 (1842); Morton J. Horwitz, The Transformation of American Law, 1780–1860 (Harvard University Press, 1977), for the subsidy reading, and Gary T. Schwartz, "Tort Law and the Economy in Nineteenth-Century America: A Reinterpretation," Yale Law Journal 90 (1981), for the case against it; Convention for the Unification of Certain Rules Relating to International Carriage by Air (Warsaw, 1929); Price-Anderson Nuclear Industries Indemnity Act of 1957, 42 U.S.C. § 2210; Communications Decency Act § 230 (1996); Digital Millennium Copyright Act, 17 U.S.C. § 512 (1998).
[liii] "The Underwriting Brake: How AI Exclusions Are Becoming the Real Constraint on Enterprise Automation," May 23, 2026, https://decision-intel.shapingtomorrow.com/scans/ai-automation/2026-05-23-ai-underwriting-brake/scan.html (ISO general-liability endorsements CG 40 47, CG 40 48 and CG 35 08, April 2026; Berkley's absolute AI exclusions in D&O, E&O and fiduciary cover; Munich Re, Armilla/Lloyd's,

Tokio Marine Kiln, CFC and Coalition products; Tufts Now on buyers unable to obtain limits). [Secondary; confirm the ISO and Berkley filings against the primaries.]

[liv] Uniform Electronic Transactions Act § 14 (1999); Electronic Signatures in Global and National Commerce Act, 15 U.S.C. § 7001(h) (2000); Restatement (Second) of Contracts § 153 (unilateral mistake); Moffatt v. Air Canada, 2024 BCCRT 149; Uniform Commercial Code §§ 4A-202 and 4A-204.

[lv] The IMF quotation can be found in Sonja Davidson and Hervé Tourpe, "How Agentic AI Will Reshape Payments"

"EPAA and HSBC launch working group on AI agent payment liability," July 20, 2026. On European Union product liability rule changes see "EU Updates its Product Liability Regime: Important Considerations for Providers of AI Systems and Software," Goodwin Law, February 2025.

[lvi] Scott Davis, "International comparisons show AI effect on productivity," Federal Reserve Bank of Dallas, July 7, 2026: U.S. labor productivity 2.4% annualized since Q1 2024 vs. 1.6% in the five pre-pandemic years; AI-exposed sectors (information; finance and insurance; professional and technical services) 3.7% vs. 1.7% for the rest; 16% of hours, 40% of gains; exposure index from Felten, Raj and Seamans; caveats as described.

[lvii] International Monetary Fund, "Generative AI and the Limits of Productivity Measurement in the System of National Accounts," Working Paper, September 25, 2026.

U.S. Bureau of Labor Statistics, "AI and productivity," Office of Productivity and Technology.

[lviii] Ian Dew-Becker and Robert J. Gordon, "Where Did the Productivity Growth Go? Inflation Dynamics and the Distribution of Income," Brookings Papers on Economic Activity 2005:2; see also Ludovica Ambrosino, Jenny Chan and Silvana Tenreyro, "Productivity and Inflation Dynamics," working paper.

[lix] Joshua Gibson, "148 gigawatts announced, 12 running: AI financing outran the grid," FXStreet, September 3, 2026, drawing on Sightline Climate (777 hyperscale projects above 50 MW announced since 2024; 12.3 GW operating, 20.6 GW under construction; the 2025 and 2026 delivery cohorts), Lawrence Berkeley National Laboratory's interconnection-queue data (about 8,200 projects and more than 2,000 GW queued; median wait above five years for projects reaching operation in 2025), and Wood Mackenzie (128-week lead times for large power transformers and 144 weeks for generator step-up units; prices 70 to 150% above 2020; more than 80% of large transformers imported); the $150 to $200 billion of 2026 capex expected to slip is the author's estimate.

[lx] "Data center boom fuels PJM power cost increases as states move to shield ratepayers," Data Center Dynamics, July 29, 2025 (bills across the thirteen PJM states rising more than 5%; Columbus residential bills up $27 a month; Ohio's 85% minimum-take settlement; Oregon's POWER Act high-load customer class; Pennsylvania's proposed Ratepayer Protection Act); see also "Energy bills likely to tick up again in 2026 after electricity auction clears at maximum price," Maryland Matters, July 23, 2025. [These figures refer to the 2025/26 and 2026/27 capacity auctions; update with the most recent PJM auction and the independent market monitor's attribution before publication.]

[lxi] Tom Klein, "As They Consider New Data Centers, States Target Water Use," National Conference of State Legislatures, October 8, 2026 (Virginia HB 496/SB 553 monthly reporting effective January 1, 2027; proposals in California, Iowa, South Carolina, Illinois and Georgia).

Rockefeller Institute of Government, "Updates on the Cloud: More Moratoriums on Data Centers," June 10, 2026 (fourteen states having considered moratoriums; New York AB 10141/SB 9144, a one-year pause passed June 4, 2026; Maine LD 307 vetoed April 24, 2026; local moratoriums or bans in Groton, Connecticut, Peculiar, Missouri, Bangor, Maine and Lysander, New York); Matthew Gooding, "Memphis council delays decision on year-long data center moratorium despite protests," Data Center Dynamics, October 12, 2026 (7–5 vote to postpone; litigation over the gas turbines at the SpaceXAI, formerly xAI,

campus; police clearing the chamber); "Memphis leaders reject contentious 18-month pause on new data center development," CNN, October 20, 2026

[lxii] Paul Berg et al., "Summary Statement of the Asilomar Conference on Recombinant DNA Molecules," Proceedings of the National Academy of Sciences 72, no. 6 (1975); National Institutes of Health, Guidelines for Research Involving Recombinant DNA Molecules, June 1976.

[lxiii] Sam Peltzman, "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments," Journal of Political Economy 81, no. 5 (1973).

[lxiv] "Three Mile Island still haunts U.S. nuclear industry" New York Times, March 2009.

[lxv] Andrew Odlyzko, "Collective Hallucinations and Inefficient Markets: The British Railway Mania of the 1840s" (working paper, 2010); Gareth Campbell and John D. Turner, "Dispelling the Myth of the Naive Investor during the British Railway Mania," Business History Review 86 (2012); D. Morier Evans, The Commercial Crisis 1847–1848 (1849), on Bank Rate reaching 8% in October 1847.

lxvi Richard White, Railroaded: The Transcontinentals and the Making of Modern America (W. W. Norton, 2011), on the failure of Jay Cooke & Co. on September 18, 1873.

[lxvii] "Private credit now dominates life insurer bonds as NAIC rewrites the rulebook," Insurance Business, August 10, 2026, citing S&P Global Market Intelligence (privately placed bonds 48.4% of life-industry bonds at year-end 2025, from 37.4% five years earlier; a general-account bond portfolio of nearly $4 trillion); NAIC Capital Markets Bureau, "U.S. Insurance Industry's Asset Mix, Year-End 2025." The 21% growth in private-credit holdings is as reported by the Wall Street Journal and cited in Kedrosky (note 71).

[lxviii] Paul Kedrosky, "Where AI Debt Is Going, and Why It Matters: Insurance Edition," August 20, 2026 (sponsor-controlled insurers; Nvidia's financing platforms with Apollo, BlackRock, Blackstone, Brookfield, Goldman Sachs and KKR targeting more than $500 billion of third-party capital; the SEC staff position that certain data-center securitizations are not asset-backed securities, with five-year anticipated repayment dates against 25-to-30-year final maturities; Delaware Life's reclassification of $17.8 billion of affiliated private credit, about 40% of invested assets, following grand-jury subpoenas).

[lxix] Life risk-based capital C-1 bond factors as revised in 2021 (approximately 2.2% for NAIC 2.C, 3.2% for NAIC 3.A, 7.4% to 9.5% for NAIC 4.A and 4.B, pre-tax); Company Action Level at 200% of Authorized Control Level capital. On negative IMR, INT 23-01 and its extension through December 31, 2027, see "Evolutionary, Not Revolutionary: The Looming Change in Bond C1" Conning, September 2021.

[lxx] "Credit braces for fallen angels as US$100 bil trades like junk," Bloomberg via The Edge, August 2026 (about $100 billion of dollar and euro investment-grade bonds trading wider than the double-B curve; an ICE BofA fallen-angel index at its highest since early 2024). Oracle's debt above $100 billion is from HB Wealth, Bubbles and Breakthroughs (2025). The size of the U.S. high-yield market (about $1.4 trillion) and Ford's roughly $36 billion of index-eligible debt at its March 2020 downgrade are Index eligibility under the Bloomberg U.S. Corporate Index turns on the middle of the Moody's, S&P and Fitch ratings.

"FSB Watchdog Flags Risks in Banks' Growing Private Credit Ties," Insurance Journal, May 6, 2026; "Federal Reserve examines U.S. banks' exposure to private credit firms and insurers," Fortune, April 10, 2026; Lustig (note 32) on the syndication of CoreWeave facilities to ten to thirteen banks since mid-2025.

[lxxi] AllianceBernstein, "Regulatory Reboot Could Transform Taiwanese Insurers' Portfolios," November 2025.

[lxxii] OpenRouter, "The Open Weight Models that Matter: June 2026," June 27, 2026: a consistent three-to-six-month gap behind U.S. frontier labs for more than eighteen months; Nemotron 3 Ultra the strongest U.S. open-weight entrant; GLM 5.2 ranked first among open-weight models.

[lxxiii] "OpenAI launches two 'open' AI reasoning models," TechCrunch, August 5, 2025 (gpt-oss-120b and gpt-oss-20b, Apache 2.0). Altman's "wrong side of history" remark was made in a Reddit discussion in January 2025.
[lxxiv] "Meta releases Muse Spark Open, a new 'open-weight' AI model" The Guardian, October 15 2026, on Muse Spark Open. The model's reported size (816 billion parameters, 42 billion active) and the spring decision to launch Muse Spark closed are
[lxxv] Nvidia's release of Nemotron training data, recipes, evaluation tooling and reinforcement-learning infrastructure alongside the weights.
[lxxvi] "AI Models Now Have a 119x Price Gap: Are You Overpaying?" Sociodigitals, September 28, 2026, on gated security tiers and the price points cited.
[lxxvii] Sarah Su, Kevin Zhu, Emily Xiao, Rohan Alur and Daniel Kang, "Learning to Replicate Expert Judgment in Financial Tasks," Bridgewater AIA Labs, June 30, 2026 (fine-tuned Qwen3-235B using Thinking Machines' Tinker; 84.7% vs. 78.2% for the best frontier model with expert prompting; ~50% with naive prompting; 13.8x reduction in inference cost per task).
[lxxviii] Gartner, "Gartner Says More Than 50% of Enterprises Will Use Industry-Specific Generative AI Models by 2027," press release, April 17, 2024.
[lxxix] "Alex Karp, frontier models and the real fight for enterprise AI," SiliconANGLE, July 5, 2026, including Jensen Huang's GTC remarks.
[lxxx] HB Wealth, Bubbles and Breakthroughs (2025), on the DeepSeek episode of January 2025 and Satya Nadella's invocation of Jevons.
[lxxxi] Anthropic's quarterly revenue disclosures and the TickerTrends estimate of a $76.8 billion annualized run-rate as of September 24, 2026, as reported in Smith (note 22)
[lxxxii] Corey Quinn, "Amazon sends chills through its workforce with mixed-message memo about AI use," The Register, October 31, 2026, https://www.theregister.com/ai-and-ml/2026/10/31/amazon-sends-chills-through-its-workforce-with-mixed-message-memo-about-ai-use-corey-quinn/5306804 (KiroRank shut down in May after gaming; Meta's Claudenomics; the Terkowitz memo circulated in September and published in late October; "tokenmaxxing"; input metrics "deeply flawed").
[lxxxiii] HB Wealth, Bubbles and Breakthroughs (2025), on high-bandwidth memory capacity sold out through 2026, TSMC advanced packaging booked through the planning horizon, and the historical distribution of surplus between users and financiers of transformative technologies.
[lxxxiv] Carlota Perez, Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages (Edward Elgar, 2002).