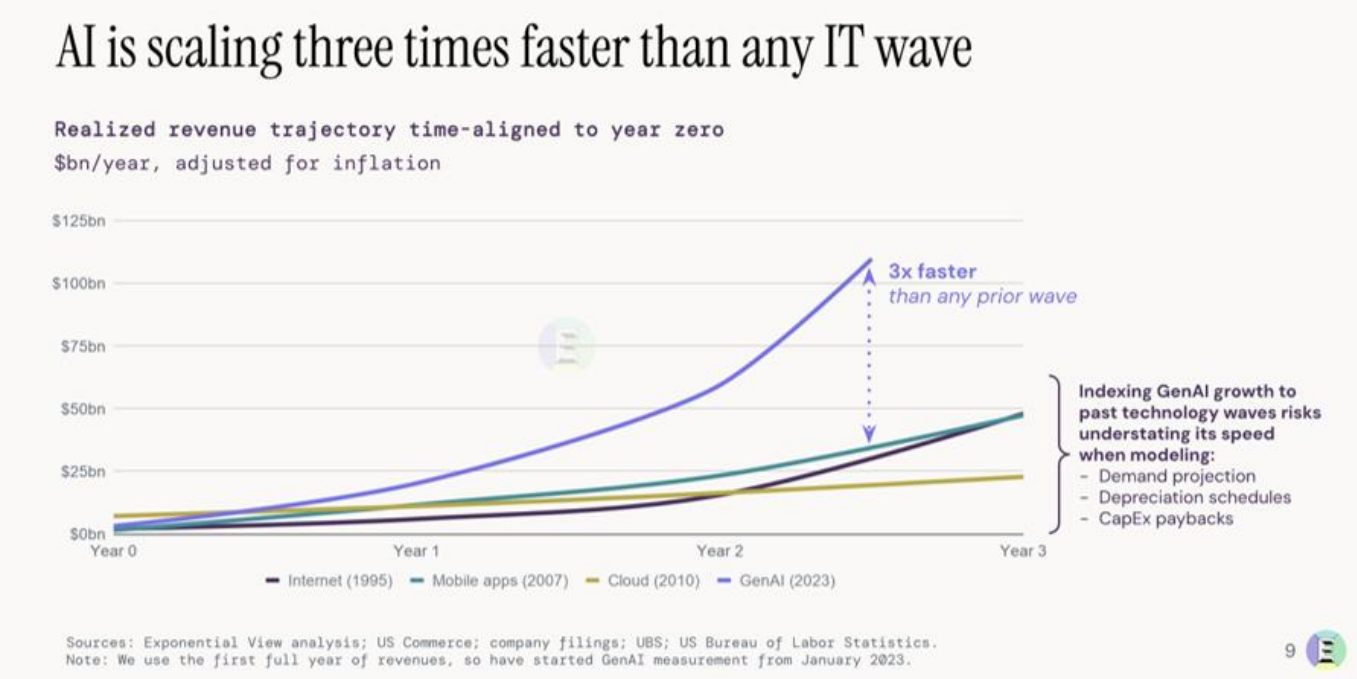
Powering the Nasdaq-100®'s AI Thematic Engine: Top Drivers and Headwinds in '26

Mark Marex, CFA, *Global Head of Index Insights*, Nasdaq Global Indexes

At a Glance:

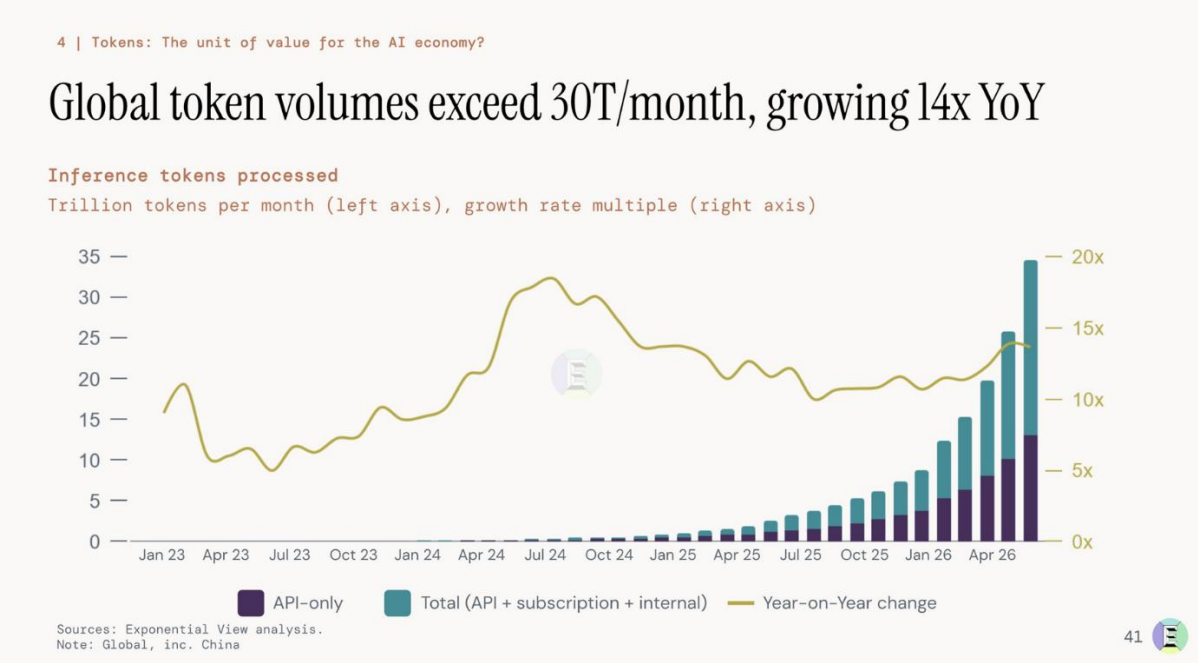
- AI is scaling at an increasingly faster exponential rate, exceeding the rates of raw adoption and monetization seen across prior major technology waves
- The uniquely powerful convergence of exponential adoption curves, exponential performance improvements, and exponentially declining costs are reinforcing each other to improve economics across multiple layers of the AI value chain
- Common headwinds cited by skeptics such as the risk of technological obsolescence, the risk of overshooting supply, and the risk of excessive leverage, remain poorly modeled and unsubstantiated
- Instead, headwinds on supply expansion are clearer and are more likely to prevent excess capacity, while also limiting the odds of an unfavorable boom/bust scenario

AI is a Series of Exponential Technological Growth Curves



Global demand for AI compute is accelerating along an exponential curve in 2026. Per the analysis in the above chart from Exponential View, the realized revenue trajectory for the AI industry is

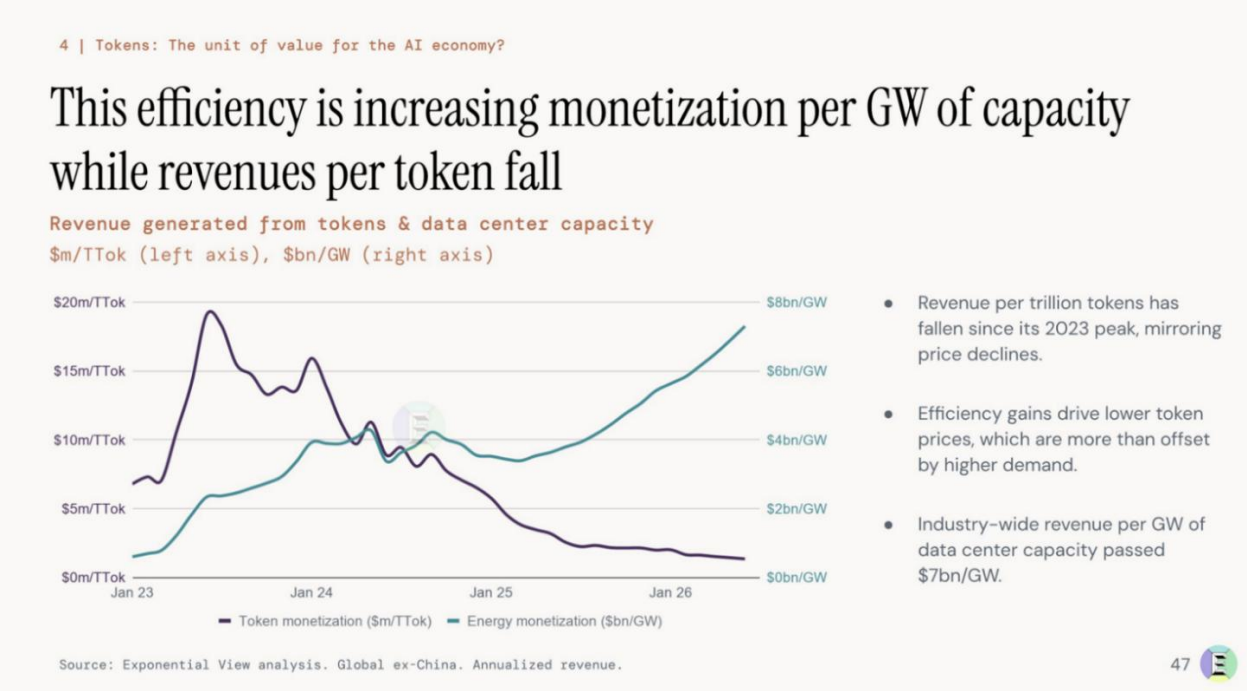
already pacing well ahead of three major technology waves (Internet, mobile apps, cloud) by 3x faster or more, with a clear upward inflection this year. As of June 2026, global token volumes have exceeded 30 trillion per month, growing 14x year-over-year, as seen in the chart below. Notice the exponential curve there as well – including how inference tokens were nearly nonexistent throughout 2024/2025 – as most AI compute demand was still concentrated in model training in the early years.



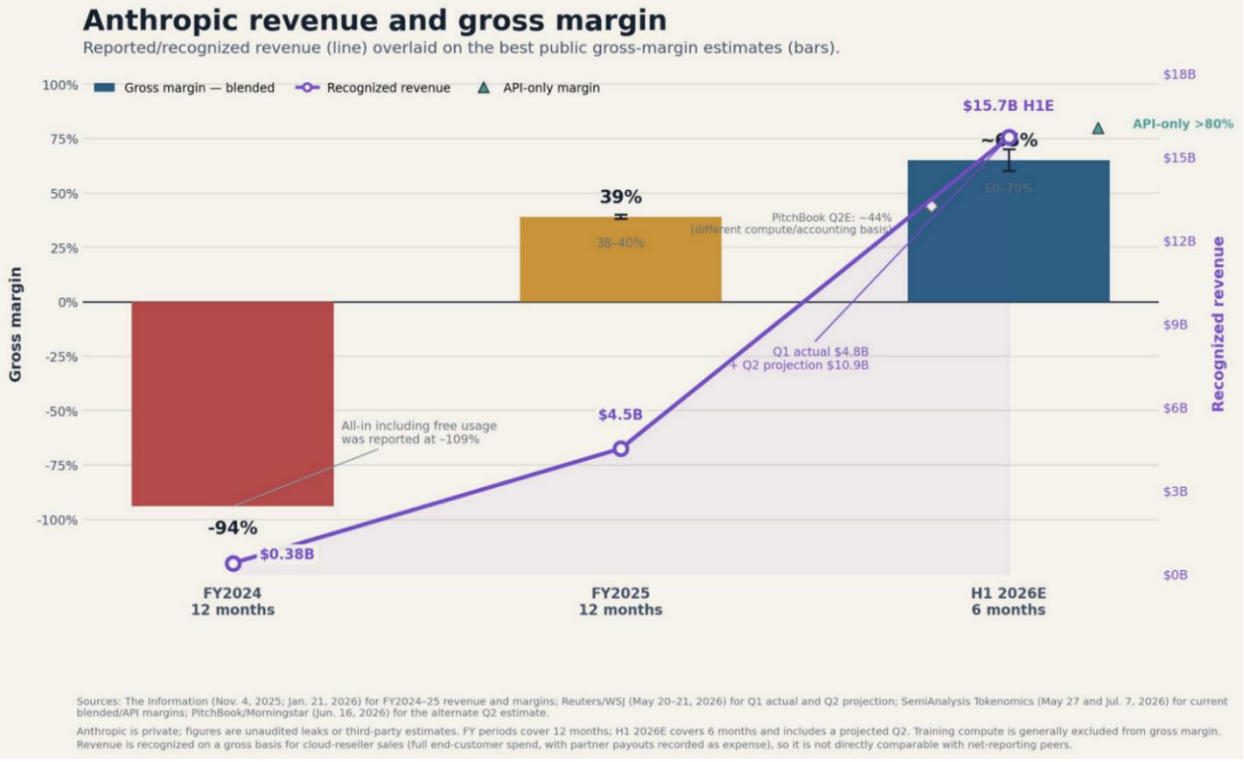
Now that models have evolved to a level of sophistication that enables the ongoing processing of tokens for hours or even days at a time, the AI industry's value proposition is shifting from *generative* output to *agentic* output. In other words, it's less about a model creating a static *output* (text, image, video, sound) in response to a prompt, and more about creating a dynamic *outcome*. Agentic AI is thus the next major leg of demand growth, and we are already seeing the highly favorable impacts on ROI extend throughout various layers of the AI value chain. At the model layer, frontier players like Anthropic have been reporting rapid acceleration in their revenue trajectories based on internally tracked metrics, growing from annualized recurring revenue of ~$9B at the beginning of 2026 to ~$65B by July, possibly exceeding $100B by end of year.

At the infrastructure layer, leading hyperscalers Microsoft and Amazon saw their share prices rally by 25% each at the end of July in response to historically strong growth rates in their cloud services businesses (43% and 37% year-over-year, respectively) alongside capex commitments of nearly $200 billion each. Google parent Alphabet reported cloud revenue that grew more than 80% year-over-year. Overall, AI revenue across the foundational model, infrastructure, and application layers is compounding at more than $1B every two days in 2026, approximately 90x faster than it was in 2023. At the same time, revenue per token continues to collapse at an exponentially downward rate. This is the first corollary to exponential AI adoption curves: as the technology gets cheaper and more efficient each year, more users and use cases flood into the ecosystem, which more than offsets price declines to drive industry revenue up its own exponential curve. Industry-wide revenue

per gigawatt of data center capacity recently passed $7B, nearly doubling in the past 18 months:

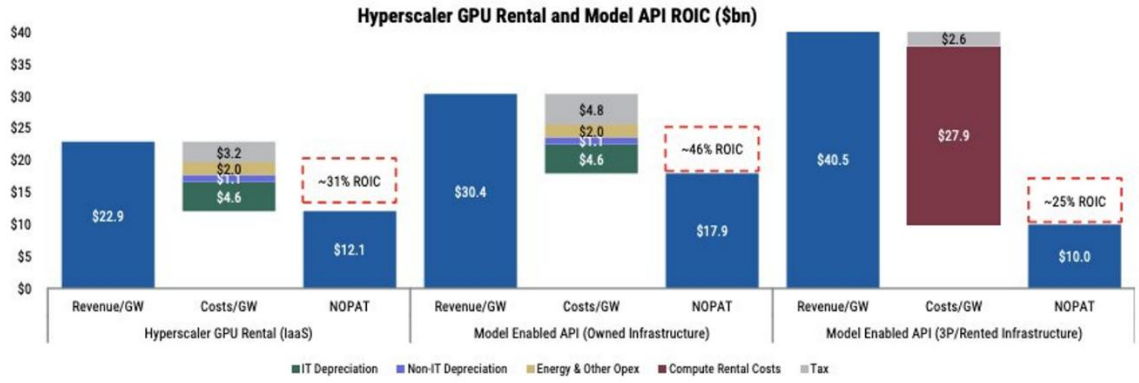


And while the revenue story is compelling on its own, the all-important ROI story is getting more compelling as well. In the first half of 2026, for example, Anthropic's gross margins have jumped to a range of ~60-70% vs. ~40% for FY2025; margins were deeply negative as recently as 2024:



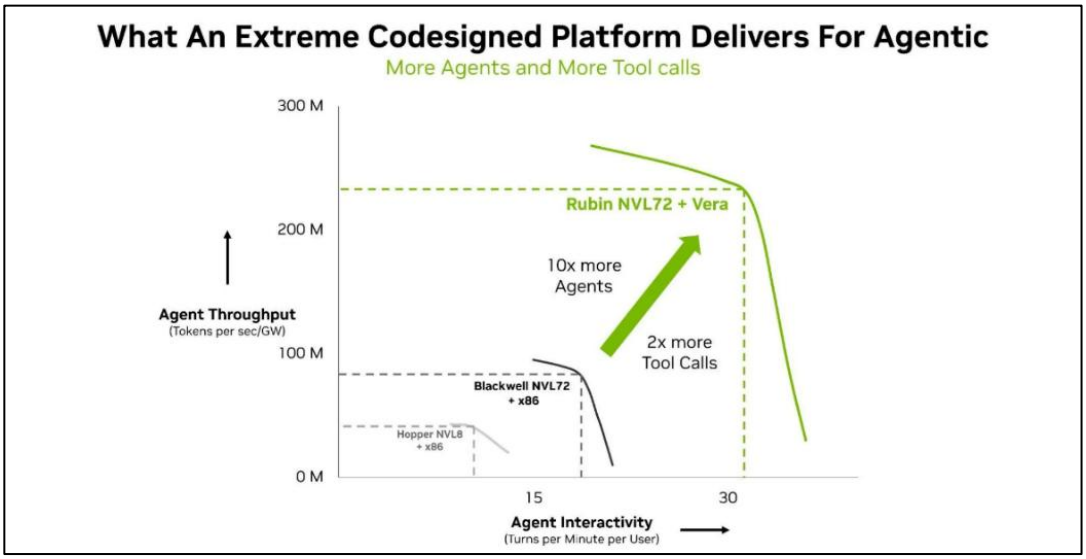
As a result of the ability to deliver higher-quality AI model outputs more quickly and at ever larger volumes, returns on investment across the value chain are going higher. Amazon recently made news by raising prices on GPU instances within its AWS unit by 20%. In July, Morgan Stanley Research released their updated framework for three different AI business models, ranging from ~25% return-on-invested-capital (ROIC) for models running on third-party infrastructure, to 31% for the hyperscaler GPU rental model, to 46% for models running on owned infrastructure. For hyperscalers specifically, Morgan Stanley sees ~67% incremental margins on new revenue – after deducting the costs of IT Depreciation, Non-IT Depreciation, Energy and Other Opex.

Exhibit 1: We see attractive incremental unit economics and ROIC (25%+) on the three main business models set to emerge through the age of AI inference.



Source: Company data, Morgan Stanley Research Estimates. Note the "ROIC" on a Model-Enabled API running on 3P infrastructure is illustrative and represents revenue, adjusted for training capacity (training costs typically expensed in opex) less the rental cost of compute (COGS), taxed at 21%.

This is all before the next generation of Nvidia GPUs hit the datacenter market, with a measured performance improvement of as much as ~10x on the Vera Rubin NVL72 rack-scale computer system vs. the prior generation of GB200s – which is expected to drive even more downward pressure on the cost to generate and serve tokens, likely driving even greater adoption, revenue and profitability.



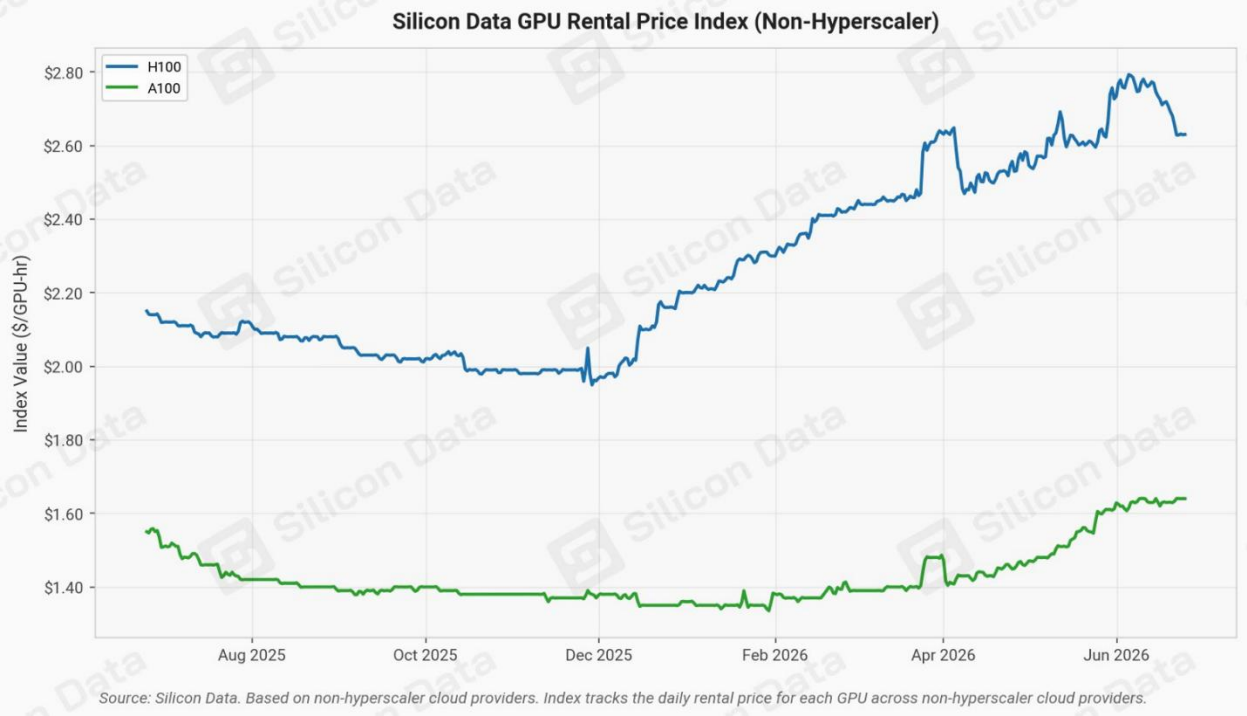
We previously wrote about this dynamic which seems unique to modern-day technological

innovation, calling it the “four factors of technology scaling.” The evolution of the AI wave is demonstrating it in real time.

How Recent Data on GPU Performance & Pricing Refutes AI Capex Skepticism

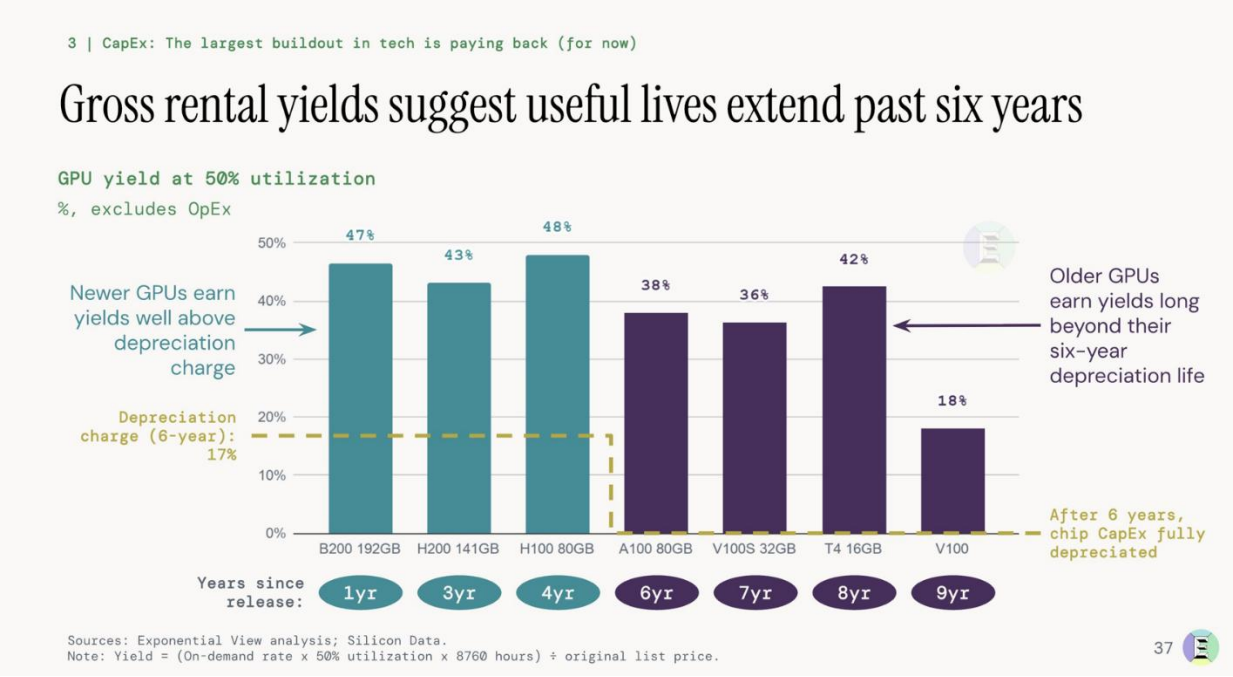
For professional and casual observers alike, the scale of the AI wave may seem too daunting to fully comprehend. Among the four major Nasdaq-listed hyperscalers alone (Microsoft, Amazon, Google and Meta Platforms), AI-related capex is approaching $800 billion in 2026 (based on reported actuals in 1Q/2Q and company guidance for 3Q/4Q) – nearly equal to the annual US defense budget. Nvidia is organizing external pools of capital to the tune of $500 billion over the next several years, solely for the purpose of lightening the financing load on the hyperscalers to enable even faster buildout of AI datacenters. For those who lived through the Internet/telecom bubble in the late 1990s, an easy analog presents itself. Surely the AI buildout is overshooting with these numbers, is it not? Surely the major spenders will be big losers tomorrow, as they were back then?

At the end of 2025, we presented analysis that did not support such comparisons at a high level, based purely on comparing index-level valuations and profitability metrics for the Nasdaq-100 between the two periods. In some respects, the picture looks even less concerning today than it did back then. The AI wave’s biggest beneficiary, Nvidia, is trading at a forward P/E of ~23 as of mid-August – a discount of more than 40% off its trailing 5-year and 10-year averages. Well then, if it’s not a valuation issue, surely there must be too much circular financing going on with Nvidia? And even it’s all real, AI’s technological advancements will render today’s investments worth relatively little in just a few years, prompting huge writedowns? Aren’t we already seeing the negative impacts of all this wasted investment reflected in the erosion of hyperscaler free cash flows this year?



Let’s start with the point around technological obsolescence, because it remains widely

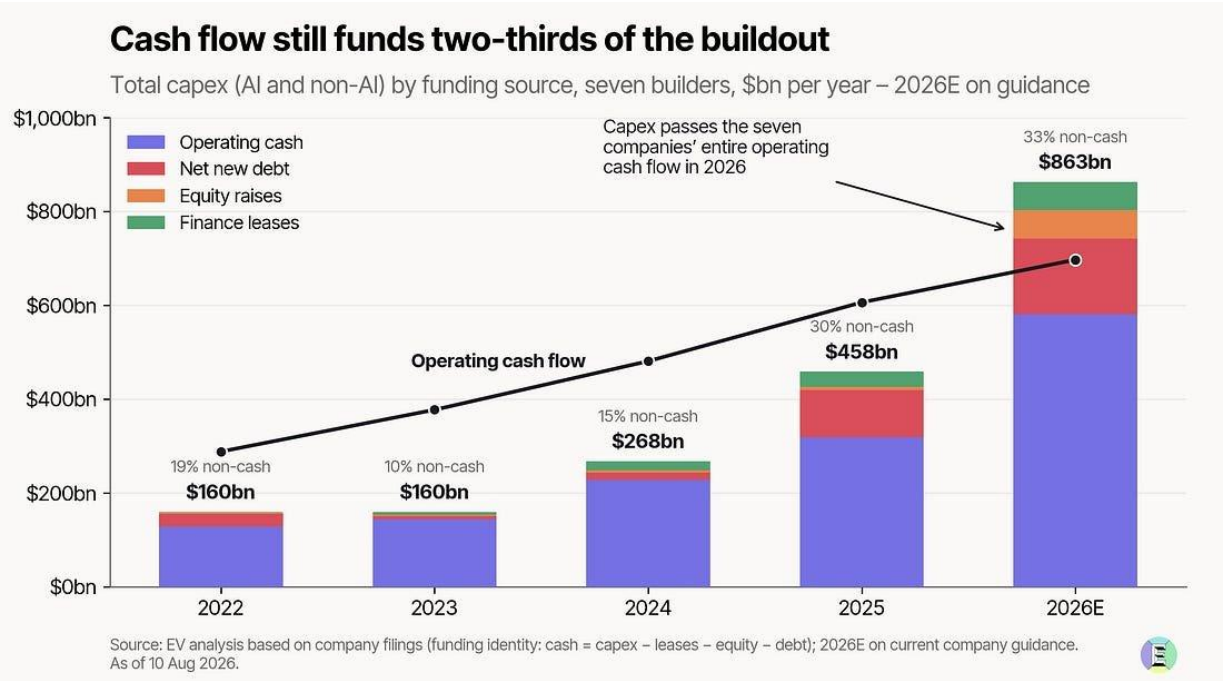
misunderstood. While there have been anecdotal datapoints in recent years to suggest that GPU depreciation is not like traditional capex depreciation, there are now institutional-quality indexes that measure just how well GPUs are retaining their value. Note in the above chart how rental rates for Nvidia's H100 chips began to surge upward at the beginning of 2026, followed by the surge in even older A100 chip rental rates in Q2. In the case of A100s, these are chips approaching or just past their fully-depreciated projected lifespan of six years. CoreWeave, a leading neocloud, recently announced on its quarterly earnings call that they plan to extend the useful life of these chips to nine years. For a fully depreciated asset, that incremental revenue can now flow through at an accelerated rate into net profits. So, not only were many companies and analysts missing extra years of useful life (and thus revenue) in their projections of AI capex economics; they were also implicitly overshooting depreciation costs, thus undershooting profit and returns on investment! The overarching driver of this phenomenon is the strength of continuing demand for AI compute observed in 2026. But it is also the fact that these chips are responding to intellectual property enhancements, such as software upgrades, that maintain their productivity over longer time horizons – as well as operational innovations by their owners who continue to optimize the mix of chips, networking, power flow and cooling to squeeze out the maximum possible yields, which continue to increase with each newer generation of chips sourced into the datacenter market:



This is, simply put, a qualitatively different type of physical asset, with an unpredictable and non-straight-line depreciation schedule. For traditional physical assets, it is a much more straightforward exercise to estimate how long a physical structure, appliance, or building component should last. Unlike GPUs, one cannot download a software update for a double-hung window, and the demand does not grow exponentially every year.

In terms of cashflow, it is true that more and more of free cash flow and operating cash flow continues to go into AI-related capex by the hyperscalers. The question investors must reckon with is: compared to what opportunity cost? Is the strongly positive ROIC on hyperscaler GPU rentals

better than the average ROIC on all other investments any of these companies can make today? Even if the hyperscalers' AI ROIC is somewhat lower, does the sheer scale of the AI investment outweigh the lower aggregate opportunities elsewhere, even if those exist at higher rates of return? It's certainly appearing more lucrative than what these companies used to do in prevailing times, which was either accumulate large piles of cash earning the risk-free rate of return on their balance sheets, or increase the return of capital to shareholders via share buybacks.



Despite the many recent headlines emphasizing the negative impact on hyperscaler cashflows, the reality is that in 2026, the capex bill is *still* being two-thirds funded by operating cash – just slightly worse vs. 2025 on a percentage basis (33% vs. 30% non-cash), as cash generation continues to increase strongly in aggregate. This is in spite of a pronounced timing mismatch in capex spending vs. the payoff, as datacenters take a substantial amount of time to construct and bring online. The cash flows generated this year are largely sourced by capex spent two or more years ago, on less efficient and less powerful GPUs. The capex spent this year – if Nvidia is to be believed about the 10x increase in throughput per second per megawatt on Vera Rubin – should in theory generate *even more favorable payoffs* in terms of revenue, profit, and cash flows in 2028 and beyond. Per leading AI research provider Epoch AI, the performance per dollar of AI chips has improved by at least 49% per year, on average, since the beginning of 2023. That is the baseline.

In the meantime, absolute leverage can continue to increase but remain low on a relative basis. We analyzed leverage dynamics in the Nasdaq-100 earlier this year, and noted no measurable change in the important debt ratios up to that point (and in some cases, an *absolute improvement* in ratios vs. multiyear trends). As of the most recent quarter ended June 30, 2026, that trend continues: per Bloomberg, NDX Index-weighted Net Debt to EBITDA dropped further to 0.85 from 0.88 as of December 31, 2025; Total Debt to Enterprise Value held at 0.07; Total Debt to Total Equity dropped to 64.4 from 66.8. Total borrowings have certainly risen YTD, but so has EBITDA; so have market capitalizations; and so have a number of other critical financial indicators that measure the capacity

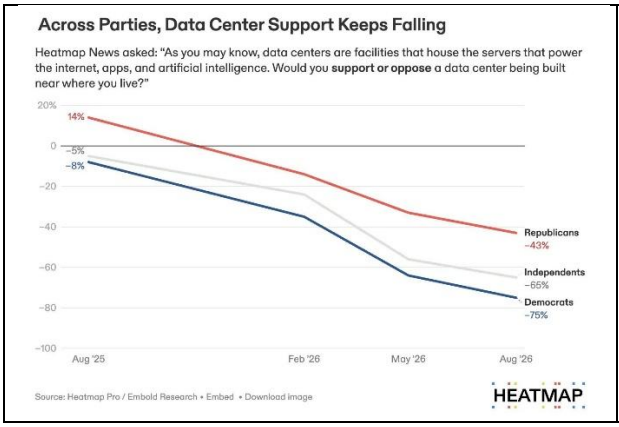
by these firms to borrow to fund more of the AI buildout – gargantuan at the beginning of this wave with minimal initial debt, and continuing to increase with seemingly every passing year.

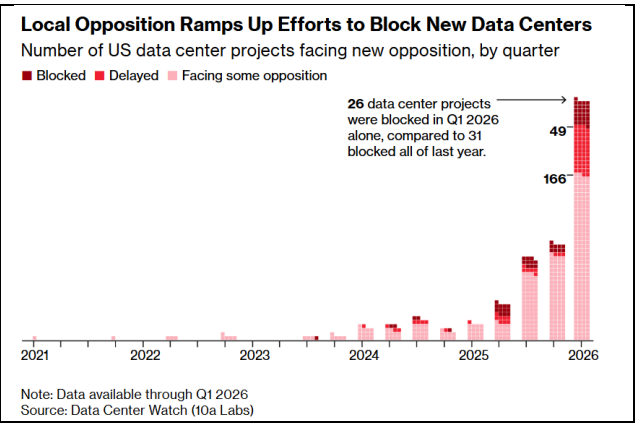
Circular financing concerns with Nvidia should abate following the recent announcement of a pool of $500 billion of external capital committing to finance AI datacenter capex, including GPU purchases. This sets a hard cap on the extent to which Nvidia leverages its balance sheet for purposes of enabling AI capex going forward.

Concerns about the competitive risk of open-source AI models will likely continue. But at a minimum, the market is potentially beginning to understand that the economic upside – for Nvidia and its peers in the semiconductor ecosystem, as well as the hyperscalers serving those models to customers, and application builders alike – actually expands no matter how tokens are consumed or produced. Whether a token is open or closed-source, it still requires the same chips and associated infrastructure to produce, and in some cases more. The battle for model sophistication, efficiency and cost will likely be fought over the course of many years, and while the biggest risk to further upside resides with the closed-model frontier players, their ability to remain at the frontier preserves their pricing power, as more and more open-source-powered AI agents proliferate and require "higher IQ" teachers, coordinators, and supervisors to help them deliver value-add.

What Remaining Headwinds vs. Upside Scenarios Exist for AI and the Nasdaq-100?

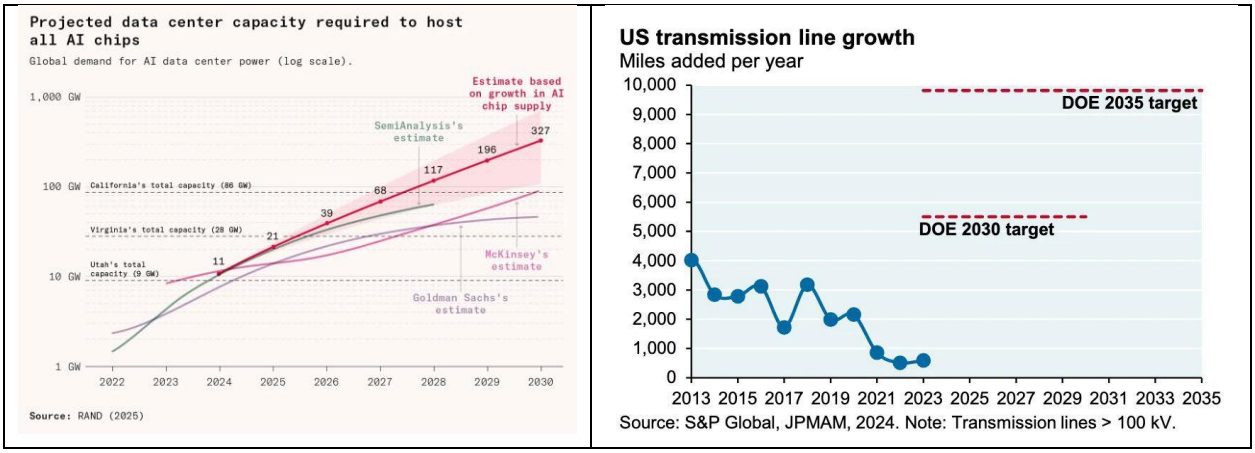
Ironically enough, of all the concerns to be "worried" about for AI investors, the regulatory one may be the biggest on its face – but in reality, a potentially useful silver lining. Just in the past few weeks, explicit or implicit moratoriums on new datacenters have been placed in three of the biggest states: New York, Pennsylvania, and Texas. A number of others have considered legislation and failed to pass it, or are in the process of advancing it. By adding an artificial brake pedal on the speed of datacenter construction, US state governments may extend the runway of AI capex spending, making it more sustainable to finance over a longer time horizon, and prevent an excessively quick buildout leading to excess supply and a potential "bust" scenario.



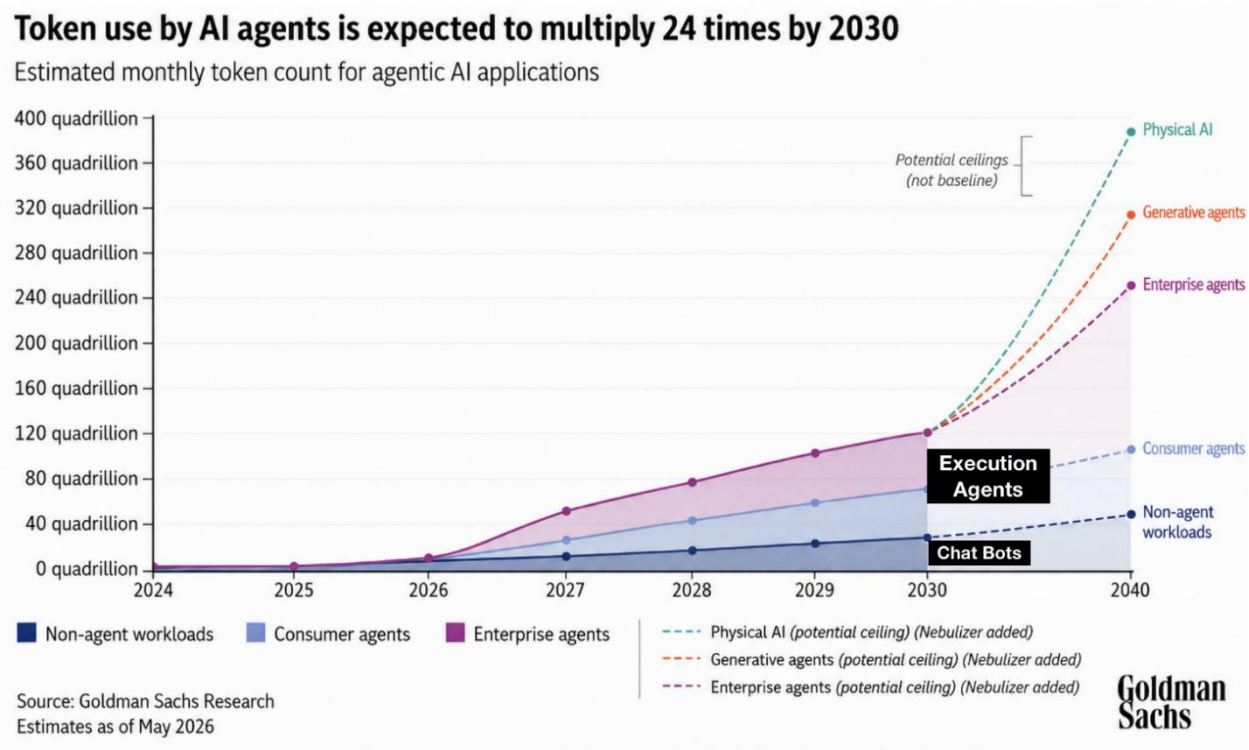


The other major headwind is, of course, power – both in terms of sourcing the raw generating capacity needed to power a California's worth of new electricity demand from projected datacenter build, and in terms of connecting enough of that power to the grid in time. After years of flatlining

growth in electricity demand in the US, national transmission line growth plummeted which, when combined with legacy regulatory hurdles, now leads to five years on average of interconnection queues across most of the country for new power generation and storage capacity. Even off-grid, "behind the meter" power generation is now severely supply-constrained, with industrial-scale gas turbines sold out for years in advance due to datacenter demand.

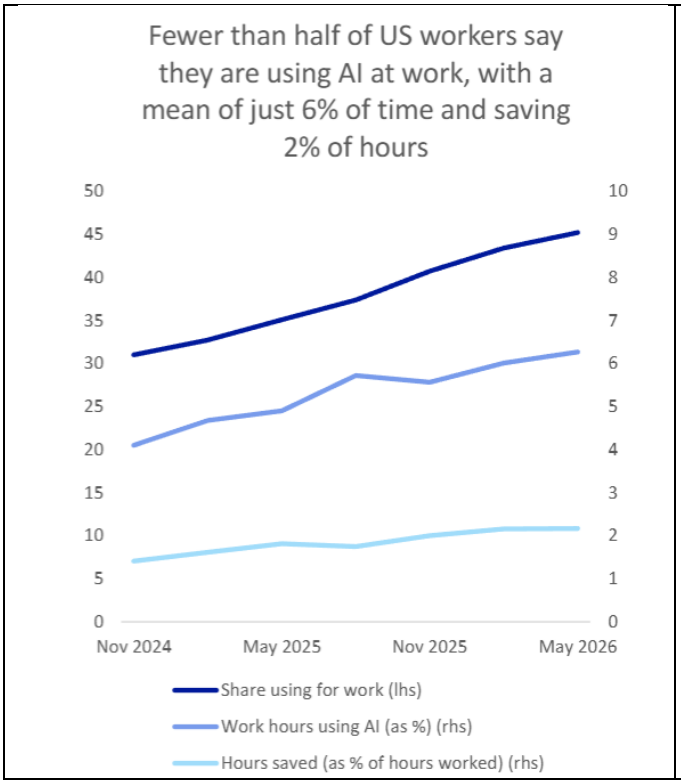


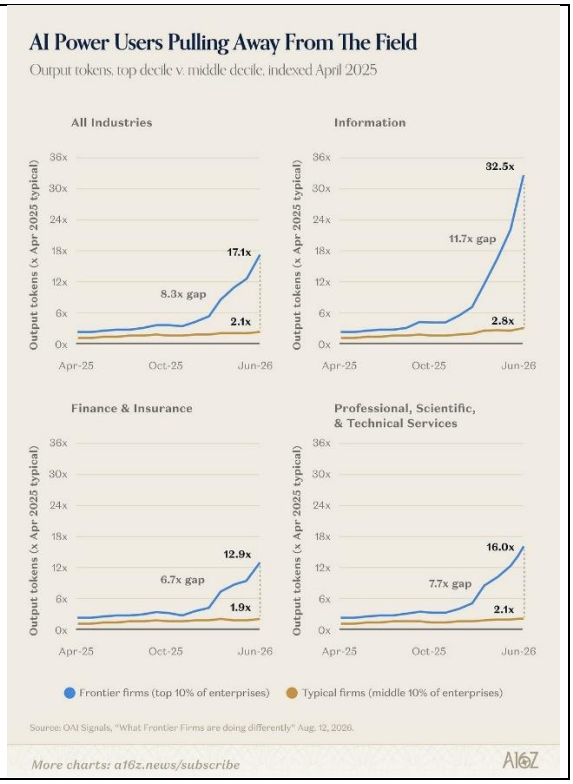
Finally, there are the headwinds from the AI supply chain itself: Taiwan Semiconductor's long ramp-up to expand capacity, the historic memory supply crunch that emerged over the past year, and all sorts of other components and equipment that go into the semiconductor industry. These factors in particular could weaken AI economics in the coming years if demand continues to significantly outstrip supply, driving input prices up more drastically.



And yet, for all of these concerns, it is worth considering how the AI buildout would evolve **without any major constraints.** It would likely much sooner reach a state of oversupply, with a resulting crash in pricing power across the value chain. As we've pointed out before, the Internet/telecom bubble's buildout of networking infrastructure led to a huge excess of supply which became known as "dark fiber," at its worst levels nearing utilization rates of only 1-2%. There is simply no analog today with AI technology, which is just nearing its first full year of evolution into the agentic stage. Per a recent estimate by Goldman Sachs Research, shown above, this evolution may alone drive token usage to expand by 24 times by 2030, to 120 quadrillion tokens per month.

After years of steady exponential growth forecasted from the proliferation of agentic workloads, there still looms the physical AI stage which may be the most data-heavy, computationally-intensive stage of them all, given the unique requirements of training robotics to understand and operate effectively within the real world. And in the meantime, enterprise adoption of AI remains modest, with only ~45% of US workers reporting AI usage at work, on average just 6% of their working hours, per Deutsche Bank Research. At the same time, the top 10% of "power users" in a variety of industries use anywhere from ~7-12x more, measured in token output, as the middle 10%, per A16Z's research – illustrating the high ceiling remaining for adoption scaling at the enterprise level:





As we approach the 4th anniversary of ChatGPT's release to the public, there appear to remain multiple, long runways for further AI adoption, and therefore growth in demand for AI compute. Continued advances in model capabilities, increases in model speed, improvements in model efficiency, and continued expansion of use cases as all of the above come to fruition – aided by multiple converging technological accelerations in the fields of chip design, networking, software, and end-application development – appear poised to further expand the AI ecosystem.

With its high concentration across companies that are integral to various layers of the AI value chain – semiconductors, model developers, AI application builders, cloud computing providers, and other infrastructure providers across networking, power and telecommunications – the Nasdaq-100 remains positioned as a prominent, accessible large cap core equity benchmark for investors to track the AI theme, with at least 70% of its exposure today centered on AI thematic names[1].

Sources: Nasdaq Global Indexes, Factset, Bloomberg, A16Z, Deutsche Bank Research, Goldman Sachs Research, Morgan Stanley Research, JPMorgan Asset Management, Enverus, RAND, CoreWeave, Reuters, Yahoo! Finance, Exponential View, Heatmap Pro, NVIDIA, Epoch AI, Silicon Data, Brookings Institute.

[1] Based on overlap of NDX vs. two Nasdaq AI Thematic Indexes (Nasdaq Global AI & Big Data™ Index, Nasdaq CTA Artificial Intelligence™ Index) , by % index weight, as of June 30, 2026.