TECHNOLOGY

INDUSTRY UPDATE

Institutional Equity Research
January 21, 2026

Price (1/21/26)
Industry: TECHNOLOGY

Alexander Platt
(503) 603-3045
AJPlatt@dadco.com

6-for-26: 6 Technology Trends to Watch for in 2026

After a year where just being in AI was a sufficient investment thesis, we expect 2026 to punish undifferentiated exposure and reward investors who can translate technical complexity into differentiated outcomes. The technology space is only getting more complex, which may leave the median take worse, not better, with misunderstandings creating larger mispricings than we saw in 2025. Below, we highlight six areas where we expect to spend most of our time this year, and where we think the most meaningful progress will be made across AI and other deep tech verticals.

2026 will be a year of constraints in the gigawatt-gigacycle. At 100MW campuses, the story could be reduced to "can you get the chips", however, at gigawatt scale, the constraint stack widens across several bottlenecks including CPUs, co-packaged optics, advanced packaging, high bandwidth memory, NAND flash, behind-the-meter power, and more. The reality is that clearing any one bottleneck will not be enough to deliver compute at scale, and the slowest components will increasingly impact timelines for these campuses, especially as we start to see more GW campuses being planned and constructed.

Models will continue to use more compute, not less. We are simply not seeing any sign of scaling laws slowing down across either pre-training, post-training, or inference. And as we start to see substantially more compute come online, we expect scaling to continue with labs spending the surplus on measurable capability and intelligence as opposed to merely cheaper tokens. If anything, the more plausible path is that effective compute per generation rises and expresses itself as deeper agentic trajectories, more tool-use, more verification, and higher tokens consumed per successful task even as cost per token falls.

Lumpy compute ramps will widen the capability gap in the most economically relevant tasks. We think 2026 is the first year we'll feel the impact of large-scale clusters as they become usable for pre-training and post-training, with access to compute increasingly acting as the binding constraint on frontier progress. Labs with abundant compute should gain ground in domains that require long-horizon, agentic performance. On the flip-side, compute-poor labs will remain competitive on many "good enough" regimes, and we expect them to lean harder on algorithmic efficiency, verification-heavy processes, and tighter domain specialization to narrow the gap.

We expect a new scaling vector to emerge from compute-constrained environments. If compute access is the binding constraint for a subset of labs, the natural response is to search for new approaches that buy more capability per unit compute. We think verification compute is the most promising candidate in the near-term, particularly in STEM domains, with several labs already pushing hard in this direction. Alongside that, we expect more attention to "model gardening" and early continual learning, which could shift pre-training economics towards longer-lived, iteratively maintained model families.

We believe a headline STEM breakthrough is in the cards, alongside the first credible AI research intern. We anticipate AI will produce at least one result on-par with solving a Millennium Prize Problem, while also expecting models to cross into genuine usefulness as research assistants on scoped tasks with verifiable output. If that threshold is crossed, we posit it creates a feedback loop where AI modestly accelerates the rate at which more capable models are developed, with implications that compound beyond a single year.

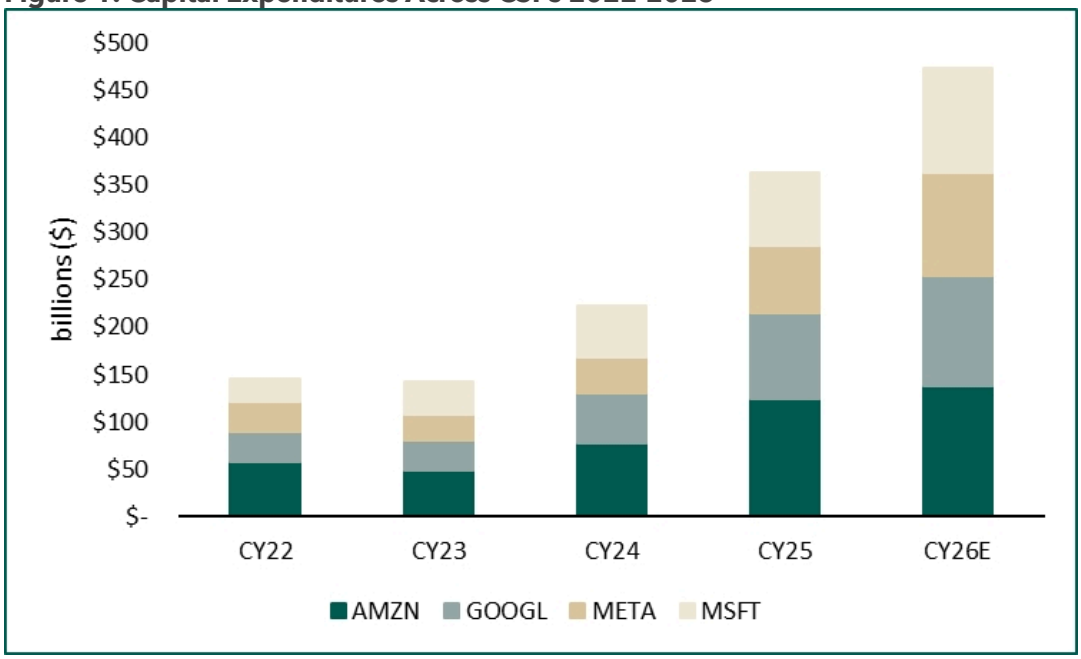
We expect a bifurcation in deep tech adoption this year. Some verticals, such as autonomous vehicles, humanoid robotics, stablecoins, and reusable rockets are approaching meaningful inflection points. While other verticals such as small modular reactors and quantum computing will likely see elongated timelines due to several gating factors that will prevent broad-based adoption this year.

Please refer to pages 39 - 40 of this report for detailed disclosure and certification information.

Introduction to Our 6-for-26: The End to Undifferentiated Exposure

The AI trade is not over, and we don't think that will change in 2026. Capital expenditures alone from hyperscalers and frontier labs remain on an upward trajectory, with aggregate spending commitments for 2026 exceeding $450B across Microsoft, Google, Amazon, and Meta alone. The infrastructure buildout thus will continue, however, we do think that this year, the market's willingness to reward undifferentiated exposure will meaningfully change. For the past three years, the dominant investment framework for AI has been categorical to say the least, with investors primarily identifying the sectors and companies that benefit from the AI buildout, taking their positions, and essentially letting the rising tide lift all boats. And the reality is that this framework has produced exceptional returns the past few years. Since the introduction of ChatGPT in November 2022, through the end of 2025, broad baskets of AI-adjacent equities (whether that be in semiconductors, neoclouds, infrastructure software, networking components, hyperscalers, etc.) outperformed the market by wide margins. For many of these investments, the logic was quite simple, and for a time, correct, which assumed that demand for AI compute would grow so fast relative to supply that anything in the supply chain would benefit.

Figure 1: Capital Expenditures Across CSPs 2022-2026



Source: S&P Capital IQ, D.A. Davidson & Co.

However, we'd argue that this logic is now insufficient. The market has moved from pricing the existence of AI demand to pricing the distribution of AI value capture, which are fundamentally different exercises. The former requires only directional conviction in the form of "will AI adoption grow?" while the latter requires a more precise understanding of competitive dynamics, technical architectures, and supply chain positioning. Furthermore, we'd argue that this dispersion matters more now because the value captured by different participants in the AI supply chain varies by orders of magnitude, with the market only beginning to price this dispersion correctly. We could take a second to just consider the range of outcomes within any given category. "AI semiconductors" encompasses NVIDIA, which earns roughly 73% gross margins on their GPUs that we could argue are still considered supply-constrained, but this categorical label also encompasses companies selling commodity components at 30% gross margins with no pricing power. Both end up being viewed as "AI winners" in the sense that their revenues grow with the adoption of AI, but the former compounds intrinsic value while the latter is merely participating in the cycle. And paying equivalent multiples for these businesses, or even similar premiums to their pre-AI baselines, reflects a failure to distinguish between exposure and value capture. This failure has persisted though because until recently, aggregate demand was so strong that it masked the underlying dispersion. When hyperscaler capex is growing rapidly on a year-over-year basis, and every supplier is capacity-constrained, the difference between a high-margin chokepoint and a low-margin commodity supplier is academic, as both are sold out, both are raising prices, and both appear to be compounding. However, the dispersion we're referring to only becomes visible when supply begins to catch demand in some segments while remaining constrained in others, which is precisely the environment we're entering this year.

The market's treatment of AI exposure as a monolithic factor has created a specific pattern of mispricing. Companies with genuine technical differentiation and durable competitive positions trade at multiples that, while elevated, often understate their earnings power in a world where AI adoption continues. Meanwhile, companies with transient AI exposure (those benefiting from a temporary supply-demand imbalance rather than a structural advantage) trade at multiples that assume the current environment persists indefinitely. The former is often undervalued on a risk-adjusted basis, the latter is more often than not, at least worth avoiding. What makes 2026 different is the maturation of supply in certain segments coinciding with continued constraints in others. We'd argue that GPUs themselves are no longer scarce or supply-constrained. Even some standard server components aren't necessarily considered scarce either. However, we'd argue that optical networking at 800G and above is scarce. Advanced packaging is clearly becoming tighter bottleneck. CPUs are emerging as a constraint in the era of AI agents. NAND flash is scarce and HBM remains constrained. Power generation and delivery for gigawatt-scale clusters remains an issue. Point being is that the companies positioned at these genuine chokepoints will continue to compound, and those positioned at the now-resolved bottlenecks will see their "AI premium" compress toward baseline multiples.

This selective supply resolution creates a stock picker's market in the most literal sense of the phrase. The returns from holding a broad basket of AI-adjacent equities will increasingly approximate market returns as the premium for undifferentiated exposure compresses. Generating alpha now requires identifying which specific bottlenecks remain binding, which companies are positioned at those bottlenecks, and more critically, how long those bottlenecks persist before supply catches up. The dispersion we're expecting is not a modest widening around a common mean but rather we anticipate that the gap between the best-positioned and worst-positioned companies within AI-adjacent sectors will exceed the gap between those sectors and the broader market. Or put differently, being in the right AI stock will matter more than being in AI stocks at all, which we'd argue is the defining ethos of a stock picker's market.

There's a second dimension to our thesis for 2026 that deserves some further explaining as well, which is the role of technical misunderstanding in creating mispricings. Financial markets have always struggled to a degree with technology transitions. The difficulty in this situation isn't stupidity by any means as market participants are largely on average, intelligent and well-resourced, but rather that technology transitions themselves require domain expertise that most haven't acquired given the nascent nature of the paradigm itself, and acquiring said expertise takes time. During these transition periods, we often see that narratives substitute for understanding, with the market trading on simplified mental models since constructing more accurate ones requires understanding that most have yet to acquire at that moment. We're not making a criticism here, but just rather observing a structural feature of how information diffuses through markets. When a new technology emerges, the pool of people who understand it deeply is small and concentrated in technical roles typically that rarely overlap with capital allocation. The pool of people allocating capital to the technology is large but lacks technical depth, which creates a predictable pattern as a result being that early narratives are directionally correct but mechanistically wrong, and the gap between narrative and mechanism creates the mispricings.

The AI revolution and its transition has followed this pattern precisely. Consider the market's initial framework for AI investment, which for a while was along the lines of "AI models require training compute, training compute requires GPUs, therefore GPU demand increases, therefore NVIDIA benefits". Now each step in this chain is correct, and the conclusion is and was correct no doubt, but the framework omits enough detail that it generates false positives and false negatives at the margin. The false positives are companies that appear to benefit from "AI demand" but whose actual exposure is to a transient or commoditized segment of the value chain. The false negatives are companies whose role in AI infrastructure is non-obvious but structurally important. Optical networking is a canonical example of the latter, as the market spent most of 2023-2024 focused on GPUs while underappreciating that interconnect bandwidth would become a binding constraint on cluster scale. The companies positioned then at that constraint (e.g. transceiver manufacturers, optical component suppliers, switch ASIC designers) were available at multiples that did not reflect their importance to the AI buildout.

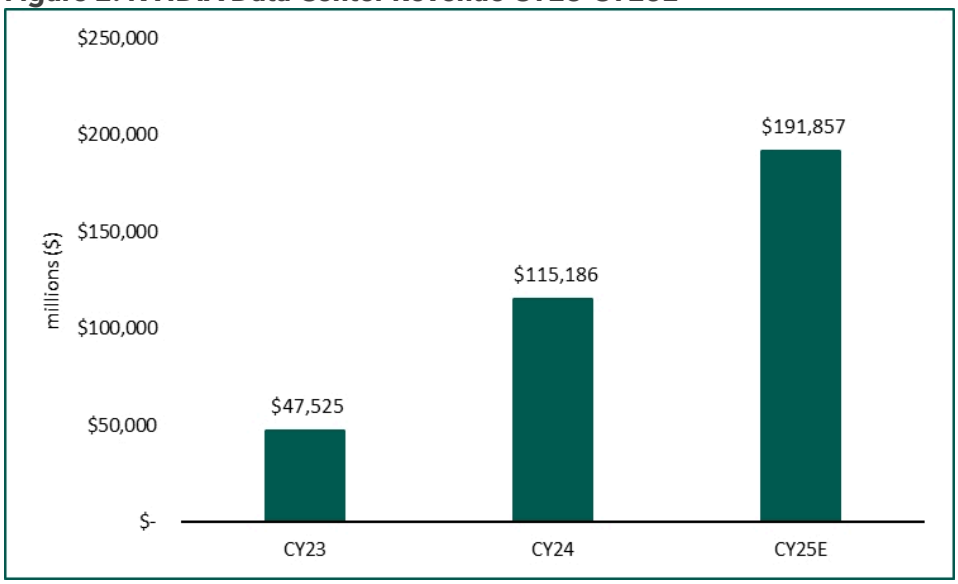
And we'd argue that this mechanism of mispricing is worth examining in detail because it will recur, starting this year. When the market lacks a detailed technical model of a system, it prices assets based on their proximity to the legible parts of the system (i.e. GPUs are legible and everyone knows that AI requires GPUs) but the specific constraints that determine how many GPUs can be deployed, how fast they can communicate, and how efficiently they can be utilized are not legible to most market participants at first. This creates a temporal arbitrage opportunity for investors who develop technical fluency before the rest of the market does. And what we're saying is that 2026 will create an environment that amplifies this dynamic for two simple reasons. First, the technical complexity of AI systems is increasing, not decreasing with architectures being deployed that are substantially more complex than the systems that defined the 2023-2024 market narrative, and understanding where value accrues in these systems requires correspondingly deeper technical knowledge. Second, the pace of technical change is creating narrative whipsaw due to mental models being based on the most recent salient data points, and those data points are arriving faster than our mental models can stabilize.

As we've said, this pattern will repeat throughout 2026. New model releases, architectural innovations, and capability demonstrations will generate headlines that sophisticated participants interpret differently than the market median. The technical details (which architectures are being used, what trade-offs they imply, how they affect the economics of inference and training) will determine the correct responses from the market. Our meta-thesis then, has two components. First, the market will stop paying a premium for undifferentiated AI exposure, creating dispersion between companies with genuine technical moats and those with transient cyclical exposure. And second, technical misunderstanding will remain the primary source of mispricing, and the complexity of the systems being deployed will make this misunderstanding more acute rather than less.

The Gigawatt Giga-Cycle

The investment narrative for AI infrastructure from 2023 through early 2025 was organized around a single constraint: GPUs. Demand for AI compute exceeded the supply of NVIDIA GPUs, and this imbalance determined pricing power, capital allocation, and equity valuations across the sector. The framework was simple and largely correct. NVIDIA's data center revenue grew from $47.5B in CY2023 to our estimate of over $191B in CY2025 because it controlled the binding constraint on AI deployment. On the flip-side of things, adjacent suppliers benefited in proportion to their proximity to that constraint. However, this year will not be the same, and the single-bottleneck model is no longer adequate. The AI infrastructure buildout has reached a scale where constraints propagate through the entire system, binding at various points depending on the specific deployment configuration and timeline. GPUs remain important, and we don't expect them to lose importance anytime soon, but they are no longer the sole determinant of who can deploy compute at scale or how fast. The binding constraint has shifted in any given quarter or year across advanced packaging capacity, HBM memory allocation, optical transceiver availability, power delivery infrastructure, or cooling system lead times. And the reality is that it's often several of these at once.

Figure 2: NVIDIA Data Center Revenue CY23-CY25E



Source: Company reports, D.A. Davidson & Co.

The shift we're witnessing from a single-bottleneck regime to a multi-bottleneck regime has a very specific cause, which is that the infrastructure being deployed has crossed a threshold of system complexity where no single component's supply can be expanded independently. A 100k GPU training cluster is not 100k independent GPUs but rather it's an integrated system where the GPUs must communicate over a fabric that requires specific networking components, draw power from delivery systems rated for specific loads, dissipate heat through cooling infrastructure with specific capacity, and be packaged using processes with specific throughput. Expanding GPU supply without proportionally expanding each of these adjacent capacities does not yield additional deployable compute, it just yields chips that are waiting for the rest of the system to catch up. This type of interdependence creates what we'd categorize as constraint propagation. When any single bottleneck is resolved, the binding constraint just shifts to the next limiting component, and that component's suppliers experience the pricing power and demand surge that previously accrued to the resolved bottleneck. The total system capacity is determined by the minimum of all component capacities, following a Liebig's law dynamic familiar from agricultural and biological systems which states that growth is dictated not by total resources available, but by the scarcest resource. Which means that investment in expanding any single component beyond the capacity of other components yields no incremental system output until those other components are also expanded. This creates significant implications on the duration of the actual AI buildout, as under a single-bottleneck model, the buildout ends when supply of the bottleneck component catches demand. But under a multi-bottleneck model, the buildout continues as long as any component remains supply-constrained because resolving each constraint merely reveals the next one. Making the total duration of the "cycle" not the time to resolve the primary bottleneck but the sum of resolution times for the sequence of bottlenecks, adjusted for whatever parallelism is achievable in capacity expansion.

Current evidence would suggest that there's substantial seriality in this sequence. For example, power infrastructure cannot be parallelized with data center construction because the power systems must be designed for specific load profiles that depend on the compute configuration. Each dependency introduces sequencing constraints that extend the total timeline beyond what a parallel-expansion model would predict. To further expand on this point, the multi-bottleneck regime produces mini-cycles within the larger buildout as different constraints bind and release. When HBM is the binding constraint, HBM suppliers capture incremental margin while GPU suppliers face inventory accumulation. When the constraint shifts to advanced packaging, the margin capture shifts accordingly. These mini-cycles create rotational dynamics within the AI infrastructure sector that are invisible to investors using aggregate "AI demand" as their primary analytical lens. To this point, we're already seeing this rotational pattern emerge in data from last year. The first half of the year saw elevated pricing power for HBM suppliers as memory allocation constrained system shipments. The second half saw that pricing power moderate as HBM capacity expanded, while optical component suppliers experienced tightening as 800G transceiver demand exceeded production capacity. Regardless of the trend, neither shift reflected a change in aggregate AI demand, which continued to grow throughout the year. Both of which reflected the internal dynamics of constraint propagation through a complex system.

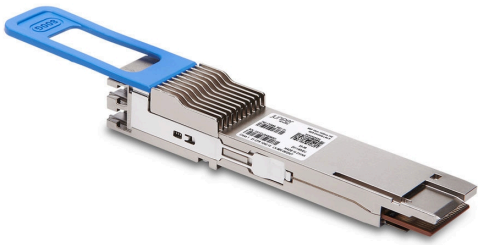
Optical Networking and Co-Packaged Optics

The networking fabric connecting GPUs within and across data centers has emerged as a binding constraint on compute cluster scale. This specific constraint operates at multiple levels being (1) the bandwidth available per link (2) the power consumed per bit transmitted (3) the physical density of connections achievable in a given form factor and (4) the reach over which high-bandwidth connections can be maintained. And each level presents their own distinct technical challenges, and the solutions we'll see being deployed this year represents a fundamental architectural transition from pluggable optical transceivers to co-packaged optics.

The Transition from Pluggables to CPO

The current paradigm for data center optical connectivity relies on pluggable transceivers which are modular units that plug into cages on the front or back panel of switches and servers, converting electrical signals to optical signals for transmission over fiber. A pluggable transceiver contains an optical engine (the components that perform electro-optical conversion), a digital signal processor (DSP) that conditions the electrical signal before conversion, and supporting circuitry for power management and thermal control. The DSP is the critical component for understanding why this architecture faces scaling limits. An electrical signal traveling from a switch ASIC or GPU to a front-panel transceiver cage must traverse 15 to 30 centimeters of copper trace on a printed circuit board. Over this distance, the signal degrades substantially due to insertion loss, crosstalk, and impedance discontinuities. The DSP's function is to recover this degraded signal through equalization, retiming, and error correction before the optical engine converts it to light. This recovery process is computationally intensive and power-hungry.

Figure 3: Breakdown of a Pluggable Transceiver



Source: Juniper Networks

In a typical 800G transceiver, the DSP accounts for approximately 50% of total module power consumption and 20-30% of the bill of materials cost. An 800G DR4 transceiver consumes roughly 16-17W, of which 6-8W is attributable to the DSP. At cluster scale, this power consumption becomes substantial. A 200k GPU cluster on a three-layer InfiniBand network requires tens of thousands of transceivers, consuming on the order of 17MW in transceiver power alone. The DSP portion of this load, approximately 8-9MW, performs no function other than compensating for the signal degradation introduced by the physical distance between the switch ASIC and the transceiver. Co-packaged optics eliminates this inefficiency by placing the optical engine on the same package substrate as the switch ASIC or GPU. The electrical signal path shrinks from tens of centimeters to tens of millimeters, reducing signal degradation to the point where DSP-based recovery is unnecessary. The optical engine can be driven directly by shorter-reach SerDes from the host ASIC, consuming substantially less power per bit.

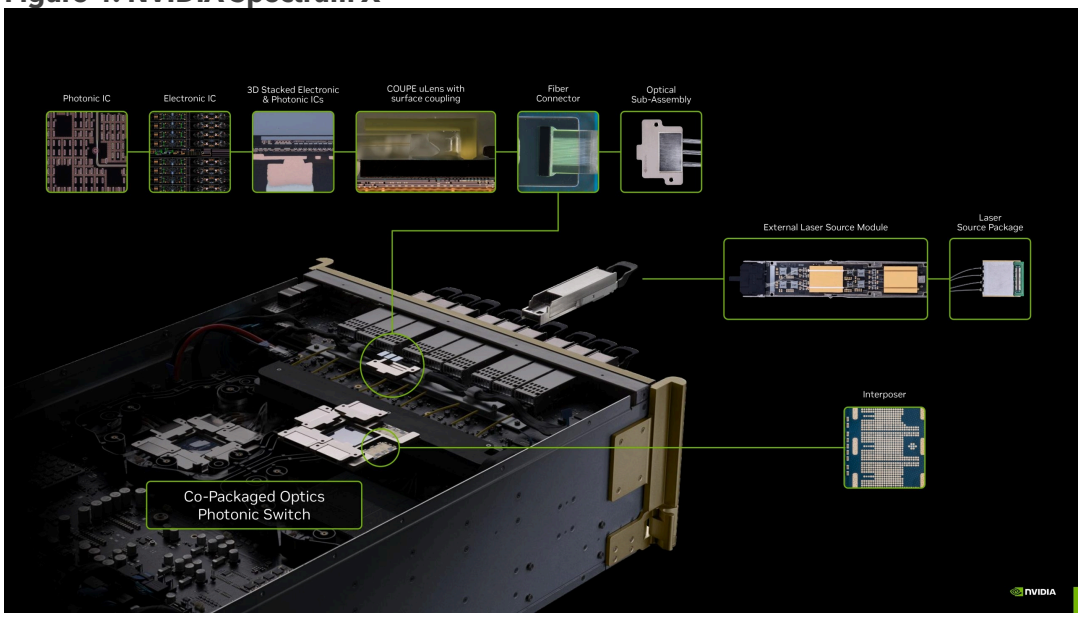
And what we'd point out is that the power savings are significant. Meta's testing of Broadcom's Bailly 51.2T CPO switch, published at ECOC 2025, demonstrated that optical engine plus external laser source power consumption is approximately 5.4W per 800G of bandwidth, compared to approximately 15W for an equivalent 800G pluggable transceiver with DSP. This represents a 65% reduction in power per bit at the optical interface. Additionally, NVIDIA's CPO implementations show similar results, with estimates of 4-5W per 800G for the optical engine and external laser source combined, representing a 70-75% reduction versus DSP-based pluggables. The total cluster-level impact is more modest because networking represents only a fraction of total cluster power. For a GB300 NVL72 cluster on a three-layer network, switching from DSP transceivers to CPO reduces total networking power by approximately 23% but reduces total cluster power by only 2-3%. The value proposition for CPO in scale-out networking (GPU-to-switch connectivity) is therefore meaningful but not transformational.

What we'd argue is that the more compelling application is scale-up networking (though scale-out networking happens first), where CPO enables capabilities that pluggable transceivers cannot achieve at any power budget. Scale-up networks connect GPUs within a coherent domain where they can share memory and coordinate at fine granularity. Current scale-up implementations, such as Nvidia's NVLink, use copper interconnects that provide high bandwidth (7.2 Tbit/s per GPU in NVLink 5.0) but are limited to approximately two meters of reach. This reach constraint limits scale-up domain size to one or two racks, which in turn limits the number of GPUs that can be interconnected in an all-to-all topology. CPO enables optical scale-up links that maintain NVLink-class bandwidth over distances of tens or hundreds of meters. This reach extension allows scale-up domains to span multiple racks or even multiple buildings, dramatically increasing the number of GPUs that can participate in a single coherent training run. The performance implications are substantial as collective communication operations that currently require traversing the slower scale-out network could instead execute over the faster scale-up fabric, reducing synchronization overhead and improving training efficiency.

Why 2026 is a Big Year for CPO

CPO has been discussed as an imminent transition for over a decade. The reason it matters specifically in 2026 is the convergence of three factors (1) products shipping in volume (2) a maturing supply chain centered on TSMC's COUPE platform and (3) accumulating reliability data that addresses customer concerns about field serviceability. NVIDIA announced two CPO-enabled scale-out switches at GTC 2025 with the Quantum X800-Q3450 for InfiniBand and the Spectrum-X 6800 for Ethernet. These are not prototypes or limited-availability products but are intended for volume deployment in production data centers. The Q3450 uses 72 optical engines at 1.6 Tbit/s each, providing 144 ports of 800G connectivity. The Spectrum 6800 offers 512 ports of 800G in its high-radix configuration. Both products integrate optical engines on the switch package substrate, eliminating front-panel transceivers for back-end network connectivity.

Figure 4: NVIDIA Spectrum X



Source: NVIDIA

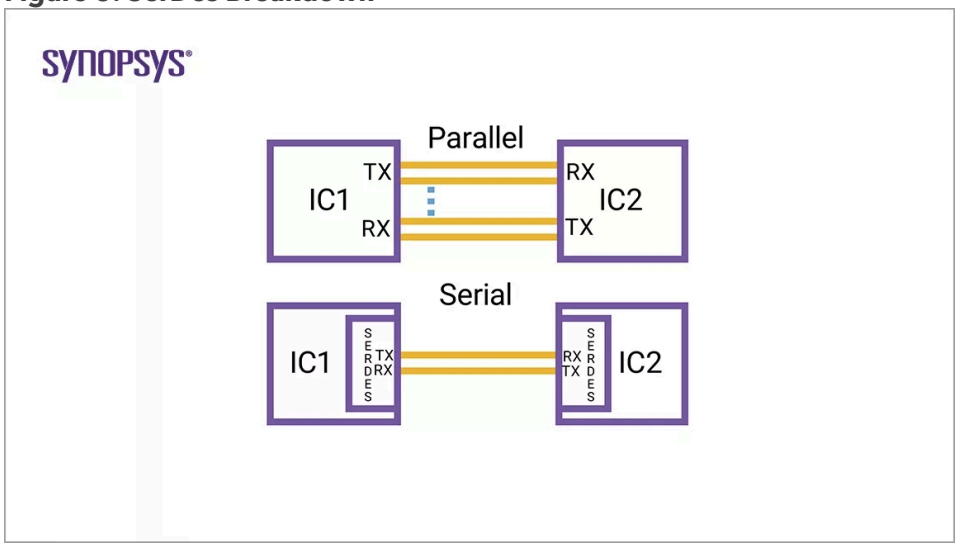
The manufacturing foundation for these products is TSMC's Compact Universal Photonic Engine (COUPE) platform. COUPE provides an integrated solution for CPO fabrication including the electrical integrated circuit (EIC) containing drivers and transimpedance amplifiers is manufactured on TSMC's N7 node, the photonic integrated circuit (PIC) containing modulators and photodetectors is manufactured on TSMC's SOI N65 node, and the two are bonded using TSMC's SoIC process, which provides a bumpless interface with minimal parasitic capacitance. The integration here is actually quite critical for performance, as the parasitic capacitance introduced by bump-based bonding limits achievable bandwidth per lane while SoIC-based bonding enables scaling to 100G per lane and beyond. Furthermore, TSMC's entry into CPO manufacturing is significant because it brings the company's established strengths in advanced logic and packaging to a domain previously served by smaller foundries with more limited capacity. NVIDIA, Broadcom, and Ayar Labs have all adopted COUPE for their CPO roadmaps, consolidating the supply chain around a single integration platform which reduces supply chain risk for customers while increasing TSMC's leverage over CPO pricing and allocation.

Reliability data is accumulating that addresses the primary customer concern about CPO which is the inability to field-replace optical components. In a pluggable architecture, a failed transceiver can be swapped by a technician in minutes, however, in a CPO architecture, a failed optical engine could render the entire switch unusable. Meta's ECOC 2025 data provides some reassurance to these problems as they stated that across 15 million 400G port-device-hours of testing (approximately 15 CPO switches operating for 11 months), the observed mean time between failure for CPO was 2.6 million device-hours, compared to 0.5-1 million device-hours for pluggable 2xFR4 transceivers. This means CPO appears to be more reliable than pluggables, not less, likely because it eliminates the mechanical connectors and contamination-prone interfaces that cause many pluggable failures. This data is helpful but not yet sufficient for broad adoption though. Fifteen switches over eleven months in a lab environment is a small sample relative to production deployments of thousands of switches in variable data center conditions. The 2026 CPO deployments from NVIDIA and Broadcom will serve partly as supply chain pipe-cleaners, generating the field reliability data that larger customers require before committing to CPO at scale. The production deployments will likely begin in scale-out networking, where the blast radius of failures is smaller, before extending to scale-up networking where reliability requirements are more stringent.

SerDes Scaling Limits

The technical pressure toward CPO is intensified by the approaching limits of electrical SerDes scaling. SerDes (serializer/deserializer) circuits convert parallel data within a chip to high-speed serial data for transmission over copper traces or cables. The bandwidth of an off-chip electrical interface is the product of the number of SerDes lanes and the data rate per lane. Increasing bandwidth therefore requires either more lanes (consuming more chip area and package pins) or faster per-lane data rates (requiring more sophisticated and power-hungry circuitry). Per-lane data rates have scaled from 25G in 2015 to 112G in 2022 to 224G in 2024-2025. NVIDIA's Blackwell architecture ships with 224G SerDes enabling NVLink 5.0's 900 GB/s bidirectional bandwidth per GPU. Broadcom has sampled 224G SerDes in its optical DSPs. This generation of SerDes represents the current state of the art for production deployment. The path to 448G SerDes is less clear though the fundamental challenge is that signal attenuation in copper traces increases with frequency. For example, a 224G signal operating at 112 Gbaud (using PAM4 modulation, which encodes two bits per symbol) experiences acceptable attenuation over short distances but requires extensive equalization for longer reaches, while a 448G signal at 224 Gbaud would experience substantially higher attenuation, potentially requiring either higher-order modulation (PAM6 or PAM8, which degrades signal-to-noise ratio) or dramatic shortening of the copper path.

Figure 5: SerDes Breakdown



Source: Synopsys

NVIDIA's approach for Rubin uses bidirectional SerDes to achieve 448G per physical channel. So rather than doubling the symbol rate, bidirectional SerDes transmits and receives simultaneously on the same conductor pair, effectively doubling the data rate per wire without increasing frequency. This approach works for scale-up interconnects within a rack but does not extend to longer-reach applications where signal propagation delays make bidirectional operation impractical. True 448G unidirectional SerDes, which would be required for longer-reach copper interconnects, remains a research challenge, with industry consensus that achieving 448G over meaningful distances will require either a breakthrough in equalization techniques, a transition to PAM8 modulation with associated SNR penalties, or abandonment of electrical transmission in favor of optics. The latter option is precisely what CPO provides. This SerDes scaling limit creates a strategic forcing function as companies that solve the interconnect bandwidth problem will be able to build larger and faster AI clusters while those that do not will face architectural ceilings on cluster scale. NVIDIA's continued investment in copper-based scale-up (NVLink over copper, bidirectional SerDes) reflects confidence that copper can be extended for at least one more generation, while their parallel investment in CPO for scale-out reflects recognition that optics will eventually be required across the full interconnect hierarchy.

The 800G to 1.6T Transceiver Transition

Independent of the CPO transition, the pluggable transceiver market is undergoing its own generational shift from 800G to 1.6T modules. This transition creates supply chain stress that compounds the broader networking bottleneck. An 800G transceiver typically uses four lanes at 200G each (DR4 or FR4 configuration) while a 1.6T transceiver uses eight lanes at 200G (DR8 or FR8) or four lanes at 400G (DR4-400G). The eight-lane approach is simpler technically but requires twice as many optical components per transceiver and the four-lane approach requires doubling the per-lane data rate, which demands more sophisticated modulators, drivers, and DSPs.

Figure 6: Comparison Between 800G Transceiver and 1.6T Transceiver

Source: AscentOptics

This merely continues to make the case that the supply chain is constrained at multiple points. For example, EML (electroabsorption modulated laser) production, which provides the light sources for DR4 and DR8 transceivers, is concentrated in a small number of facilities with limited expansion capacity. DSP silicon for 1.6T transceivers requires advanced process nodes (7nm or below) and competes for wafer capacity with other high-demand products. Testing and qualification throughput is rate-limited by the availability of specialized equipment that must itself be manufactured and deployed. All these constraints translate directly into cluster deployment timelines so a hyperscaler planning a new training cluster must secure not only GPU allocation but also transceiver allocation, and transceiver lead times for 1.6T modules currently extend to multiple quarters. And the reality is that the companies that locked in transceiver supply early in 2025 are likely the ones that will be deploying clusters in 2026, while those that didn't will have to wait.

Another thing that's key to mention is that the transceiver bottleneck also influences the economic case for CPO. If pluggable transceivers were abundant and cheap, the incremental complexity of CPO integration would be harder to justify. In a world where pluggables are scarce and expensive, CPO becomes more attractive even if its technical advantages were marginal. So we'd argue that the current supply situation therefore accelerates CPO adoption beyond what a pure technology comparison would predict.

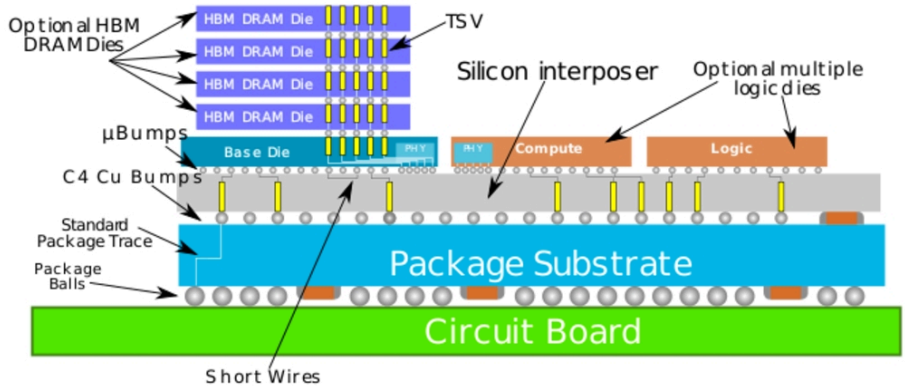
CoWoS and Advanced Packaging

The discourse around AI chip supply has focused predominantly on wafer fabrication capacity, so "how many wafers can TSMC produce at the N4 or N5 node", "how NVIDIA's allocation competes with Apple and AMD", or "does capacity expansion at Arizona or Kumamoto relieve shortages?" We'd argue that for the most part this framing is rather incomplete, as for most advanced AI accelerators, the binding constraint is not wafer fabrication itself (though this is still important) but advanced packaging.

CoWoS as a System-Level Chokepoint

Modern AI accelerators are not monolithic chips but multi-die systems integrated on a shared substrate. NVIDIA's Blackwell GB200 combines two GPU dies with eight HBM3e memory stacks on a single package. AMD's MI300X integrates multiple compute chiplets with HBM stacks on a unified base die. Google's TPU v5p and Amazon's Trainium 2 follow similar multi-die architectures. The technology enabling this integration is advanced packaging, and more specifically, TSMC's CoWoS (Chip on Wafer on Substrate), which is the dominant platform for AI accelerator packaging. CoWoS creates a silicon interposer that serves as a high-bandwidth interconnect layer between dies. The GPU or accelerator dies are placed on the interposer along with HBM memory stacks, with the interposer providing the dense wiring (thousands of interconnects per millimeter) required for high-bandwidth memory access. The interposer-plus-dies assembly is then bonded to an organic package substrate that provides power delivery and external I/O, resulting in a system-in-package that achieves memory bandwidths impossible with conventional packaging.

Figure 7: Breakdown of CoWoS Architecture



Source: GitHub

The constraint arises because CoWoS capacity scales differently than wafer fabrication capacity. Wafer fabs can increase output by adding tools and running additional shifts, and while the capital intensity is high, the expansion physics are well understood. CoWoS capacity on the other hand is constrained by the availability of specialized equipment (bonding tools, testing systems), the yield of the multi-die assembly process, and the physical throughput of operations that cannot be parallelized in the same way as wafer processing. While TSMC's CoWoS capacity has expanded substantially since 2023, expansion has consistently lagged demand regardless. The company has repeatedly doubled CoWoS capacity over the past couple of years, yet customers continue to report allocation constraints. To this point, the gap between wafer output and packaging throughput is visible in inventory dynamics, as GPU dies accumulate waiting for packaging slots while finished package units remain supply-constrained.

The Substrate and Interposer Scaling Challenge

CoWoS packaging faces technical challenges that intensify as AI accelerators grow larger. Two issues are particularly relevant, with the first being interposer size limits and the second being yield degradation from multi-die integration. Silicon interposers are fabricated using semiconductor lithography, which imposes a reticle size limit of approximately 26mm by 33mm (858 square millimeters) per exposure. Interposers larger than this limit must be created by stitching multiple exposures together, a process that introduces alignment challenges and potential defect sites at stitch boundaries. For instance, the Blackwell GB200 requires an interposer substantially larger than a single reticle, necessitating multi-exposure stitching with associated yield implications.

As interposers grow larger, mechanical stress becomes increasingly problematic. The interposer, dies, and substrate have different coefficients of thermal expansion meaning as that package heats and cools during operation and testing, these materials expand and contract at different rates. Larger packages experience greater absolute dimensional change, increasing the risk of warpage, delamination, and interconnect failure. Managing this stress requires careful co-design of materials, die placement, and thermal management, adding engineering complexity that does not scale linearly with package size. The yield mathematics of multi-die packaging compound these challenges. A CoWoS package containing n known-good dies has a yield ceiling equal to the product of individual die yields raised to the nth power, multiplied by the packaging process yield. If individual GPU dies have 95% yield and the packaging process has 90% yield, a two-die package like Blackwell has a theoretical yield ceiling of approximately 81% ($0.95 \times 0.95 \times 0.90$). Adding HBM stacks, each with their own yield, further reduces the ceiling.

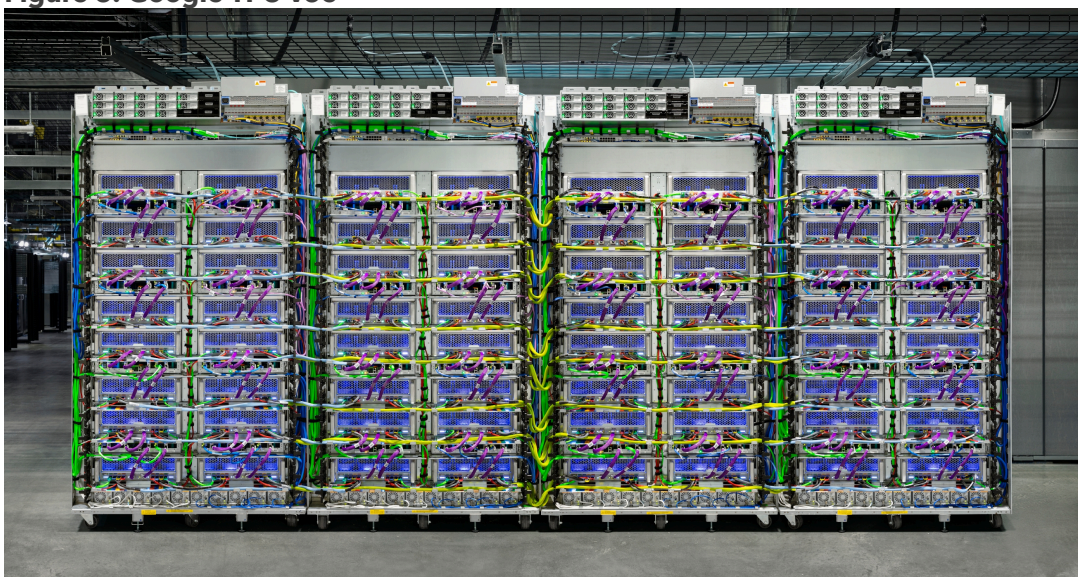
In practice, sophisticated binning and redundancy schemes recover some yield loss, but the fundamental dynamic remains which is that packages containing more dies have lower yield than packages containing fewer dies, all else equal. This yield penalty acts as a tax on the multi-die architectures that enable the highest-performance AI accelerators, making them disproportionately expensive relative to their component costs. The CPO transition discussed in the previous section intensifies these packaging challenges. Co-packaged optics requires placing optical engines on the package substrate alongside the switch ASIC or accelerator die. NVIDIA's Spectrum-X Photonics switch package for example measures 110mm by 110mm to accommodate 36 optical engines surrounding the switch ASIC, compared to Blackwell's 70mm by 76mm package. This larger substrate must maintain signal integrity for high-speed electrical connections to each optical engine while managing the thermal load of both the switch ASIC and the optical components. The engineering complexity is substantially higher than conventional packaging, and the yield implications of bonding 36 additional known-good optical engines onto each package are significant.

Implications for Custom Silicon Timelines

The advanced packaging constraint has direct implications for the timeline of custom AI silicon from hyperscalers. Including but not limited to Google's TPU, Amazon's Trainium, Microsoft's Maia, and Meta's MTIA all require advanced packaging for their highest-performance configurations. These chips compete with NVIDIA and AMD for the same TSMC CoWoS capacity. The allocation dynamics favor high-volume, high-margin customers. NVIDIA, as TSMC's largest CoWoS customer by revenue, receives priority allocation. Hyperscalers developing custom silicon face a structural disadvantage: their volumes typically are lower (many of them barring Google with the TPU are developing their chips for internal uses only), their design iterations are less frequent (reducing learning curve benefits), and their packaging requirements are often more demanding (custom configurations optimized for specific workloads rather than general-purpose designs).

Because of this, custom silicon programs have a higher risk of slipping their announced timelines. A hyperscaler announcing a new AI accelerator for deployment in 2026 is making an implicit assumption about CoWoS allocation that may not hold if NVIDIA's demand increases or if TSMC's capacity expansion encounters delays. The history of custom silicon announcements includes numerous examples of products that were technically ready but supply-constrained due to packaging limitations. This dynamic creates an information asymmetry that sophisticated investors can exploit. Announced timelines for custom silicon should be interpreted as aspirational targets conditional on packaging availability, not firm commitments. The companies with the strongest packaging partnerships (long-term agreements, prepaid capacity, co-investment in expansion) will meet their timelines; those without such partnerships will experience delays that may not be disclosed until they affect reported metrics.

Figure 8: Google TPU v5e



Source: Google

The packaging constraint also influences architectural decisions. Some hyperscalers have opted for designs that use less advanced packaging (standard flip-chip rather than CoWoS) to avoid the capacity bottleneck, accepting lower memory bandwidth in exchange for supply certainty. Others have invested in alternative packaging approaches (Intel's EMIB, proprietary solutions) to reduce dependence on TSMC. These strategic responses are visible in product specifications and partnership announcements for those who know where to look. For 2026, we expect the packaging constraint to remain binding for the highest-performance AI accelerators. TSMC's capacity expansion will absorb much of the incremental demand from Blackwell ramp and next-generation products from AMD and hyperscalers, leaving limited slack for unexpected demand increases. The companies with secured allocation will execute on their roadmaps; those without will find their ambitions constrained by the physical throughput of bonding tools in Taichung.

High Bandwidth Memory & NAND Flash

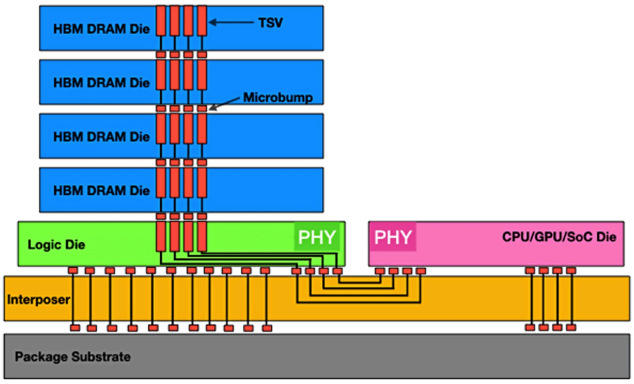
The memory system of an AI accelerator determines what workloads it can execute efficiently. Training and inference place different demands on memory capacity and bandwidth, and as models have grown larger and inference patterns have shifted toward longer contexts and reasoning-intensive workloads, the memory hierarchy has emerged as a binding constraint on AI deployment. This constraint operates at two levels (1) the HBM attached directly to accelerators and (2) the storage systems used for model weights, checkpoints, and intermediate state.

HBM as an Inference Bottleneck

High Bandwidth Memory (HBM) provides the memory capacity and bandwidth for modern AI accelerators. An NVIDIA H100 SXM for instance contains 80GB of HBM3 with approximately 3.35 TB/s of memory bandwidth while a GB200 increases this to 192GB of HBM3e with approximately 8 TB/s of bandwidth per GPU. Simply put, these specifications determine the size of models that can be served from a single accelerator and the throughput achievable for a given batch size. And the relationship between memory bandwidth and inference throughput is direct for large language models, as each token generated requires reading the model weights from memory, performing matrix multiplications, and writing intermediate activations. For a model with *p* parameters at *b* bytes per parameter, generating a single token requires approximately *p*b* bytes from memory at minimum. The achievable tokens per second is therefore bounded above by memory bandwidth divided by model size, before accounting for compute utilization and other overheads.

For a 70B parameter model stored in FP16 (2 bytes per parameter), the minimum memory read per token is 140GB. On an H100 with 3.35 TB/s bandwidth, this implies a theoretical maximum of approximately 24 tokens per second per GPU for single-batch inference, assuming perfect memory bandwidth utilization. Practical throughput is lower due to memory access patterns, activation storage, and KV cache overhead. Larger models face proportionally tighter constraints like if we were to take a 405B parameter model like Llama 3.1 405B which requires distributing inference across multiple GPUs not because any single GPU lacks sufficient compute but because no single GPU has sufficient memory capacity or bandwidth. The transition from Hopper to Blackwell relaxes this constraint meaningfully as the combination of 2.4x higher memory capacity (192GB versus 80GB) and 2.4x higher memory bandwidth (8 TB/s versus 3.35 TB/s) enables serving larger models on fewer GPUs and achieving higher throughput for memory-bound workloads. For inference providers, this translates directly to cost per token because fewer GPUs per model instance means lower capital and operating costs per unit of inference output.

Figure 9: High Bandwidth Memory Architecture



Source: Semiconductor Engineering

Another point we feel important to make is that this constraint will reassert itself at the frontier, as models designed for Blackwell-class memory systems will be larger than models designed on previous generations of chips, essentially absorbing the capacity and bandwidth gains. The memory bandwidth requirement scales with model size, and model size scales with available memory, creating a co-evolution dynamic where hardware improvements enable larger models rather than making existing models cheaper to serve. The memory constraint does not disappear; it migrates to a new equilibrium.

Yield and Capacity Dynamics

HBM production is primarily concentrated among three suppliers: SK Hynix, Micron, and Samsung. SK Hynix has maintained a consistent lead in both capacity and yield, supplying the majority of HBM for NVIDIA's data center GPUs. Samsung has faced yield challenges that have limited its qualification for high-volume NVIDIA orders, while Micron entered the HBM3e market later but has achieved qualification for select NVIDIA products and supplies HBM for other accelerators including AMD's MI300 series.

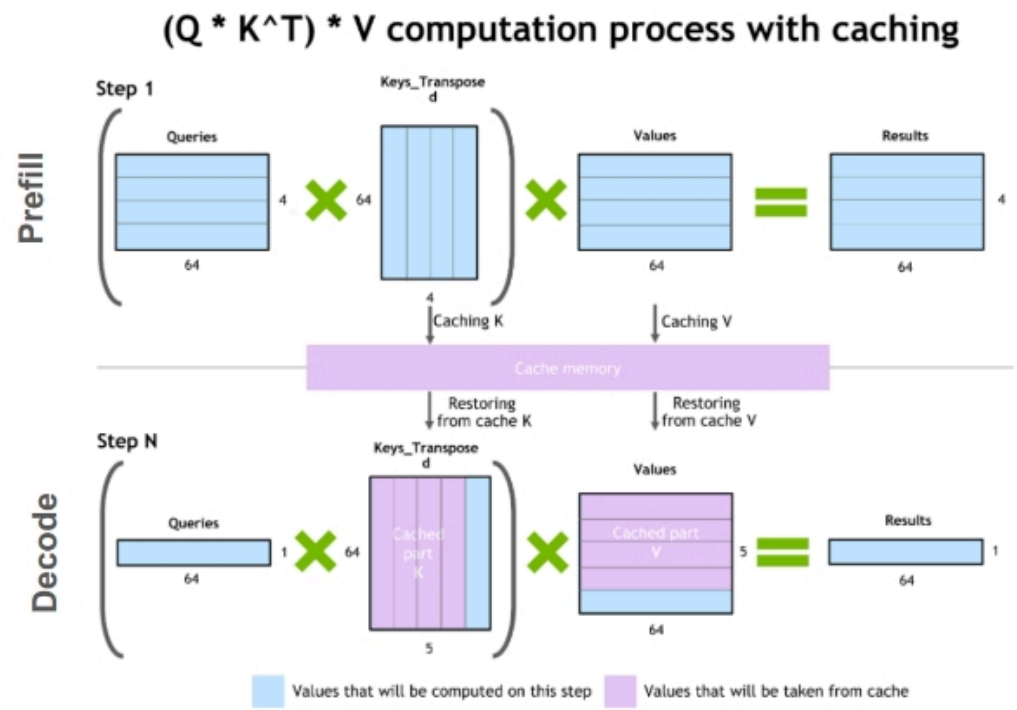
The yield dynamics of HBM are particularly challenging because HBM stacks multiple DRAM dies vertically, connected by through-silicon vias (TSVs). An HBM3e stack typically contains 8 or 12 DRAM dies, while the stack yield is the product of individual die yields multiplied by the stacking and bonding yield. If individual DRAM dies have 95% yield and the stacking process has 90% yield, an 8-high stack has a theoretical yield of approximately 66% (0.95^8 × 0.90). Higher stacks, required for higher capacity per package, face steeper yield penalties. The yield differential between suppliers translates into capacity and allocation constraints. SK Hynix's yield advantage means it can produce more functional HBM stacks per wafer start, making it the preferred supplier for volume applications. Samsung's yield challenges have resulted in limited allocation from NVIDIA despite Samsung's substantial DRAM manufacturing capacity. While Micron's position is intermediate, its yields have improved sufficiently for qualification but not for displacing SK Hynix as the primary supplier.

For AI system availability, HBM allocation can be the binding constraint independent of GPU die supply. A GB200 requires eight HBM3e stacks, so if HBM supply is constrained while GPU dies are available, the result is unshippable systems despite adequate GPU production. This dynamic has been visible in multiple quarters where Nvidia's reported data center revenue was limited by memory availability rather than GPU production. The geographic concentration of HBM production also introduces additional risk given all three major HBM suppliers have their primary production facilities in East Asia, with SK Hynix and Samsung concentrated in South Korea and Micron's HBM production primarily in Taiwan and Japan. This concentration creates supply chain vulnerability to regional disruption, a factor that hyperscalers increasingly weigh in their infrastructure planning.

KV Cache as a First-Class Problem

Beyond model weights, inference workloads must store key-value (KV) cache which is the accumulated context from previous tokens in a sequence that enables the model to attend to earlier parts of the conversation. KV cache size scales linearly with sequence length and is multiplicative across attention layers. For a transformer with l layers, h attention heads, d dimensions per head, and sequence length s, the KV cache requires storage of $2 \times l \times h \times d \times s$ values, with the factor of 2 accounting for both keys and values. For a model like Llama 3.1 70B with 80 layers, 64 heads, and 128 dimensions per head serving a 128K context window in FP16, the KV cache per sequence is approximately 2 × 80 × 64 × 128 × 128,000 × 2 bytes, or roughly 167GB. This exceeds the HBM capacity of a single H100 for the KV cache alone, before accounting for model weights or activations. Serving long-context workloads therefore requires either distributing the KV cache across multiple GPUs, compressing the cache through quantization or pruning, or offloading portions of the cache to slower storage tiers.

Figure 10: Explanation of KV Caching in LLMs

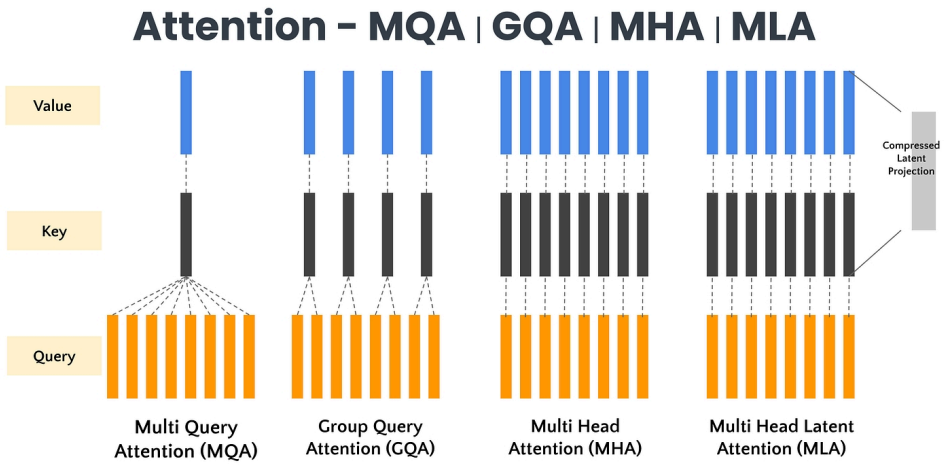


Source: Medium

The growth of context windows from 4K tokens in early GPT-3 deployments to 128K or 1M tokens in current frontier models has transformed KV cache from an incidental overhead to a primary memory consumer. Inference providers report that for long-context workloads, KV cache memory consumption often exceeds model weight storage, inverting the traditional assumption that weights dominate memory usage. This inversion changes the optimization target as memory efficiency techniques must address cache management, not just weight compression.

Several architectural responses have emerged. Multi-query attention (MQA) and grouped-query attention (GQA) reduce KV cache size by sharing key-value projections across multiple query heads. DeepSeek's multi-head latent attention (MLA) compresses the KV cache through learned projections. Sliding window attention limits the context each token attends to, bounding cache growth at the cost of reduced long-range dependency modeling. Each approach trades some capability for memory efficiency, and the optimal trade-off depends on the workload distribution. For inference infrastructure, KV cache dynamics affect provisioning and pricing. A server optimized for short-context, high-throughput workloads (many concurrent users with brief interactions) has different memory requirements than one optimized for long-context, low-throughput workloads (few concurrent users with extended reasoning). Inference providers increasingly differentiate pricing based on context length, reflecting the real resource cost difference rather than treating all tokens as equivalent.

Figure 11: Multi-Query Attention (MQA) vs. Grouped-Query Attention (GQA)



Source: Medium

Storage Entering the Narrative

The constraints discussed above concern HBM, the fastest tier of the memory hierarchy. But AI workloads increasingly stress lower tiers as well, pulling NAND flash storage into the infrastructure bottleneck narrative. KV cache offloading represents one driver of storage demand. When KV cache exceeds HBM capacity, portions can be offloaded to NVMe SSDs and retrieved when needed. The latency penalty is substantial (microseconds for NVMe versus nanoseconds for HBM) but acceptable for workloads where the alternative is failing to serve the request at all. Offloading requires high-bandwidth, low-latency storage with sufficient write endurance to handle the continuous churn of cache eviction and retrieval.

Video generation represents a second driver. Generating video requires producing sequences of frames, each represented as a spatial grid of tokens that must be stored and processed in temporal order. A single minute of generated video at 24 frames per second with 1080p resolution can require storing and manipulating hundreds of gigabytes of intermediate activations. Training video generation models requires even larger storage for checkpoints, gradient accumulations, and dataset shards. The storage requirements scale with video length, resolution, and frame rate, creating demand growth that substantially exceeds the growth in text and image workloads. Retrieval-augmented generation (RAG) represents a third driver. RAG systems supplement model generation with retrieved passages from external knowledge bases, which must be stored, indexed, and accessed with low latency. Enterprise RAG deployments commonly involve knowledge bases of hundreds of gigabytes to terabytes, stored on SSDs for fast retrieval. The growth of RAG as a deployment pattern creates storage demand proportional to the knowledge base sizes organizations choose to index.

The flash endurance question becomes relevant under these access patterns. SSDs have finite write endurance, typically specified in drive writes per day (DWPD) over a warranty period. KV cache offloading, checkpoint storage, and video generation all involve sustained high-write workloads that stress endurance limits. Enterprise-grade SSDs with high DWPD ratings command significant price premiums over consumer-grade drives, and inference providers must factor endurance-limited drive replacement into their operating cost models. For 2026, we expect storage to transition from a commodity input to a considered constraint for specific workloads. The overall NAND market is not supply-constrained in the way that HBM is; flash memory is a mature commodity with multiple suppliers and adequate capacity. But the specific categories of storage optimized for AI workloads (high-endurance NVMe, low-latency enterprise SSDs, high-capacity drives for checkpoint storage) may face tighter supply as demand from AI deployments grows faster than these segments' traditional markets anticipated.

Energy and Power Generation

The constraints we've discussed thus far operate within the data center with packaging, memory, and networking determining what systems can be built and how they perform. However, energy operates upstream of all of these, as a data center cannot deploy compute it cannot power, and the availability of power at the scale required for gigawatt or even just multi-hundred megawatt clusters has become a binding constraint on where and how fast clusters can come online.

The Arithmetic of Gigawatt Scale

The power requirements for frontier AI clusters have grown by roughly an order of magnitude since 2022. A DGX A100 system (eight A100 GPUs) consumes approximately 6.5kW. A DGX H100 system (eight H100 GPUs) consumes approximately 10.2kW. A GB200 NVL72 rack (72 Blackwell GPUs in a liquid-cooled, high-density configuration) consumes approximately 120kW. The per-GPU power has increased, and the packaging density that enables high-bandwidth interconnects has concentrated that power into smaller physical footprints. A 100k GPU training cluster built on GB200 NVL72 racks requires approximately 1,400 racks, consuming roughly 168MW of IT load before accounting for cooling and facility overhead. Applying a power usage effectiveness (PUE) ratio of 1.2 (achievable with liquid cooling) yields total facility power of approximately 200MW. A cluster targeting 500k Blackwell-class GPUs requires on the order of one gigawatt of facility power. These figures assume current-generation hardware. The trend in accelerator power consumption continues upward. NVIDIA's has already announced that Rubin will have higher per-GPU power draw to support increased compute density. So each hardware generation increases the power required per unit of deployed compute, even as efficiency (FLOPS per watt) improves. The efficiency gains reduce the power required per unit of useful work, but the total work demanded grows faster than efficiency improves, resulting in net power growth.

Cooling is inseparable from power at these densities. A 120kW rack dissipates 120kW of heat in a footprint of approximately 5 square meters. Air cooling cannot remove heat at this density; the thermal gradient between chip surface and ambient air is insufficient to drive adequate convective heat transfer regardless of airflow volume. Liquid cooling is required, either through direct-to-chip cold plates (as in the GB200 NVL72) or through immersion in dielectric fluid. Liquid cooling reduces PUE by eliminating the energy spent moving large volumes of air, but it introduces its own infrastructure requirements: coolant distribution systems, heat exchangers, and often cooling towers or dry coolers sized for the full thermal load. These systems require space, capital, and lead time to deploy. A data center designed for air-cooled servers cannot be trivially retrofitted for liquid cooling at GB200 densities; the mechanical and plumbing infrastructure must be purpose-built.

The gap between announced cluster capacity and actual power availability reflects these compounding requirements. A hyperscaler can announce a 500MW AI cluster, but delivering that power to racks requires: securing a power purchase agreement or utility interconnection, building or upgrading substation capacity, installing switchgear and power distribution within the facility, deploying cooling infrastructure sized for the thermal load, and commissioning the integrated system. Each step has its own lead time, and the total timeline is the critical path through these dependencies.

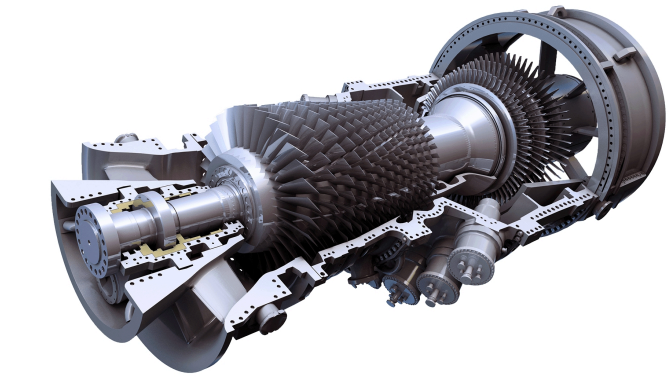
Grid Interconnection vs. Behind-the-Meter

The traditional model for data center power involves grid interconnection: the facility connects to the local utility's transmission or distribution system and draws power as needed. This model faces two challenges at AI cluster scale: interconnection queue delays and transmission capacity constraints. Interconnection queues in major U.S. markets have extended to multi-year timelines. A large load seeking to connect to the PJM Interconnection (serving the mid-Atlantic and Midwest) may wait three to five years between application and energization. The queue backlog reflects both the volume of applications (driven heavily by renewable energy projects and data centers) and the studies required to assess grid impact. Each large interconnection requires analysis of how the new load affects power flows, voltage stability, and fault currents throughout the surrounding network. These studies cannot be parallelized indefinitely, and the engineering resources to conduct them are finite.

Transmission capacity constraints compound the queue delays. Even after completing the interconnection process, a data center may face curtailment if the local transmission network cannot deliver the contracted power during peak periods. Building new transmission lines requires rights-of-way, environmental review, and construction timelines measured in years to decades. The grid infrastructure in most regions was not designed for concentrated gigawatt-scale loads appearing in locations chosen for land cost and fiber connectivity rather than proximity to generation. These constraints have driven interest in behind-the-meter power: generation assets located at the data center site, connected directly to the facility load without passing through the utility grid. Behind-the-meter generation avoids interconnection queues (the data center connects to the grid at a smaller capacity for backup and supplemental power) and avoids transmission constraints (power is generated where it is consumed).

Gas turbines represent the most mature behind-the-meter option for AI datacenters. A utility-scale gas turbine installation can deliver hundreds of megawatts with a construction timeline of 18 to 24 months, substantially faster than grid interconnection in congested markets. The economics depend on natural gas prices and carbon emission costs, but in many jurisdictions, gas generation remains cost-competitive with grid power while offering faster deployment and greater supply certainty. Several announced AI clusters are being designed around behind-the-meter gas generation. The power architecture involves on-site turbines providing base load, grid interconnection providing backup and supplemental capacity, and potentially battery storage for load smoothing and peak shaving. This hybrid approach allows clusters to energize on the turbine timeline rather than the grid interconnection timeline, accelerating deployment by years in some cases.

Figure 12: GE Vernova 9F Gas Turbines



Source: GE Vernova

Fuel cells represent an emerging alternative, particularly for facilities seeking lower carbon intensity than gas turbines while maintaining behind-the-meter deployment speed. Solid oxide fuel cells and molten carbonate fuel cells can achieve electrical efficiencies above 60%, higher than simple-cycle gas turbines, and can operate on natural gas or hydrogen. The installed base and manufacturing capacity for utility-scale fuel cells remains smaller than for gas turbines, limiting near-term deployment at gigawatt scale, but several hyperscalers have announced fuel cell installations for AI data centers. The nuclear option frequently appears in discussions of AI data center power but faces timeline constraints that exclude it from 2026 deployment relevance. Small modular reactors (SMRs) are not yet licensed for commercial deployment in the United States; the NRC licensing process requires years of review before construction can begin. Even after licensing, construction and commissioning timelines for nuclear facilities are measured in years to decades. Announcements of nuclear-powered AI data centers should be understood as statements about the 2030s, not the 2020s.

Figure 13: Bloom Energy Fuel Cells

Source: Bloom Energy

Implications for Cluster Geography and Timelines

Power availability has become a primary driver of site selection for frontier AI clusters, often superseding traditional factors like fiber connectivity, labor markets, and tax incentives. The question is not where a hyperscaler would like to build but where sufficient power can be delivered on the required timeline. This dynamic explains the geographic clustering of announced AI infrastructure projects. Texas, particularly the ERCOT grid region, offers faster interconnection timelines than PJM or California ISO due to its different regulatory structure and available transmission capacity. The announced Stargate project in Abilene, Texas reflects this calculus: the site offers access to power that would be unavailable on the same timeline in more traditional data center markets.

Similarly, locations with existing heavy industrial infrastructure offer advantages. Sites with retired or underutilized power plants may have transmission interconnections already in place, dramatically reducing the timeline to energize a new load. Sites adjacent to large generation facilities (hydroelectric dams, nuclear plants, industrial cogeneration) may have access to power that is not available to the broader grid. The Fairwater facility in Atlanta represents this pattern: leveraging existing infrastructure to accelerate deployment. International locations are increasingly competitive. Regions with surplus generation capacity, streamlined permitting, or state-owned utilities willing to prioritize AI infrastructure can offer timelines unavailable in the United States. The Middle East, Scandinavia, and parts of Asia have attracted AI infrastructure investment partly on the basis of power availability.

For investors, power constraints create both risk and opportunity. Announced cluster deployments should be evaluated against realistic power availability timelines, not aspirational commissioning dates. A cluster announcement without a credible power strategy (identified site, secured generation or interconnection, plausible construction timeline) is an intention, not a plan. The companies that have locked in power agreements and begun infrastructure construction will deploy on schedule; those still searching for sites or waiting in interconnection queues will experience delays that may not be disclosed until they affect financial guidance. The power constraint also creates long-duration competitive advantages. A hyperscaler that secures a gigawatt of power capacity with a long-term agreement has an asset that cannot be quickly replicated by competitors. Power agreements and site infrastructure represent committed capital that raises barriers to entry, unlike software or model weights that can be copied or approximated. The AI infrastructure leaders of the late 2020s will be partly determined by who secured power in 2024 and 2025.

Models Will Use More Compute, Not Less

Our last theme established that compute supply is expanding across multiple dimensions this year, with new compute clusters coming online that will be available to frontier labs and hyperscalers despite the bottlenecks that constrain the pace of deployment. The question we're aiming to address in this theme is just how will that compute be used, now and in the future as we start to see the first gigawatt clusters come online. Our answer is simple, which is that frontier labs will consume more compute in 2026 than they did in 2025, across every scaling vector. Models will not only use more compute in pre-training, but they will use more compute at test-time, and they will also use more compute in post-training as well. And we believe that the aggregate effect here is that models deployed at the frontier in mid to late 2026 will be substantially more capable than those that were deployed in early 2025 for example, with this capability gain being purchased with compute.

For us to make this claim, it requires us to address a persistent counter-narrative that has persisted for a couple of years now, which is this idea that scaling has reached diminishing returns and that future progress will come from algorithmic efficiency rather than increased compute. This narrative gained a lot of traction in 2024 and early 2025, particularly after DeepSeek demonstrated that frontier-competitive models could be trained at costs dramatically lower than previous estimates. The interpretation from supporters here (which even included us at times) was that the era of scaling was ending and the era of efficiency was beginning. However, this interpretation was incorrect, or more precisely, it conflated two distinct phenomena. Both efficiency gains and scale gains are complements, not substitutes. So when algorithmic improvements reduce the compute required to achieve a given level of intelligence, the equilibrium response is not to achieve that capability at lower cost but to pursue higher capability at similar or greater cost. And over the past year or so, this has become the revealed preference of every frontier lab, which is that when efficiency improves, they train larger models, not cheaper ones. Even for someone like DeepSeek, their efficiency gains were immediately absorbed into training DeepSeek-V3 at a scale that would have been economically prohibitive at prior efficiency levels, with this same dynamic applying to OpenAI, Anthropic, DeepMind, xAI, and every other lab operating at the frontier.

The reality is, that none of the core scaling vectors show any evidence of hitting a wall. Pre-training continues to show that it's yielding capability gains across model sizes, with the frontier defined by the largest models that available compute can train. Test-time scaling continues to yield gains across compute budgets, with no observed ceiling on the returns to additional "thinking time". Furthermore, post-training scaling continues to yield gains in domains where verification is more readily available, with ongoing research extending verification to broader task categories with techniques such as rubrics. So the question we should be asking isn't whether scaling works, because it does, but how much compute is available to scale with. This is where we believe the connection to our first theme becomes more explicit, as the bottlenecks described in our first theme constrain how fast compute supply can grow, but they do not constrain the direction. Every additional unit of deployable compute will be absorbed into one or more of these scaling vectors. Frontier labs have demonstrated their willingness to spend, with the core constraint being availability as opposed to demand. As the clusters described in our first theme start to come online (some already have including Amazon and Anthropic's Rainier), they will be used to train larger models, run more inference compute per query, and invest more in post-training refinement which we'd posit will result in capability growth that is commensurate with compute growth.

In our view, this means that in 2026, models available at year-end will be materially more capable than those available at year-start. We'll see intelligence and capability gains concentrated at the frontier, where the largest compute investments are being made, but will propagate through the model ecosystem as frontier techniques are adapted to smaller scales and as open-weight models incorporate advances pioneered at the frontier. This means that we should plan for an intelligence trajectory that continues on and upwards.

The Edge Narrative as a Sentiment Hedge

A recurring feature of the AI investment discourse is the periodic resurgence of the "edge AI" narrative, which is this claim that AI compute will shift from centralized data centers to distributed devices, reducing the need for infrastructure buildout and transferring value from data center suppliers to device manufacturers and edge chip designers. This narrative has appeared multiple times since 2023, and its timing follows a consistent pattern that reveals its function as a sentiment hedge rather than a technical prediction.

When Edge Appears

The short answer is that the edge narrative gains traction when sentiment on the data center buildout is at its lowest. We can merely observe this from last year, where in late January 2025, following the media's attention over DeepSeek-R1 and DeepSeek-V3 and the subsequent market's reaction, commentary shifted rapidly toward edge deployment. The argument ran as follows. If models can be trained efficiently at small scale, then inference can be run efficiently on devices, reducing dependence on centralized compute. This is really where we started to see the edge narrative takeoff, however, the narrative receded as the market recognized that DeepSeek's efficiency gains did not imply reduced aggregate compute demand, while also, the actual DeepSeek-V3 model itself was so large that there was no viable way for it to fit on any local device. We could even look to just a few months ago, where we've witnessed a slight resurgence of the narrative, as concerns over the AI data center buildout began to grow. Again, the argument was that smaller, more efficient models would enable device-local inference, obviating the need for continued data center expansion.

We'd argue that the correlation between low data center sentiment and high edge enthusiasm isn't coincidental at all, but that edge AI functions as a hedge position for investors who want AI exposure but are skeptical of the infrastructure buildout thesis. When the buildout thesis appears weakest, edge becomes most attractive as an alternative. When the buildout thesis strengthens, edge loses its appeal as a differentiated bet. This dynamic explains the narrative's periodicity better than any underlying technical shift. Additionally, the edge narrative also serves a rhetorical function for those that wish to appear contrarian without abandoning AI exposure entirely, as calling for edge displacement of data centers allows one to be bearish on compute providers while remaining bullish on AI.

Why Frontier Capabilities Stay Centralized

The technical case for centralized compute rests on three factors (1) memory constraints (2) economic efficiency (3) and capability requirements. Each favors data centers over edge devices for frontier AI workloads, and the gaps are widening rather than narrowing. We can start with memory constraints which are the most fundamental. A frontier model with hundreds of billions of parameters requires hundreds of gigabytes of memory to store its weights, plus additional memory for KV cache during inference. The largest smartphone memory configurations typically offer 16-24GB of DRAM. Even with aggressive quantization (reducing weights from 16-bit to 4-bit representation or lower), a 70-billion parameter model requires approximately 35GB of memory for weights alone, exceeding device capacity before accounting for the operating system, applications, or inference overhead. This means that running frontier models on edge devices is not a matter of optimization, but rather that it's physically impossible given current memory architectures. The models that can run on edge devices are necessarily smaller and less capable. A 7-billion parameter model quantized to 4 bits fits in 3.5GB, leaving room for KV cache and system overhead on high-end smartphones. These models are still very useful, but more in specific tasks such as text completion, simple question answering, and basic code assistance. They are not competitive with frontier models on complex reasoning, long-context understanding, or agentic task execution. And the reality is that this gap is not going to close, simply because frontier models are scaling and iterating far faster than available memory in local devices is.

Additionally, economic efficiency favors centralization for most inference workloads, as the cost per token for inference is a function of hardware cost, utilization, and throughput. Data center GPUs achieve higher utilization than edge devices because they serve many users concurrently, amortizing fixed costs across more queries, and they achieve higher throughput because they have more memory bandwidth, more compute, and more optimized software stacks. The result is that cost per token at data center scale is typically an order of magnitude lower than cost per token on edge devices for equivalent model quality, and the edge cost advantage only materializes when the alternative is zero. If a user would not have made a cloud API call because of latency, privacy, or connectivity constraints, then edge inference at any cost is preferable to no inference. But for users who have the option of cloud inference, the economic comparison favors centralization in nearly all cases. The edge narrative implicitly assumes that edge-specific constraints are widespread enough to shift the aggregate compute distribution, but the evidence suggests these constraints apply to a small minority of inference volume at the moment.

Finally, capability requirements for the most valuable AI applications favor centralization decisively. Agentic workflows that execute multi-step tasks, maintain state across long interactions, and invoke external tools require models with strong reasoning capabilities and large context windows. Long-context understanding for document analysis, code review, and research assistance requires models that can attend to hundreds of thousands of tokens. These capabilities are compute-intensive by construction, with test-time scaling meaning that harder problems require more compute, and the compute required often exceeds what edge devices can reasonably provide. The reality is that a user who has experienced an AI assistant that can reason through a complex problem step by step does not revert to a simpler assistant that cannot. The revealed preference data from API usage shows that users consume more tokens per task over time, not fewer, as they learn to use models more effectively, and this demand growth will flow to centralized compute because that is where the most capable models will run.

Where Edge Matters

One point we want to make sure to note is that our analysis does not imply that edge AI is irrelevant, but that it merely implies that edge AI serves a specific use-case with specific constraints and those use-cases do not constitute the majority of AI compute demand or value creation at least in the near-term. Automotive applications represent the strongest case for edge inference as autonomous driving systems require real-time perception and decision-making with latency budgets measured in milliseconds. For instance, cellular connectivity is insufficiently reliable for safety-critical functions, as a vehicle cannot wait for a cloud response while approaching an obstacle. This means that the compute must be local. And this creates genuine demand for edge AI hardware in vehicles, a market that is growing as autonomous capabilities expand. Industrial IoT applications have similar characteristics. A manufacturing robot that must respond to sensor inputs in real time cannot tolerate network round-trip latency. Meaning a quality inspection system processing images at production line speeds requires local inference. Cost-sensitive high-volume inference represents a third category. Applications like voice assistants, smart cameras, and IoT devices that perform simple classification or detection tasks can achieve acceptable quality with small models running on low-cost edge hardware. The economics favor edge deployment when the alternative is paying per-query API costs for millions of devices.

The common thread across these use cases is that they involve constraints that make cloud inference infeasible or uneconomical whether that be latency requirements below network round-trip times, reliability requirements exceeding network availability, or cost requirements below API pricing floors. These constraints are in fact real, and the markets they define are real. But they do not describe the majority of AI compute demand, which involves workloads where cloud inference is feasible, economical, and provides access to capabilities that edge devices cannot match. Meaning that for this year, we expect edge AI to continue growing in its natural markets while remaining peripheral to the frontier capability story. The companies building edge inference chips will find demand from automotive, industrial, and IoT applications, however, the companies building data center infrastructure will find demand from the much larger market of users who want frontier capabilities and are willing to pay for centralized compute to access them. Regardless, the edge narrative will resurface the next time data center sentiment weakens, and it will recede again when the evidence reconfirms that frontier AI runs in data centers.

The Capability Implications of Scaling Compute

Our previous writing has established that frontier models are growing across all scaling vectors, and what we aim to do below is just briefly address what these increases might mean for model capabilities in the new year.

Larger Pre-Training Bases

The simplest assessment of pre-training compute is that it determines the breadth and depth of knowledge encoded in a model's weights. A model trained on more data with more parameters learns more facts, more relationships between facts, and more nuanced representations of concepts. This means improved performance on tasks requiring broad knowledge, accurate recall, and generalization to novel combinations of known concepts. The scaling laws for pre-training are also very well known at this point. Loss decreases as a power law in both model size and training data, with the optimal allocation between parameters and tokens following predictable ratios. The Chinchilla scaling laws suggested that models should be trained on approximately 20 tokens per parameter for compute-optimal training. However, post-Chinchilla practice has shifted toward over-training (more tokens per parameter than the compute-optimal ratio) because inference cost depends on parameter count while training cost depends on total compute. Meaning a smaller model trained longer is cheaper to serve than a larger model trained to the compute-optimal point.

For 2026, the models emerging from new clusters will have been trained on token counts in the tens of trillions, with effective training compute reaching into the 10^26 FLOP range for the largest runs. We think the implications on intelligence and capabilities of these models are rather straightforward, which is that these models will know more, make fewer factual errors, and generalize more reliably than their predecessors. In our view, the improvements will be most visible on knowledge-intensive tasks such as question answering, research synthesis, and technical explanation. That being said, the subtler implication is improved base capability for downstream adaptation. A model with a stronger pre-training foundation responds better to fine-tuning, produces better results with in-context learning, and provides a higher ceiling for post-training refinement. And since the returns to post-training and not independent to pre-training quality, a larger pre-trained base ends up multiplying the effectiveness of subsequent training stages.

More Test-Time Compute

Inference-time compute scaling allows models to allocate variable effort to problems based on their difficulty through the mechanism of chain-of-thought (CoT) reasoning that extends to an arbitrary depth meaning the model generates intermediate reasoning steps, each step conditioning the next, until it arrives at a final answer. Harder problems receive more steps, consuming more compute and producing more reliable answers. The scaling behavior of inference-time compute differs from pre-training scaling, as pre-training improvements are baked into weights and apply uniformly to all queries while inference-time improvements are allocated dynamically and apply selectively to queries that benefit from extended reasoning. This selectivity makes inference-time scaling particularly effective for tasks with variable difficulty such as mathematical problem solving, code debugging, multi-step planning, and complex analysis.

Originally, we first saw this in OpenAI's o1 model family that demonstrated that test-time scaling yields log-linear returns across the tested range, as doubling inference compute produces consistent accuracy improvements on reasoning benchmarks. In our view, the models being trained this year will extend this paradigm with larger reasoning budgets, more sophisticated search procedures, and better-calibrated allocation of compute to problem difficulty. This means that models that can solve harder problems when given time to think, with the definition of "harder" expanding as compute budgets increase. Tasks that previously required human expertise because models could not reason through them reliably become tractable when models can allocate sufficient inference compute, which has been visible in various benchmarks like AIME and SWE-bench where test-time compute scaling has driven accuracy from levels that were barely useful to levels that approach human expert performance.

More Post-Training Compute

Post-training encompasses the techniques applied after initial pre-training to align model behavior with desired outcomes which include but are not limited to supervised fine-tuning on curated examples, reinforcement learning from human feedback (RLHF), reinforcement learning from AI feedback (RLAIF), and reinforcement learning with verifiable rewards (RLVR). The compute invested in post-training has grown substantially, from a minor refinement step to a major training phase in its own right. And most notably RLVR has emerged as particularly important because it enables scaling in domains where correctness can be automatically verified. In mathematics, code execution, and formal reasoning, a model's output can be checked programmatically, allowing training on millions of examples without human annotation. DeepSeek's R1 model demonstrated that extensive RLVR can produce reasoning capabilities competitive with models trained with far more human supervision, while also providing the implication that post-training compute scales in proportion to the availability of verifiable domains, and the set of verifiable domains is expanding through improved verification tools and environment design.

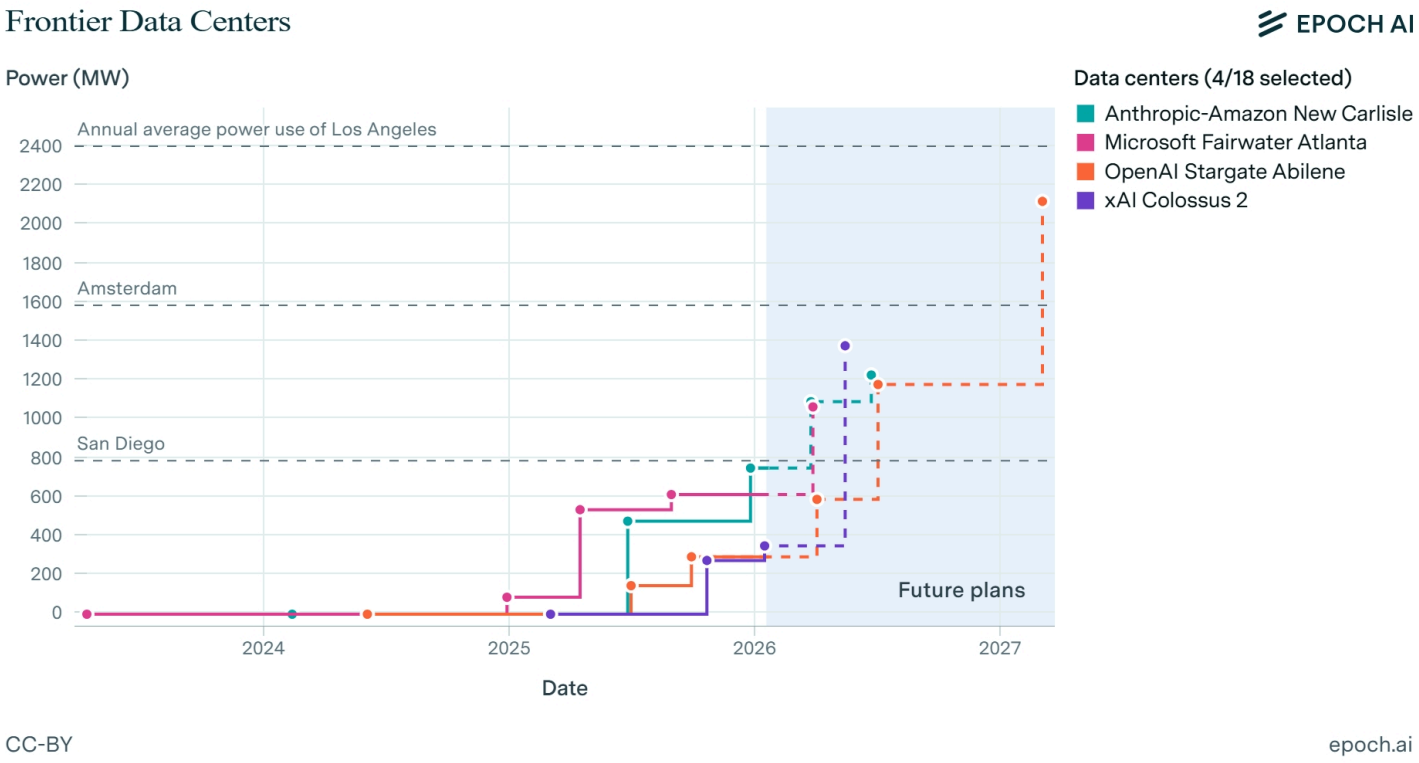
On the capability side of things, increased post-training means improved reliability, better instruction following, and more consistent behavior across task variations. A model with extensive post-training produces fewer errors on tasks it has been trained to perform, follows complex instructions more precisely, and exhibits fewer failure modes under adversarial or unusual inputs. And these improvements end up mattering disproportionately for deployment, where reliability often matters more than peak capability (albeit there's discussion and ongoing debate as to the benefits of reinforcement learning and its impact on economically valuable work). Additionally, the interaction between post-training and the other scaling dimensions is multiplicative as a larger pre-trained model provides more capability to align and more test-time compute provides more reasoning to refine.

Lumpy Compute Ramps Widen the Capability Gap

The typical distinction between closed-source and open-source is often framed as one in terms of philosophy, business model, or safety approach. That being said, we'd argue that these framings often obscure the more fundamental distinction between the two, particularly in AI, which is access to compute. More often than not, closed-source frontier labs have access to compute resources that open-source labs cannot match, and this gap is widening as the largest compute clusters are coming online. And as we've stated in our previous theme, the compute differential can translate directly into capability differentials, and those capability differentials concentrate in the dimensions most relevant to economic value creation. DeepSeek, which is widely regarded as the most capable open-source AI lab, has been explicit about this constraint, as in some of their most recent published papers, DeepSeek researchers have identified compute access as the binding limitation on their ability to match closed-source frontier models. Their algorithmic innovations which are their efforts to close that gap through efficiency rather than scale have worked, and while they are genuine technical achievements, they operate within a compute envelope that is fundamentally smaller than what closed-source competitors can access.

The reality is that the compute gap has structural causes, with closed-source frontier labs being funded by organizations with the capital to build or lease infrastructure at scale. As we know, OpenAI operates with the backing of Microsoft, Amazon, and Google, including access to vast infrastructure across Microsoft and Amazon, and dedicated training clusters. Anthropic has secured billions in funding from Google and Amazon, with corresponding infrastructure access including direct access to place TPUs in their own data centers as opposed to renting them through GCP. Google DeepMind operates within Alphabet's infrastructure footprint, with access to TPU capacity that no other frontier lab has access to at that kind of scale. Meta is maybe the only lab that one could argue is still doing open-source, but it has become clear with Llama 4 and reports with Llama 5 that they do not intend to continue their open-source crusade. On the flip side, open-source labs outside these arrangements face capital constraints that limit their compute access. Training a frontier model can sometimes require hundreds of millions to billions of dollars in compute costs, and the organizations capable of dining such expenditures are precisely the hyperscalers and well-capitalized startups that constitute the closed-source frontier. Independent open-source efforts operate at compute budgets one to two orders of magnitude smaller, which translates directly into capability limitations given established scaling laws.

Figure 14: Frontier Data Centers



Source: Epoch AI

Our thesis for this theme, is that in 2026, the compute clusters that are coming online and being deployed will widen the gap substantially. Facilities such as Anthropic's Project Rainier, Microsoft's Fairwater in Atlanta, OpenAI's Stargate Abilene, and xAI's Colossus 2 are just some representations of compute capacity that has no open-source equivalent. Each of these clusters provides training capability in the range of 10^26 to 10^27 FLOPs for a single training run and open-source labs operating on cloud compute or smaller dedicated clusters have access to maybe 10^24 FLOPs which represents a gap of two to three orders of magnitude that is growing rather than shrinking.

Why Compute Ramps Create Compounding Capability Gaps

The relationship between compute access and model capability is not merely additive as compute advantages compound over time through mechanisms that extend beyond the direct effect of training larger models. Understanding these compounding dynamics explains why the capability gap between compute-rich and compute-poor labs widens faster than the raw compute gap would suggest.

Training Windows as Competitive Moats

A training run for a frontier model is not instantaneous. The largest runs currently take three to six months of continuous cluster operation, and this duration will extend as model scale increases, which means that the time required to complete a training run creates a temporal structure to competition that favors labs with earlier access to large-scale compute clusters. Consider two labs, one with access to a 100k GPU cluster starting in January 2026 and another with access to an equivalent cluster starting in July 2026. Both labs have the same peak compute capacity, but the first lab can complete a six-month training run by July, evaluate results, incorporate learnings, and begin a second run. By the time the second lab completes its first run in January 2027, the first lab is potentially finishing its second run and beginning a third, thus even the six-month head start ends up translating into a full generation of model iteration advantage.

This temporal compounding is amplified by the learning that occurs between training runs. Like any kind of learning, each training run generates insights about architecture, hyperparameters, data composition, and training dynamics that inform subsequent runs. A lab that completes more runs should theoretically accumulate more learning, which will improve the efficiency of their future training runs. And we want to make sure we're emphasizing how important iteration is, because take DeepSeek as an example. Their efficiency gains did not emerge from theoretical analysis alone and academic papers, but from empirical experimentation across numerous training runs that each built on observations they gained from the previous ones. The iteration velocity advantage also extends to post-training as well. RL requires substantial compute and elapsed time, and a lab that finishes pre-training earlier can being post-training earlier, ship models earlier, collect user feedback earlier, and incorporate said feedback into subsequent model generations. So we see this compounding occur across the entire model development cycle not just in the pre-training phase.

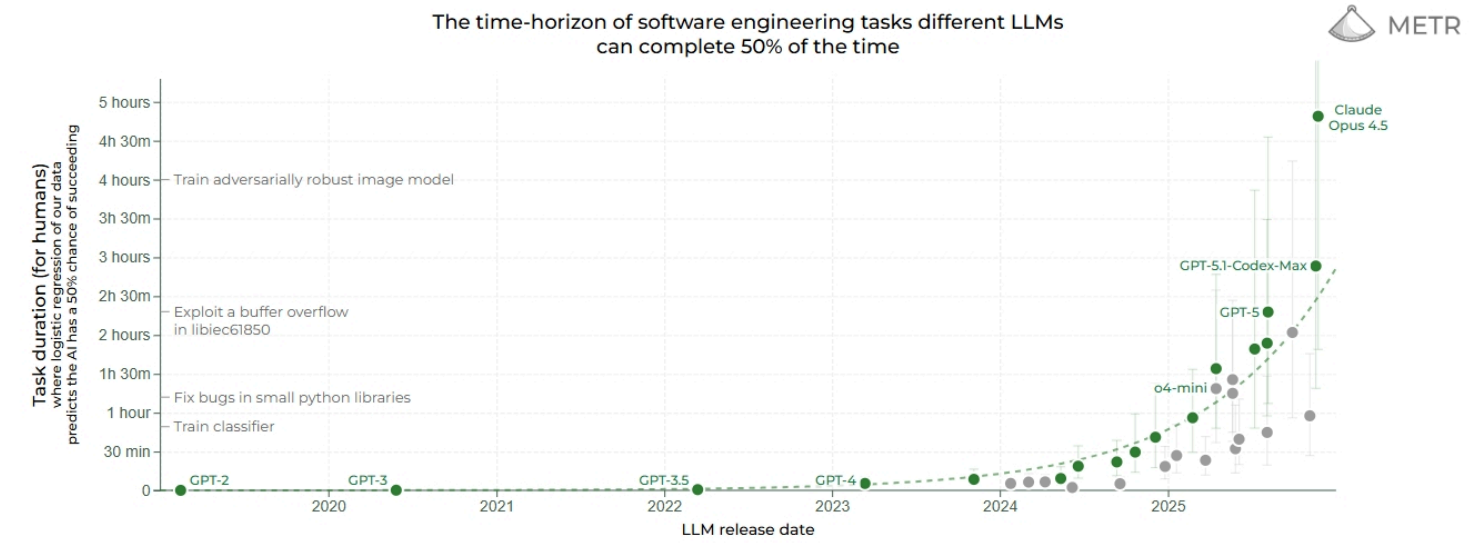
The main takeaway here then is that compute access timing can sometimes matter as much as compute access magnitude. As we outlined, a lab that secures cluster access six months before a competitor gains a significant advantage that can persist and potentially widen even if the competitor eventually matches raw capacity. Thus, it's our view that the compute clusters coming online throughout this year will enable training runs that complete before competitors' clusters are operational, and those early completions translate into shipped models, accumulated learning, and iteration cycles that later entrants cannot easily recover.

Dimensions Where Capability Gaps Will Widen

The compute advantage we're talking about does not manifest uniformly across all capabilities, as certain capability dimensions exhibit steeper scaling with compute than others, and these dimensions correlate with economic value in ways that concentrate returns at the frontier. For instance, agentic task execution requires models to decompose complex goals into subtasks, execute actions in external environments, observe outcomes, and adjust plans accordingly. Meaning the capability to perform agentic work reliably depends heavily on several factors that scale with compute.

First of all, the base model must have sufficient world knowledge and reasoning ability to formulate sensible plans. And as we've noted in our previous thoughts, capability scales with pre-training compute following established scaling laws. Secondly, the model must be trained to use tools and APIs reliably, which requires extensive post-training on tool use tasks. This post-training is compute-intensive because it involves running the model in simulated or real environments, observing outcomes, and updating based on success or failure. Finally, the model must allocate appropriate inference-time compute to planning and verification, which requires training to use inference compute effectively. So quite literally every one of these requirements favors labs with large compute budgets. Long-horizon reasoning presents similar dynamics. Tasks requiring coherent reasoning across many steps, such as mathematical proofs, complex code architectures, or multi-stage analysis, benefit from inference-time compute scaling. The models must be trained to use that inference compute effectively, which requires post-training with compute budgets that scale with the complexity of target tasks. A model trained with limited post-training compute exhibits degraded performance on long-horizon tasks even if its base capabilities are strong, because it has not learned to allocate inference compute appropriately. Let's just take a look at METR's long-horizon task evaluation benchmark, which some would argue is one of the most economically relevant benchmarks especially when looking at software development. A trend that you'd notice is that the best performing models in this benchmark, that are well above trendline in terms of length of task they can complete at 50 or 80% accuracy, are always closed source models, with the models that are below trendline almost always open-source models that have access to less compute in all three stages of scaling.

Figure 15: METR Long-Horizon Evaluation of SWE Tasks



Source: METR

The training data for long-horizon reasoning is also inherently scarce. Human-generated examples of extended reasoning chains are rare compared to examples of short question-answer pairs. Synthetic data generation can address this scarcity, but generating high-quality synthetic reasoning traces requires running capable models for extended inference, which consumes compute. Labs with more compute can generate more synthetic training data, creating a feedback loop where compute advantage enables data advantage which enables further capability advantage. Multimodal capabilities exhibit similar scaling dynamics with additional complexity. Training models that can reason across text, images, audio, and video requires datasets that are larger and more expensive to curate than text-only datasets. The compute required to train on multimodal data exceeds text-only training by factors that depend on the resolution, duration, and diversity of non-text modalities. Video understanding is particularly compute-intensive because video combines spatial resolution (pixels per frame), temporal resolution (frames per second), and duration into token counts that grow multiplicatively.

The economic significance of these capability dimensions is not incidental as enterprise AI adoption is concentrated in applications that require agentic execution (automated workflows, code generation, data analysis), long-horizon reasoning (research, strategy, complex problem solving), and multimodal understanding (document processing, visual inspection, content generation). These are the applications where willingness to pay is highest because they substitute for expensive human labor on complex tasks, and the capability gaps are largest precisely where the economic value is greatest. This correlation ends up creating a winner-take-most dynamic in frontier AI markets because the labs that can deliver agentic, long-horizon, and multimodal capabilities capture the high-value use cases. Labs that cannot deliver these capabilities are relegated to lower-value applications where capability requirements are less stringent and price competition is more intense.

The Chinese Open-Source Dynamic

The compute gap between closed-source frontier labs and open-source alternatives has a specific geographic dimension that warrants separate analysis. Chinese AI labs, such as DeepSeek, have produced open-weight models that approach frontier performance despite operating under hardware constraints that limit their compute access. Understanding how this is possible, and what limits it implies, requires examining the specific mechanisms of Chinese AI development under export control restrictions.

What DeepSeek Demonstrated

DeepSeek's technical reports for V3 and R1 documented a set of architectural and algorithmic innovations that achieved frontier-competitive performance at reported training costs dramatically below Western frontier labs. The headline figure of $5.6M in training compute costs for DeepSeek-V3 captured market attention, but the technical substance lies in the specific innovations that enabled this efficiency. Multi-head latent attention (MLA) reduces the memory footprint of the key-value cache by projecting keys and values into a lower-dimensional latent space before storing them. The compression ratio is substantial, reducing KV cache memory requirements by approximately 90% relative to standard multi-head attention while maintaining model quality. This innovation directly addresses the memory bandwidth bottleneck discussed in our first theme, allowing larger batch sizes and longer contexts on fixed hardware. Efficient mixture-of-experts routing reduces the computational overhead of expert selection in MoE architectures. DeepSeek's approach minimizes the communication costs that typically limit MoE scaling by optimizing how tokens are assigned to experts and how expert outputs are combined. The result is that total parameter counts can scale to trillions while active parameters per forward pass remain in the tens of billions, achieving the capacity benefits of scale without proportional compute costs.

FP8 mixed-precision training reduces memory and compute requirements by using 8-bit floating point representations where precision loss is tolerable, while maintaining higher precision for numerically sensitive operations. The technique requires careful analysis of which operations can tolerate reduced precision, but when properly implemented, it approximately halves memory requirements and increases throughput relative to FP16 training. Reinforcement learning with verifiable rewards, which DeepSeek scaled extensively in R1, demonstrated that post-training compute can substitute for some pre-training compute in domains where correctness is automatically verifiable. By training models to reason through mathematics and code problems with outcome-based rewards, DeepSeek achieved reasoning capabilities competitive with models trained at substantially higher pre-training cost.

The aggregate effect of these innovations is a shift in the compute-capability frontier. DeepSeek demonstrated that the same capabilities achievable with X compute using 2023-vintage techniques could be achieved with X/5 or X/10 compute using optimized techniques. This efficiency gain does not eliminate the importance of compute, but it changes the exchange rate between compute and capability. Critically, DeepSeek has been explicit that compute remains their binding constraint. In technical reports and public statements, DeepSeek researchers have indicated that they would train larger models if they had access to more compute. Their efficiency innovations represent adaptation to constraint, not a claim that compute no longer matters. The innovations allow them to extract more capability per FLOP, but they do not change the fundamental relationship between compute and capability at the frontier.

The Lag Structure

The compute and efficiency dynamics produce a characteristic lag structure between closed-source frontier capabilities and Chinese open-source availability. When a closed-source lab introduces a new capability, Chinese labs can eventually match it, but with a delay determined by the time required to develop efficiency improvements that compensate for the compute gap. The lag is shortest for capabilities that are well-defined and "benchmarkable". For example, when OpenAI released o1 with test-time scaling for reasoning, the core technique was identifiable from the model's behavior even without technical documentation (despite the fact that it was well-documented even in press releases how this was done, relatively speaking), which allowed Chinese labs to observe what the model did, hypothesize how it worked architecturally, and develop their own implementations. The result was that DeepSeek-R1 demonstrated competitive reasoning performance within months of o1's release, suggesting a lag of roughly one to two quarters for capabilities where the target was very clear from the jump.

That being said, the lag is typically far longer for capabilities that are less observable or require extensive infrastructure development. For instance, agentic capabilities that depend on tool use training, environment simulation, and real-world deployment feedback require investment beyond the core training run. A lab that lacks the compute to train frontier base models also typically lacks the compute to run extensive agentic training, creating compounding delays, with the lag being longest for capabilities that depend on cumulative organizational learning rather than discrete technical innovations. This applies to labs that aren't even Chinese or open-source per se. Take a look at Meta Superintelligence Labs, which has taken considerable time since its formation despite the talent and compute resources they have available to even have a shot of developing a model that is within reach of the frontier. Our point being, is that it takes a lot of effort to maintain your position at the frontier, especially if you're playing catch-up relatively speaking with American frontier labs. And the reality is that Chinese labs complete fewer training runs per year than compute-rich frontier labs, accumulating less empirical knowledge about training dynamics, failure modes, and optimization opportunities which just results in a widening knowledge gap with each iteration cycle, creating a structural disadvantage that efficiency innovations cannot fully address.

For commoditized capabilities where open-source models are already competitive, the lag is essentially zero. Chinese open-source models match or exceed Western open-source alternatives on standard benchmarks like MMLU, HumanEval, and GSM8K. The market for "good enough" inference on routine tasks is competitive, with Chinese models often offering favorable price-performance ratios due to lower operating costs. Our 2026 outlook is for this bifurcated pattern to continue. On well-defined benchmarks and commoditized tasks, Chinese open-source models will remain competitive, but on frontier capabilities requiring large compute budgets, particularly agentic execution and long-horizon reasoning, the gap will widen as Western labs bring gigawatt-scale clusters online. The efficiency innovations that allowed DeepSeek to approach the frontier in 2024-2025 will continue, but they are unlikely to fully offset the order-of-magnitude compute expansion occurring at closed-source frontier labs.

Token Growth as a Downstream Effect

While we've been focusing on training compute and the capability advantages it confers in some of our previous sections, we want to point out that training compute is only an input to model development, and isn't necessarily a direct measure of economic activity (though building out data centers is certainly contributing its fair share to economic growth in the United States). The actual economic activity in AI occurs at inference and where deployed models are actually processing tokens in response to user queries. So essentially, token volume is the throughput metric of the AI economy, and its growth trajectory has numerous implications for infrastructure demand, competitive dynamics, and value capture that extends well beyond training considerations.

The Capability-Adoption Feedback Loop

Our first order of business is to mention that token growth is not independent of model capability. More capable models enable more use cases, which generates more users, which produces more queries, which consumes more tokens. This relationship creates a feedback loop where capability improvements drive demand expansion rather than merely capturing existing demand more efficiently. This mechanism operates through use-case unlocking, as a model that cannot reliably perform a task generates no token demand for a specific task, while a model that can perform the task reliably generates demand from every user who values that task. The transition from "cannot do" to "can do" is rather discrete, but the demand unlocked by that transition can be substantial. When Sonnet-3.5 demonstrated reliable code generation, it unlocked demand from millions of developers who would not have used a less capable model, along with the emergence of vibe coding, while a development like reasoning models demonstrated reliable multi-step problem-solving that unlocked demand for users with more complex analytical needs.

This unlocking dynamic means that token demand is a step function of capability thresholds, not a smooth function of capability level. Below a threshold, demand is zero or negligible. Above the threshold, demand can be very large. The threshold varies by use case, with some applications requiring only modest capability and others requiring frontier performance, and as models improve, they cross more thresholds, unlocking more use cases, generating more demand. This dynamic explains why token growth projections have consistently underestimated actual growth as projections based on extrapolating current usage patterns fail to account for use cases that do not yet exist because capability thresholds have not yet been crossed. And what we'd argue here is that the current projection methodology captures demand from current users doing current tasks but misses demand from future users doing tasks that current models cannot perform.

The Math on Inference Scaling

Token consumption per query is also increasing, independent of query volume growth. Test-time compute scaling means that complex queries consume more tokens than simple queries. A reasoning model that "thinks" before responding generates internal tokens that contribute to compute consumption even if they are not visible to the user. And this ratio of internal to external tokens can be substantial, with some reasoning queries generating ten or more internal tokens per external token. Furthermore, context length expansion increases tokens per query through a different mechanism, as simply put, users with access to longer context windows use them. Document analysis queries that previously required summarization or chunking can now process full documents in a single context while conversations that previously lost context over multiple turns can now maintain coherent state. The token consumption scales with context utilization, and context utilization scales with context availability. The combination of more queries, more tokens per query, and more internal reasoning tokens creates multiplicative growth in total token volume. For instance, if query count grows at 50% annually, tokens per query grow at 30% annually, and reasoning overhead adds another 20% to average token consumption, the composite growth rate exceeds 100% annually. And while these figures are illustrative, the actual multiplicative structure is real and implies growth rates that substantially exceed any single component.

Early Innings

The characterization of AI adoption as being in "early innings" is sometimes dismissed as promotional language, but the token volume data supports the claim in a specific sense. Current token consumption represents a small fraction of the token consumption that would occur if AI capabilities were applied to all economically valuable tasks they could address. Consider the potential token demand from software development alone. Global developer count is approximately 30M and if each developer used an AI coding assistant for four hours per day, generating an average of 1,000 tokens per minute of active use, the daily token demand from this single use case would be approximately 7T tokens. Current actual usage is a small fraction of this potential, reflecting both adoption lag and capability gaps that limit utility for complex coding tasks.

Similar calculations can be performed for customer service, content creation, data analysis, research, education, and other knowledge work categories. In each case, the potential token demand assuming full adoption and adequate capability substantially exceeds current actual demand. The gap between potential and actual represents the growth opportunity that improves as capabilities advance and adoption spreads. Again, the actual implication for infrastructure planning is that current demand is a poor guide to future demand. Capacity built for current token volumes will be insufficient for token volumes two or three years hence if the capability-adoption feedback loop continues operating, and this is something we've mentioned numerous times over the past year, especially as people doubt the capabilities of AI systems and what AGI could actually look like.

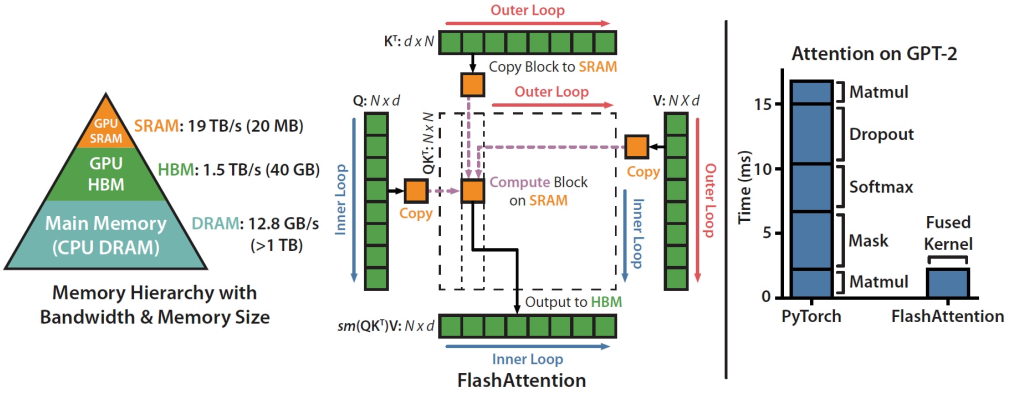
A New Scaling Paradigm Emerges

The preceding themes established that compute access determines capability at the frontier and that closed-source labs hold substantial compute advantages over open-source alternatives. A natural inference from these claims might be that open-source progress will stall as the compute gap widens. And while there may be some shifting of the dynamics as the gap in compute widens, historical evidence would suggest that we shouldn't be concerned. Resource constraints, rather than halting progress, often results in algorithmic and architectural innovations that shift the compute-capability frontier itself, which has repeated itself across the history of machine learning and has manifested clearly in the past couple of years.

The Historical Pattern

When compute is abundant and scaling is the path of least resistance, organizations allocate engineering effort toward scaling rather than efficiency, however, when compute is constrained, the same engineering talent redirects toward extracting more capability per FLOP. This just means that the constraint changes the optimization target, and the changed target produces different innovations. Let's take the Transformer architecture as an example of this dynamic, though not in the direction that is often assumed. Google Brain developed the Transformer in 2017 when it had access to substantial compute (relatively speaking), and while this innovation wasn't necessarily driven by the constraint, it was driven by the opportunity to design an architecture that could exploit parallelism better than the alternatives. However, the subsequent efficiency innovations that made Transformers practical for widespread deployment came largely from organizations that had far less compute than Google. For instance, Flash Attention, which reduces the memory and compute requirements of attention mechanisms by restructuring the computation to minimize memory bandwidth, was developed at Stanford and Princeton. The technique enables training longer sequences on fixed hardware by exploiting the memory hierarchy more effectively. Further, sparse attention variants, mixture-of-experts (MoE) architectures, and quantization techniques followed similar patterns, with efficiency innovations often emerging from academic labs or smaller companies rather than the compute-rich frontier labs.

Figure 16: Flash Attention Architecture



Source: Medium

And this pattern that we see does follow some kind of economic logic. The reality is that frontier labs with abundant compute face opportunity costs when allocating engineering talent to efficiency improvements. An engineer working on architectural efficiency is not working on scaling the next training run or building product features. When scaling produces reliable capability gains, the expected value of efficiency work is lower than the expected value of scaling work. However, constrained-environments inverts this calculation, as an organization that cannot scale has no opportunity cost from efficiency work, because scaling is not an option. This does not mean that frontier labs ignore efficiency, rather they just invest substantially in making their training and inference more efficient. But their efficiency work is complementary to scaling rather than a substitute for it and they can optimize to maximize capability at a given scale, then increase scale. On the other hand, constrained organizations optimize to maximize capability at fixed scale because fixed scale is their only option.

The 2024-2025 Precedent

Over the past two years, we've seen this dynamic play out visibly. An example that we've used a few times is that in 2024, OpenAI released the first reasoning model o1, which demonstrated that test-time compute scaling could produce substantial capability gains on reasoning tasks. The model's chain-of-thought reasoning, trained through reinforcement learning, set a new benchmark for mathematical and analytical problem-solving. At the time, OpenAI appeared to be pulling ahead of competitors through a novel scaling dimension that required both architectural innovation and extensive compute for training, however, DeepSeek's response illustrated this exact constraint-driven innovation pattern. Unable to match OpenAI's compute for either pre-training or large-scale reinforcement learning, DeepSeek developed techniques that achieved competitive reasoning performance at dramatically lower compute cost. Multi-head latent attention reduced memory requirements, efficient expert routing reduced communication overhead in mixture-of-experts architectures, and targeted reinforcement learning with verifiable rewards achieved reasoning capabilities without the extensive human feedback data that OpenAI could afford to collect.

The DeepSeek-V3 and R1 releases demonstrated that the techniques worked as models trained at reported costs of single-digit millions achieved benchmark performance competitive with models trained at costs of hundreds of millions. This meant that at the time, the gap between frontier closed-source models and the best open-source alternatives narrowed, not because open-source compute access improved but because constrained organizations found ways to extract more capability per FLOP. And these exact innovations did not remain constrained to their originators, as we know that multi-head latent attention, efficient MoE routing, and RLVR techniques have been adopted by numerous labs across the industry, including by compute-rich frontier labs. This means that the efficiency improvements developed under constraint become industry-standard techniques that benefit all participants, and this diffusion pattern is characteristic of algorithmic innovations, which unlike compute cannot be monopolized through capital expenditure.

Our Projection for the New Year

This year, we're expecting this exact same dynamic to play out, but this time with even greater intensity, as the compute gap between the closed-source frontier labs and their constrained counterparts is widening as gigawatt-scale clusters come online. This widening gap only increases the pressure that constrained organizations will feel to innovate, since the alternative is falling further behind on the capabilities that matter the most. And while we do believe that the gap in the most economically relevant tasks will widen between these labs, we think at the very least it will spur accelerating progress in the architecture and algorithmic spaces, with innovations emerging that target current techniques that have inefficiencies or unexploited structure. To this end, we'd argue that architecture remains a productive area, as the Transformer at this point is seven years old, but as many have stated this past year, it is not necessarily optimal. A breakthrough in alternative architectures could enable training at scales currently infeasible with standard transformers, benefiting constrained organizations disproportionately, or it could require less training but higher forms of intelligence through other mechanisms of knowledge acquisition. We'd also argue that training dynamics offers another area for innovation, as current practices involve extensive hyperparameter tuning, learning rate schedules, and training stability techniques developed through expensive experimentation. Better theoretical understanding of training dynamics could reduce the experimentation required, lowering the effective compute cost of developing new models, while data efficiency improvements could multiply the effective value of available compute, and techniques for curriculum learning, data mixing, and synthetic data generation determine how much capability each training token contributes. Any improvements in these techniques would greatly benefit organizations that cannot simply scale token count through brute-force data collection.

So our take for this year isn't that any specific innovation will emerge, but rather that the constraint-innovation dynamic will produce something of value, and below, we're putting out a few candidates that we feel are worth watching, though we may be wrong altogether and be surprised by innovation in a new area.

Candidate Innovations to Watch

Predicting which specific innovations will emerge from constrained organizations is inherently speculative. However, the categories where innovation might occur are identifiable based on current technical limitations and unexploited structure in existing approaches. Three categories merit particular attention for 2026 as potential sources of efficiency gains or capability improvements that could shift the compute-capability frontier.

Verification Compute as a Scaling Vector

As a result of all this algorithmic and architectural improvement, we believe we'll begin to see more emphasis on verification compute as the primary engine of capability improvement in domains where correctness can be meaningfully adjudicated. And we'd contend that this is the natural next step once the field has learned two lessons the way. First being that raw scale is not enough to make reasoning reliable and then secondly, the bottleneck in turning impressive behavior into trustworthy behavior is not always the model's capacity. When the reward is based on a checkable notion of correctness rather than on imitation or subjective preference, training becomes less dependent on scarce human annotation and more dependent on the ability to generate, critique, and filter candidate solutions at scale. Which we'd argue is close to post-training but deserves to have its own independent recognition. In the verification regime, the scarce resource is no longer simply the amount of gradient compute you can spend on a model, but rather the scarce resource is grading capacity. The system that improves fastest is not necessarily the one with the largest base model, but the one that can scale the act of judging. On this point, meta-verification is the second piece that makes this vector durable because once verification becomes valuable, the immediate failure mode is rather obvious. Verifiers become new sources of error, bias, and overconfidence, and a naive verifier can be gamed by a generator that learns the quirks of its grader rather than the substance of the domain.

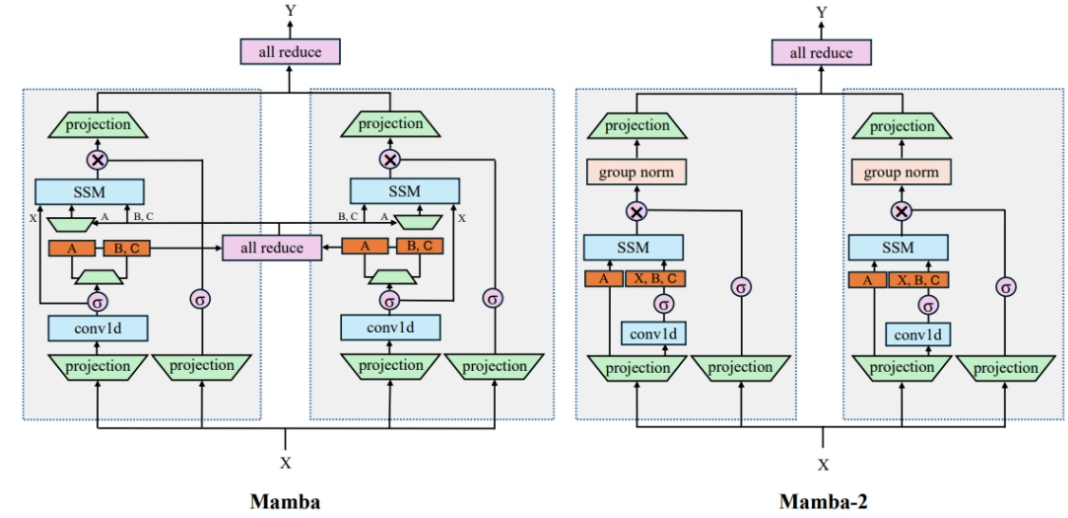
Meta-verification then is applied to the verifier itself, a layer that asks whether a critique is real, whether a score is justified, and whether the evaluation process is behaving coherently. Which in practice allows the system to escape the human grader bottleneck without surrendering to an ungrounded self-referential loop. The verifier then can scale because it is being policed, and the generator can improve because it is being graded by a mechanism that is harder to exploit and easier to audit. A reason why we might be more inclined to consider this as a candidate for a new scaling vector is because it can be scaled in two directions at once. It scales during training because you can label and filter vast quantities of candidate solutions with an automated grading stack, and it scales at inference because the model can spend compute on generating multiple candidates and letting verification select the one that survives scrutiny. Another reason why we think the shape of progress in 2026 will look different from the last few years, as we'll see the largest gains accrue first in verifiable domains such as mathematics, coding, and structured scientific workflows, not because those are the only domains that matter (which we could argue they are), but because they are the ones where verification can supply dense feedback and where dense feedback is the fastest path to truly impressive and game-changing capabilities.

Architectural Candidates

As we stated previously, the Transformer architecture has dominated for seven years, but it is not necessarily optimal for all computational characteristics of interest. We'd note that several alternative or complementary architectures address specific limitations of Transformers and could provide efficiency advantages if they prove scalable.

State-space models for instance, particularly the Mamba family, offer linear-time sequence processing as an alternative to the quadratic scaling of attention. The core mechanism is a selective state-space layer that compresses sequence history into a fixed-size state vector, updating the state based on each new token. And unlike attention, which computes pairwise relationships between all tokens in a window, state-space models maintain a running summary that is updated incrementally. Mamba-2, which was released in 2024, improved on the original architecture by introducing constraints that enable recurrent updates to be reformulated as matrix multiplications optimized for GPU tensor cores. This hardware-algorithm co-design increased training speed by 2-8x compared to Mamba-1 while expanding the state dimension from 16 to 128, bringing expressiveness closer to transformers. The selective mechanism allows the model to choose which information to retain and which to discard based on input content, addressing a key weakness of earlier linear-time models that compressed indiscriminately.

Figure 17: Mamba Architecture

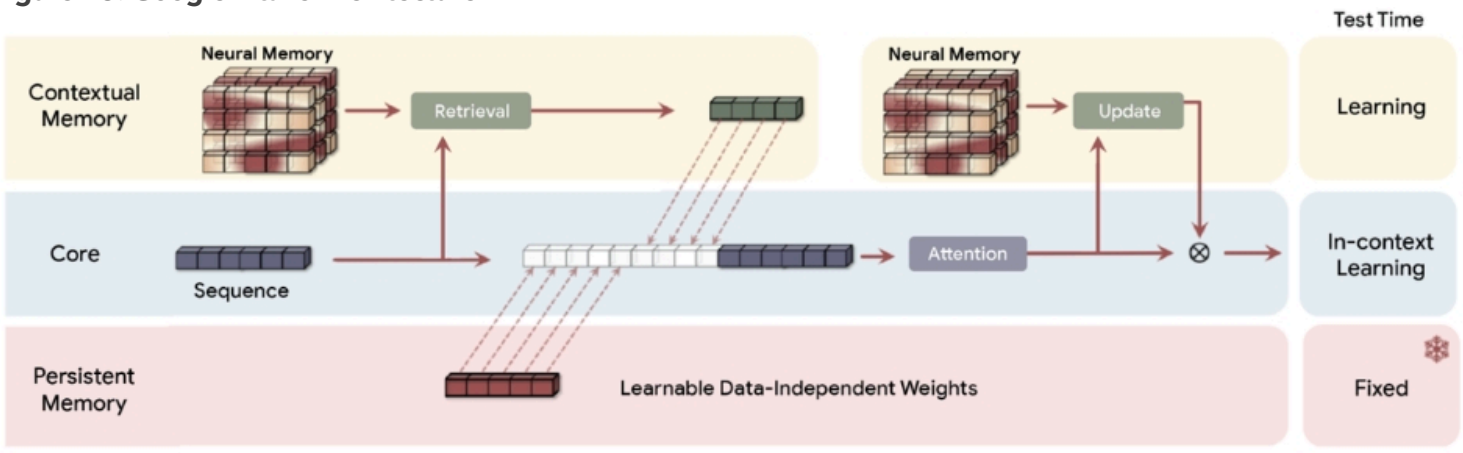


Source: NVIDIA

Google's Titans architecture, which was published in late 2024 and announced formally in late 2025 alongside the MIRAS framework, addresses a different kind of limitation. Transformers and traditional recurrent models both struggle with very long contexts, either because of quadratic compute scaling in attention or because fixed-size state compression loses information. Titans introduces a neural long-term memory module that learns to memorize during inference. Unlike the fixed-size vector memory in traditional recurrent networks, this memory is itself a neural network, specifically a multi-layer perceptron, that updates its own weights as data streams through. The memory module uses a "surprise metric" to prioritize what to store, so when a new input differs substantially from what the memory predicts, the gradient of the prediction error is large, signaling that the input is surprising and should be retained. When an input matches predictions, the gradient is small, and the memory can safely skip permanent storage. This selective memorization enables the architecture to scale to context windows exceeding 2M tokens while maintaining accurate retrieval on needle-in-haystack tasks.

Titans presents three architectural variants for incorporating long-term memory. Memory as Context uses the memory module's output as additional context for attention, Memory as Gate combines memory and attention outputs through a learned gating mechanism, and Memory as Layer stacks memory and attention layers sequentially. Each offers different trade-offs between efficiency and capability, with the Memory as Layer variant enabling use of the long-term memory module without any attention, functioning purely as a sequence model.

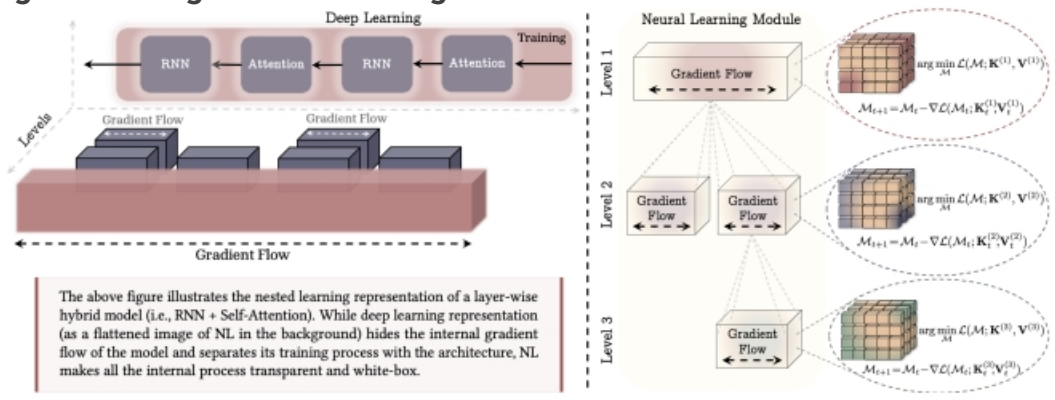
Figure 18: Google Titans Architecture



Source: "Titans: Learning to Memorize at Test Time", Behrouz et al.

Finally, Google's Nested Learning paradigm, also introduced in late 2025, extends the Titans concept by unifying architecture and optimization into a single framework. The core idea is that both neural network architectures and training optimizers can be formalized as associative memories operating at different timescales. For instance, attention maps tokens to tokens within a context window, while momentum in SGD maps gradients to parameter updates across training steps, and weight updates in backpropagation map loss gradients to parameter changes. The HOPE architecture, built on Nested Learning principles, implements a hierarchy of persistent memory experts, each governed by different optimizers and update frequencies. Some components update rapidly to handle transient information while others retain information for extended periods, shaping longer-term behavior. The architecture can learn to modify its own update rules through self-referential optimization, creating recursive learning loops with early results show improved performance on long-context tasks and continual learning benchmarks compared to standard transformers and other modern recurrent models.

Figure 19: Google Nested Learning



Source: "Nested Learning: The Illusion of Deep Learning Architecture", Behrouz et al.

Model Gardening and the Economics of Mid-Training

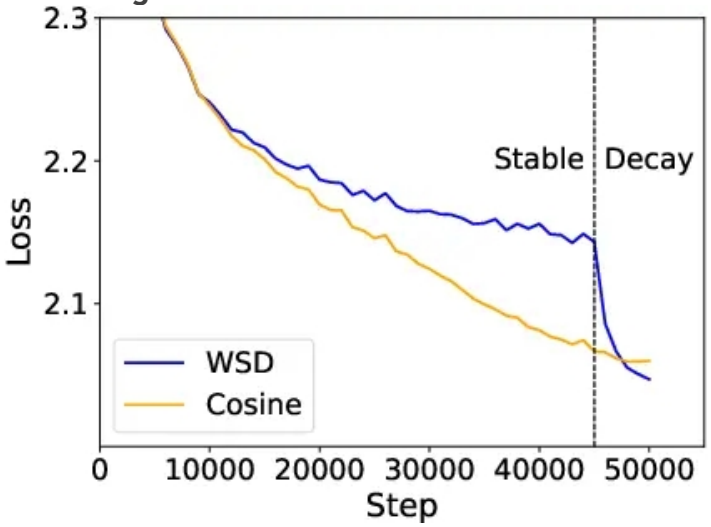
The traditional view of model development treats pre-training as a discrete phase that produces a single base model, which then proceeds through post-training to become a deployable system. However we'd argue that this view is becoming more obsolete, as the emerging paradigm is likely to treat pre-training as an ongoing process that produces not a single model, but a family of related models that are branching from shared training trajectories. We think the easiest way to categorize this is by calling it "model gardening", because like cultivating plants from a common rootstock, model gardening involves maintaining a living pre-training run from which future versions of models can be built from and extended from.

The Premise of Model Gardening

The economic logic of model gardening follows from the cost structure of large-scale pre-training. A frontier training run consumes tens to hundreds of millions of dollars in compute. The vast majority of this cost goes toward the initial phase of pre-training, where the model acquires broad knowledge and general capabilities from diverse data while the later phases of training, where models are specialized for particular tasks or optimized for specific deployment scenarios, consume a much smaller fraction of total compute (though we'd highlight various models like Grok 4 that spent as much compute on post-training as they did on pre-training). This cost asymmetry creates an opportunity because if the expensive initial phase can be shared across multiple model variants or generations, the marginal cost of producing an additional model can drop dramatically. Rather than running separate pre-training processes, a lab can extend any single pre-training process and branch it from any given checkpoint, allowing primarily compute-bound labs to squeeze more juice out of the theoretical base model lemon.

This approach does require slightly rethinking how pre-training runs are structured because traditional runs typically use cosine learning rate schedules that assumed a fixed total training budget. The learning rate would warm up, remain high through most of training, then decay smoothly to near zero as the run concluded. This schedule produces good final models but creates a problem for continuation. Once the learning rate has decayed, resuming training requires either accepting suboptimal learning dynamics or restarting from an earlier checkpoint and recomputing the decayed portion. The Warmup-Stable-Decay schedule, adopted by DeepSeek, Kimi Moonshot, and increasingly by other labs, addresses this limitation. After an initial warmup, the learning rate remains constant through an extended stable phase, and decay occurs only at the end, when the training run is genuinely concluding. The stable phase can be extended indefinitely, with checkpoints saved at regular intervals, so when developing a new model or an extended version (i.e. DeepSeek-V3.2), training branches from a stable-phase checkpoint, applies the appropriate decay schedule, and produces a final model, while the main branch continues at constant learning rate, available for future branching.

Figure 20: Cosine Learning Rate Schedule vs. WSD Scheduling



Source: "Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective", Wen et al.

Technical Details

The mechanics of checkpoint branching involve several interrelated choices. The first is checkpoint frequency during the stable phase. More frequent checkpoints provide finer-grained options for branching but increase storage costs and I/O overhead. Typical intervals range from every few billion to every few tens of billions of tokens, depending on the scale of training and storage constraints. The second choice is the branching strategy itself. A branch can modify multiple aspects of training simultaneously, and data composition can shift to emphasize particular domains. The Allen AI OLMo approach treats mid-training as a distinct phase where high-quality, instruction-adjacent data is upsampled while maintaining the pre-training objective. This targeted data exposure during the final portion of pre-training, rather than waiting for supervised fine-tuning, improves downstream task performance with minimal additional compute.

Context length extension represents another common branching objective. Models trained on sequences of 4,096 or 8,192 tokens can be extended to 128,000 or longer through continued training on long documents with modified position encoding. The context extension phase often precedes other post-training stages, as the extended context capability benefits subsequent instruction tuning. However, learning rate management during branching presents subtle tradeoffs. The standard approach applies a decay schedule to the branch, reducing the learning rate from its stable-phase value to near zero over the course of specialization training. Recent work on checkpoint merging, however, suggests that decay may not be strictly necessary. The WSM framework demonstrates that averaging checkpoints from continued constant-learning-rate training can approximate the effect of learning rate decay while providing greater flexibility.

The checkpoint merging insight connects to broader questions about the geometry of training. The loss landscape of large language models appears to have a "river valley" structure, where progress occurs along a relatively narrow channel surrounded by higher-loss regions. Learning rate decay serves to settle the model into the valley floor, reducing oscillation and improving generalization. But averaging across checkpoints achieves a similar effect by canceling out the oscillations arithmetically. The Schedule-Free optimization approach takes this logic further, maintaining implicit averaging throughout training and eliminating the need for explicit decay phases entirely. For model gardening, these techniques expand the design space, as a lab maintaining a long-running pre-training process can branch at any point without committing to a decay schedule, with branches themselves be capable of being averaged, merged, or extended.

Implications on Model Training

In 2026, while we expect clean-sheet pre-training to continue and to grow even larger in scale, we think that this phenomena of model gardening will mean that large-scale pre-training runs will occur less often. This means the inter-release interval should widen because each new base model is becoming a major capital project in its own right, and because the opportunity cost of starting over rises as post-training, evaluation, and test-time scaffolding consume a larger share of what it takes to ship frontier capability. That cadence shift though has a pretty simple consequence, which is that the space between base models becomes the real battleground and the dominant work in that space is a continuation of mid-training and layered post-training that steadily expands the capability, reliability, and commercial shelf life of a single base model. This pattern essentially exists in embryonic form across much of the field, however, we'd argue that most labs are still leaving too much capability on the table because they treat a babes model as a stepping stone rather than as an asset to be cultivated. Compute-bounded labs do not have that same sort of luxury as they are forced to squeeze more juice from the same foundation, which results in them pursuing successive rounds of mid-training continuation, targeted data curation, reinforcement learning regime where feedback can be made stable, and repeated post-training passes that convert a general model into a more reliable tool for specific classes of work. When this process is executed well, the gains can be surprisingly durable as a compute-bounded lab that cannot simply leap to the next enormous pre-training run instead stretches the useful life of a base through repeated refinements, sometimes to the point where the latest iteration remains competitive with closed models in narrow but economically meaningful subdomains.

While this has been going on through last year, what changes this year is that this approach stops being a niche adaptation and becomes an obvious economic strategy even for the compute-privileged. Compute-rich labs will still build the next giant base model but the point is that they will increasingly behave as if the base is only the beginning rather than the main event. While they've been doing this to some extent, they will spend more time in the in-between, and they will do so deliberately, because the marginal return on additional mid-training and post-training is often higher than the marginal return on launching an entirely new backbone before the current one has been fully exploited. It's merely a recognition that pre-training scale produces a platform while refinement produces a product. And as competitive pressure intensifies and investors become more sensitive to the efficiency of capital deployment, it becomes rational to treat each base model as a multi-release franchise where you're trying to maximize the useful life of each base model.

Headline STEM Breakthrough and the AI Research Intern

This theme makes two related predictions for 2026, which we think are the most "out there" predictions in this piece. First, we're predicting that at we will see a headline STEM breakthrough on the order similar to that of a Millennium Prize Problem that was solved either with the assistance of an AI system, or by an AI system outright. Our second prediction is that we believe models will reach a threshold of usefulness in research workflows where they're capable of functioning as credible research interns, capable of executing well-specified research tasks with quality sufficient for a senior researcher to build upon their output rather than merely correct it. We'd note that these predictions are related because the same underlying capabilities drive both outcomes, which we will get into on the following sections.

On one hand, the AI research intern seems like a natural endpoint of the shifts we've been describing the past few trends. Once progress is concentrated in verifiable, tool-mediated regimes, capability where the model can finally start to look like a system that can run a workflow. Meaning a credible research intern is simply the point where that workflow becomes dependable enough to be delegated. The AI research intern can propose an approach, implement it in code, run the experiment or proof attempt, interpret the output, and then iterate while maintaining coherence and context across steps. So what is essentially the consensus view of what an agent should be capable of doing, but limited to research in STEM fields as opposed to an enterprise agent running around various day-to-day white collar workflows. And because the work being done in STEM can typically be objectively checked with lower cost of iteration, we believe those domains will be where we see the earliest credible versions of this intern and even why deployments might be more internal-facing rather than a broadly deployed, available to the average consumer version. Our contention is that frontier labs will use these AI research interns and future iterations to accelerate their own research loops because the ROI becomes nearly immediate and because the failure modes can be relatively contained.

On the other hand, we believe the same dynamics we've been mentioning meaningfully raise the probability of significant breakthroughs in STEM fields to the point where our second prediction within this theme is that we will see a headline STEM breakthrough on the likes of a Millennium Prize Problem, which are often considered some of the most difficult unsolved mathematics equations pertaining to various field such as computer science and fluid dynamics. This would be done either by AI directly or as a meaningful collaborator, which is what leads us to believe that the potential headline breakthrough for the year could be Google DeepMind providing either a genuine proof or counterexample to the existence and smoothness of Navier-Stokes, one of the Millennium Prize Problems. If this does occur, we believe it will be the first widely legible evidence that the compute being poured into AI is not only buying better chatbots but also buying a step-change in the rate at which humanity can do verifiable science.

What a Millennium Prize Solution Would Validate

Just for context, since many may not be aware of what the Millennium Prize Problems are, they represent a specific class of mathematical challenges deemed sufficiently important and difficult that the Clay Mathematics Institute offered $1M for each solution. Of the seven original problems posed in 2000, only one has been solved by human effort, the Poincaré Conjecture by Grigori Perelman in 2003, while the remaining six problems have resisted decades of work by the world's strongest mathematicians. In short, an AI solution to a Millennium Prize Problem would provide strong evidence about AI reasoning capabilities as these problems were specifically selected to require deep mathematical insight that cannot be achieved through brute-force search or pattern matching on known techniques. Meaning a solution would demonstrate that AI systems can achieve genuinely novel mathematical reasoning, not merely recombine existing ideas.

More specifically, a Millennium Prize solution would validate several capability claims. It would demonstrate sustained reasoning over extremely long inference chains, as these problems require building complex arguments with many interdependent steps, while also demonstrating creative hypothesis generation, as solutions will require new mathematical ideas not present in training data. Additionally, it would demonstrate robustness to distribution shift, as the problems are specifically chosen to resist known techniques, and the solving of a problem at this level would clearly show the ability to integrate ideas across mathematical subfields as most problems require combining insights from multiple areas. In short, we believe that solving such a problem would serve as a massive validation for the entire AI field, and tangentially, all the investments that have been made over the past few years into research and compute buildout, as these are clearly defined problems that essentially no human being or group has been capable of solving beyond one problem for the past 26 years.

Why Verifiable Domains are the Main Focus

The predictions in the preceding section cluster around domains with a shared characteristic. Mathematics, code, formal proofs, and computational chemistry all permit automated verification of correctness. The reason this happens to be the case is because verifiable domains offer fundamentally different training dynamics than domains where evaluation requires human judgment. And understanding why verification matters explains both where AI research capabilities will emerge first and why frontier labs are concentrating investment in these areas.

The Validation Problem

Training AI systems to perform complex tasks requires feedback signals that indicate whether outputs are correct. In reinforcement learning frameworks, these signals take the form of rewards. In supervised learning, they take the form of labeled examples. In both cases, the quality and availability of feedback determines what a system can learn and how quickly it can improve. Non-verifiable domains impose severe constraints on feedback quality. Consider training a system to generate novel scientific hypotheses. Evaluating whether a hypothesis is good requires domain expertise, takes substantial time, and produces subjective assessments that may differ across evaluators. The feedback loop operates on a timescale of hours to days per evaluation, limits training to samples that human experts have reviewed, and introduces noise from evaluator disagreement. Scaling up training data requires proportional scaling of expert evaluation effort, which is expensive and slow. Verifiable domains eliminate these constraints. A mathematical proof either verifies in Lean or it does not. Code either passes its test suite or it fails. A chemical synthesis route either produces the target compound or it does not. The feedback is binary, immediate, and consistent, and no human needs to evaluate each sample. The same compute that generates candidate outputs can verify them, enabling training on millions or billions of samples rather than thousands.

The difference in training dynamics compounds over time. Consider two hypothetical research projects with equal starting capabilities. Project A operates in a non-verifiable domain where each training sample requires 30 minutes of expert evaluation. Project B operates in a verifiable domain where samples can be evaluated in milliseconds by an automated checker. Assuming equal compute budgets, Project B can evaluate roughly one million times more samples per unit time. And over months of training, this difference translates into qualitatively different final capabilities. AlphaProof extends the verification advantage further. By operating in the formal language Lean, where proofs can be mechanically verified, the system can train on millions of proof attempts. Each failed attempt provides signal about which proof strategies do not work, while each successful proof provides positive reinforcement and can be decomposed to identify which intermediate steps were productive. The formal verification environment enables a form of curriculum learning where the system generates and solves progressively harder problems, with correctness guaranteed at each step. This training regime produced a system capable of solving IMO competition problems, including problems that only five of 609 human competitors solved.

Why Labs are Investing Here

As we've stated a few times now, one of the most important things to understand is that frontier AI labs face a resource allocation problem. Training frontier models requires billions of dollars in compute, data, and engineering effort. The return on this investment depends on the capabilities that result, so labs must chose which capabilities to prioritize given limited resources. In this light, verifiable domains receive disproportionate investment for reasons beyond their tractability for training, especially since these verifiable domains include the core activities of AI research itself. Machine learning is fundamentally a computational discipline, developing new architectures requires writing code, analyzing training dynamics requires mathematical reasoning, and designing experiments requires formal specification of hypotheses and success criteria. If AI systems can accelerate these activities, they accelerate the pace at which labs can develop better AI systems.

This creates a compounding dynamic, and one that has been outlined several times including in the infamous AI 2027 paper. A lab that develops AI tools capable of assisting with AI research gains a significant advantage that grows quickly over time because each improvement in AI research capability enables faster development of the next generation of models. A lab with superior AI research tools can iterate more quickly, explore more architectural variants, and identify promising approaches before competitors. And since models are capable of doing research on a much shorter timescale than humans, after a couple of times through this compounding loop, it would not be unreasonable to assume that years worth of research can be done in weeks or days even. All of this also means that the concentration of investment in verifiable domains has implications for where breakthroughs will occur. Capabilities emerge where training is most effective while training is most effective where feedback is most reliable, and feedback is most reliable in verifiable domains. This logical chain we believe predicts that headline breakthroughs and research intern capabilities will appear first in mathematics, code, formal proofs, and computational science, then potentially transfer to domains where verification is harder.

Deep Tech Progress Report

We feel like for the longest time, deep tech has been treated as some far-away technology that for the foreseeable future will be full of demos and little adoption. However, in 2026, we expect deep tech to stop being viewed as this collection of interesting prototypes and to start being viewed like a set of adoption curves that have visible slopes. This year, we believe the clearest acceleration should come from autonomous vehicles, reusable rockets, stablecoins, and humanoid robotics, where deployment expands in ways that are tangible to end users and customers and that increasingly shows up in market narratives as real volume rather than promise. We also expect meaningful progress to be made in small modular reactors and quantum computing, however with limited broad adoption beyond niche use-cases because the binding constraints sit upstream of deployment and do not compress into a single year. The important distinction we believe is worth making is not that winners are fully solved or that the laggards are stagnant, but rather that the distinction is the first group is positioned to scale real world usage in 2026 in a way that becomes visible in daily life and enterprise operations, while the later two remain gated by timelines, regulatory pathways, and technical prerequisites that can advance materially without translating into widespread adoption on the same calendar.

Autonomous Vehicles

Autonomous vehicles enter this new year with a different kind of momentum than the category has had in prior years because adoption is no longer confined to a narrow set of demos or pilot programs. Last year, Waymo was the clearest proof point we've seen yet as paid rides continue to scale and the footprint continues to broaden within their current locations and even to new geographies across the United States. What we think makes 2026 especially investable though is that this stops being interpreted as a single winner story. Waymo remains the most visible consumer deployment, but the field is widening with credible competition in the form of Tesla's robotaxi ambitions, China-linked operators such as WeRide and Pony.ai pursuing commercial expansion, and a growing enabling layer where Uber functions far less like a pure ride share company and more as distribution and demand aggregation for autonomous fleets that want access to riders without building a full consumer marketplace from scratch.

Figure 21: Autonomous Vehicle Adoption in Major Metropolitan Cities

Source: World Economic Forum, Boston Consulting Group

The more counter intuitive part of this belief is that we expect adoption to accelerate alongside rising scrutiny rather than because scrutiny as we've found out, eventually fades. As autonomous services become more common, regulators have stronger incentives to formalize oversight and the public has more occasions to notice failures, outages, and edge case incidents. And while those sound a lot like major roadblocks, they still do not necessarily slow the adoption curve. We'd argue they can actually accompany steepening adoption when the service is already delivering real convenience and when operators respond by improving safety processes, tightening operational envelopes, and scaling methodically into geographies where the regulatory pathway is tractable. Waymo's own experience has already shown that operational setbacks can coexist with continued expansion, which is exactly the pattern we think becomes more normal in 2026. Or in other words, the category's maturation does not look so much like a smooth reduction in controversy, but rather looks like a service that becomes more routine while the social and regulatory apparatus around it becomes more demanding, and the winners are the operators and platforms that can keep scaling through that friction rather than waiting for the world to become comfortable first.

Humanoid Robotics

Humanoid robotics in 2026 is best viewed as the year the category as a whole begins to earn the right to be discussed in terms of utilization rather than teleoperated choreography. Over the past couple of years, the public image of humanoids was dominated by staged demonstrations that proved these kinds of robots could move, balance, and perform various scripted tasks, but they did not prove that they could operate day after day in a way that created economic value. Thus, the shift we expect in 2026 is that a subset of humanoid platforms begin crossing over from just being impressive demos into early deployment loops that run daily, primarily in constrained commercial environments where verification is straightforward and safety envelopes can be more tightly defined. This means use-cases in warehouses, logistics settings, manufacturing support tasks, and other structured spaces which are by no means glamorous, but they are the natural beachheads because they provide repeatable work, consistent instrumentation, and fast feedback when performance starts to degrade. This is where humanoids can start accumulating the only dataset that matters for adoption which is real world operation under constraints that resemble real world production.

Figure 22: Tesla Optimus Generations

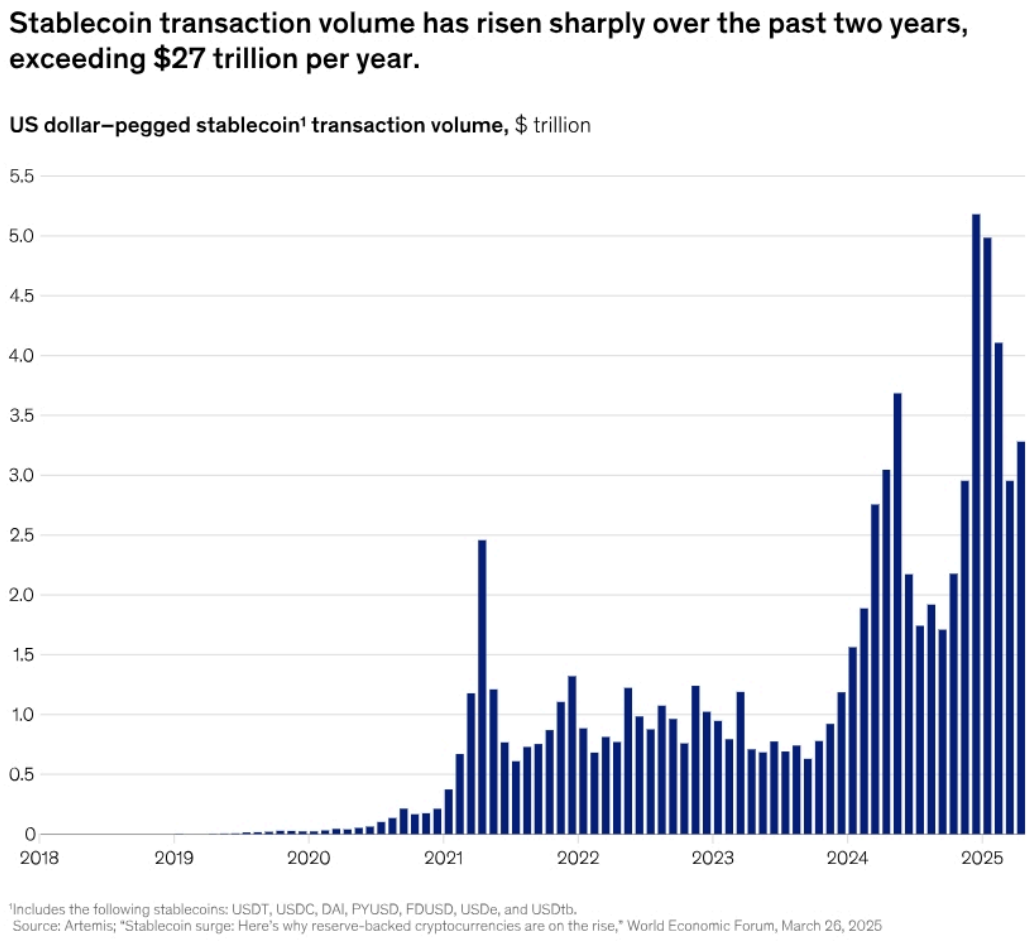
Source: Tesla

At the same time though, we expect the entire humanoid robotics category to become more culturally tangible through small niche domestic use-cases, even if those use-cases remain narrow and heavily constrained. The importance nuance though is that cultural visibility isn't the same as adoption even though it sometimes feels like it. A handful of household demonstrations can make the category feel real to the public without implying that the product is ready for broad home deployment. What changes in 2026 is that we think it delivers both a modest increase in public visibility and the first credible utilization data points that indicate these kinds of robots are beginning to do real work. That being said, we're expecting this numbers to be small and even feel trivial, with the capabilities being narrower than the hype would imply, but the presence of repeatable operational loops is what changes the entire conversation.

Stablecoins

This year, it's our belief that stablecoins will finally be understood as a payments rail that is being pulled into the institutional perimeter rather than as a purely crypto native phenomenon. Adoption accelerates when the marginal friction for an incumbent to engage falls below the threshold where ignoring the technology is the riskier option. That is why regulatory progress matters so much more than any single product or feature announcement, and why we believe this year will look so much different than the previous ones. The recent GENIUS Act functions as the anchor event because it moves the United States closer to a coherent framework for payment stablecoins and pushes reserve quality, disclosure, and compliance expectations toward something institutions can underwrite without improvising their own standards, which has been a bottleneck in the past. And it's our view that when this kind of uncertainty declines, the adoption curve will steepen almost mechanically with more issuers that will be capable of operating with clearer constraints, more counterparties will be capable of accepting stablecoin flows without bespoke legal frameworks, and more enterprises will be able to consider stablecoin settlement as a treasury and working capital tool rather than as an experimental cryptocurrency workflow.

Figure 23: Stablecoin Adoption Trends



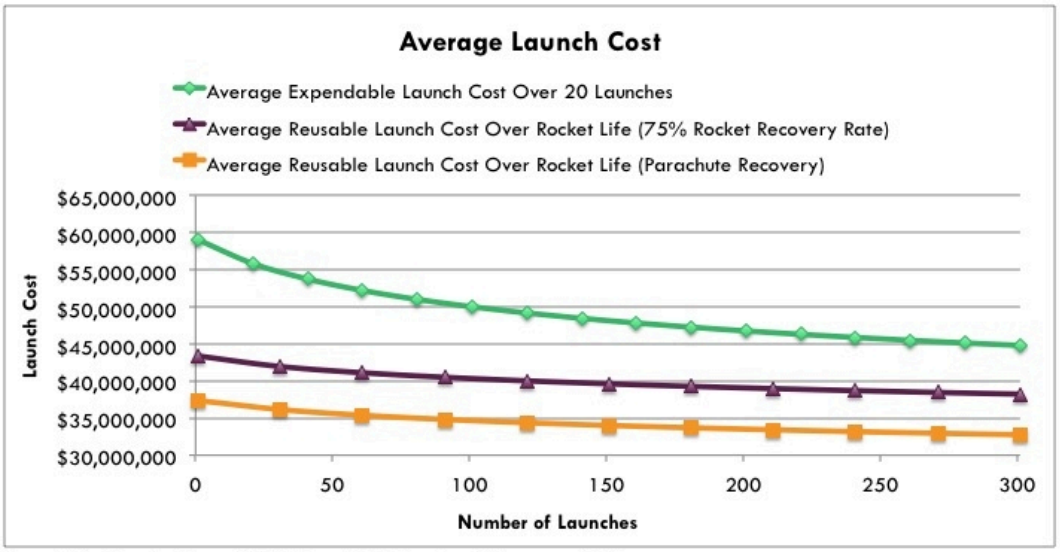
Source: McKinsey & Co.

The second part to this story and maybe the most important is that incumbents are no longer treating stablecoins as optional infrastructure because once the regulatory pathway becomes more legible, larger networks and platforms that already sit in the flow of commerce have incentives to integrate stablecoins as a settlement layer and as a programmable extension of existing payment primitives. Mastercard's stablecoin initiatives are emblematic of the direction of travel because they are reflecting this world where stablecoins are being normalized inside mainstream network logic as opposed to being taped off as a separate ecosystem. And at the end of the day, the entire bull-case on stablecoins has always been about moving them, not holding them. They're about settlement speed, cross border transfers, treasury operations, merchant integration, and the gradual embedding of stablecoin rails into payment stacks where the user experience can remain familiar even as the settlement layer underneath changes. In that world, stablecoins do not need to replace traditional systems overnight, as they only need to become a trusted a widely available option that progressively captures the flows where they are structurally superior.

Reusable Rockets

Reusable rockets will be best understood through two variables that matter more than any particular vehicle design. The first being launch cadence and the second being reuse, as those two together determine whether the category looks more like manufacturing or like bespoke aerospace. In the private markets, we've already seen SpaceX demonstrate what happens when reuse becomes more routine and cadence begins to compound, but in the public markets for 2026, we believe the story is that the reusable launch stack broadens beyond just a single dominant supplier. That matters for adoption because launch is a throughput business and as we have more reliable capacity and more providers, that'll translate into more missions, faster timelines, and a lower friction path for satellite constellations, defense payloads, and emerging in space infrastructure to scale.

Figure 24: Reusable Rocket Economics (2016)



Source: Ark Invest

It is also worth mentioning that with credible reporting that SpaceX is planning on moving towards an IPO in late 2026, we believe this will function as a major catalyst even before any formal filing. A potential public listing gives investors a benchmark anchor that the market has never had for arguably the most important deep tech franchise of the last decade. As we've seen with other IPOs that involve bellwethers in their respective spaces, a SpaceX IPO will pull attention onto the entire ecosystem around launch, including public comparable, key suppliers, satellite manufacturers, and the downstream services that become more viable when access to orbit is cheaper and more frequent. We'd also mention that the operational substance behind the broader attention is relatively straightforward. In 2025, the industry reached new records for global orbital launch cadence, with SpaceX accounting for a large share, and the path in 2026 is still one of increasing throughput rather than a plateau. To this end, we believe the next leg of expansion comes from incremental new capacity sources that make the ecosystem less single threaded, which would include rockets like Blue Origin's New Glenn reaching orbit which is an important milestone since it adds another credible heavy lift pathway, while Rocket Lab's Neutron represents the next wave of reusable economics entering the medium to heavy segment with a design built around higher cadence.

Small-Modular Reactors

Small modular reactors this year is where we are deciding to be more disciplined around timelines rather than narratives, though by no means are we trying to take an anti-nuclear stance. The long run case for small modular reactors remains compelling precisely because firm, dispatchable power becomes more valuable in a world of rising electricity demands, especially demands that are being significantly accelerated by our need for AI compute. However, our claim for this year is simply that adoption, in the sense most investors implicitly mean it, will lag expectations because the binding constraints sit in licensing, construction sequencing, supply chain readiness, and project financing structures that do not compress into a single year and we would have to have already made significant progress in these areas for us to see meaningful adoption in 2026. Even with strong demand pull from data centers and industrial users, one cannot shortcut regulatory processes, nor can they pour concrete and complete commissioning on venture timelines, and it's near impossible to scale a new nuclear supply chain without multi-year coordination.

All of this is why we expect this year to look like real progress in the SMR space but without broad output. The visible milestones should look more like approvals, site preparation, permitting and licensing advances, and deeper institutional alignment between developers, utilities, regulators, and capital providers. Those are the prerequisites that determine whether SMRs can eventually scale, but they do not translate into meaningful adoption beyond any kind of niche demonstrations in the same calendar year. That being said, we don't want investors to confuse momentum with deployment, as momentum will likely increase as the energy conversation becomes more urgent, however, deployment will remain limited because the system is still upstream constrained.

Quantum Computing

Quantum computing in 2026 needs some kind of category correction because much of the public discourse still assumes a replacement narrative that we believe is structurally wrong at this point in time. Quantum right now isn't on the path to take over broad classical workloads in the near to medium term, and the relevant question for adoption we should be asking is not whether quantum becomes generally useful, but whether it becomes selectively decisive at the upper edge of computational difficulty where classical methods are not sufficient. The plausible adoption arc here is that quantum becomes valuable when it can deliver a narrowly defined advantage in specific problem families that are both economically meaningful and technically aligned with what quantum systems can plausibly do under realistic error budgets. Chemistry and materials discovery have been the canonical examples because they map naturally into quantum, while certain optimization classes and carefully constructed simulation settings also belong in this conversation. So while we do think there will be a lot of progress made in the quantum space this year, it will not look like a broad workload migration in the same way that cloud adoption has looked nor do we think we'll see the broad, legible quantum advantage that many investors are expecting to see in the near-term.

Another point we'd like to bring up that we've only brought up in passing, is that adoption is not only gated by the actual hardware and devices. We'd argue that quantum progress is gated by the surrounding ecosystem that turns a device into a usable tool chain, and that ecosystem depends on talent, capital, and a credible promise of payoff. This is why we believe the field likely requires a breakthrough moment that is legible enough to reset expectations and attract a step change in attention. Meaning a result that can convince academics that the most important problems are now within reach, and that convinces venture capital and corporate R&D that scaling the software, error correction stack, and application layer is worth the cost. A lot of what we saw in the AI space immediately following the ChatGPT-moment. And without this catalytic event, progress can continue in a steady and respectable way while adoption remains niche since the incentives to build the full surrounding stack remain too weak to accelerate the ecosystem fast enough.

Potential Risks

Required Disclosures

D.A. Davidson & Co, or any of its affiliates, does or seeks to do business with companies covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision.

D.A. Davidson & Co. is a full service investment firm that provides both brokerage and investment banking services. Alexander Platt, the research analyst principally responsible for the preparation of this report has received and is eligible to receive compensation, including bonus compensation, based on D.A. Davidson's overall operating revenues, including revenues generated by its investment banking and institutional equities activities. D.A. Davidson & Co.'s analysts, however, are not directly compensated for involvement in specific investment banking transactions.
I, Alexander Platt, attest that (i) all the views expressed in this research report accurately reflect my personal views about the common stock of the subject company, and (ii) no part of my compensation was, is, or will be, directly or indirectly, related to the specific recommendations or views expressed in this report.

Rating Information
D.A. Davidson & Co.'s Institutional Research Rating Scale Definitions (maintained since October 10, 2017); information regarding our previous definitions is available upon request:
BUY: Expected to produce a total return of over 15% on a risk adjusted basis over the next 12-18 months
NEUTRAL: Expected to produce a total return of -15% to +15% on a risk adjusted basis over the next 12-18 months
UNDERPERFORM: Expected to lose value of over 15% on a risk adjusted basis over the next 12-18 months

Rating Distribution (as of 12/31/25)	Coverage Universe Distribution			Investment Banking Distribution		
	IR	WMR	Combined	IR	WMR	Combined
BUY (Buy)	60%	85%	63%	8%	0%	8%
NEUTRAL (Hold)	40%	13%	36%	4%	0%	3%
UNDERPERFORM (Sell)	0%	2%	1%	0%	0%	0%

IR denotes Institutional Research; WMR denotes Wealth Management Research whose rating scale is Buy/Add, Neutral, Sell/Reduce. Investment Banking Distribution denotes companies from whom D.A. Davidson & Co. has received compensation in the last 12 months. Best-of-Breed: Expected to outperform on a risk adjusted basis over a five-year time horizon.

Target prices are our Institutional Research Department's evaluation of price potential over the next 12 months, based upon our assessment of future earnings and cash flow, comparable company valuations, growth prospects and other financial criteria. Certain risks may impede achievement of these price targets including, but not limited to, broader market and macroeconomic fluctuations and unforeseen changes in the subject company's fundamentals or business trends.
While the Best-of-Breed designation does not contain a separate rating and/or price target from that of the standard ratings system referenced above, the expectation is that the security, based on the 12 criteria utilized in assessing the "Best-of-Breed" designation, will outperform over a five-year time horizon, not the standard 12-18 month time horizon.

For a copy of the most recent reports containing all required disclosure information for covered companies referenced in this report, please contact your D.A. Davidson & Co. representative or call 1-800-755-7848.

Other Disclosures

Information contained herein has been obtained by sources we consider reliable, but is not guaranteed and we are not soliciting any action based upon it. Any opinions expressed are based on our interpretation of data available to us at the time of the original publication of the report. These opinions are subject to change at any time without notice. Investors must bear in mind that inherent in investments are the risks of fluctuating prices and the uncertainties of dividends, rates of return and yield. Investors should also remember that past performance is not necessarily an indicator of future performance and D.A. Davidson & Co. makes no guarantee, express or implied, as to future performance. Investors should note this report was prepared by D.A. Davidson & Co.'s Institutional Research Department for distribution to D.A. Davidson & Co.'s institutional investor clients and assumes a certain level of investment sophistication on the part of the recipient. Readers, who are not institutional investors or other market professionals, should seek the advice of their individual investment advisor for an explanation of this report's contents, and should always seek such advisor's advice before making any investment decisions. Consensus estimates are obtained from Capital IQ. Further information and elaboration will be furnished upon request.

Other Companies Mentioned in this Report

Company Name	Ticker	Rating	Price
Apple Inc.	AAPL	NEUTRAL	$246.70
Amazon.com, Inc.	AMZN	BUY	$231.00
CoreWeave, Inc.	CRWV	NEUTRAL	$95.22
Alphabet Inc.	GOOGL	NEUTRAL	$322.00
IonQ, Inc.	IONQ	NEUTRAL	$50.66

Company Name	Ticker	Rating	Price
Meta Platforms, Inc.	META	BUY	$604.12
Microsoft Corporation	MSFT	BUY	$454.52
Nebius Group N.V.	NBIS	BUY	$99.29
NVIDIA Corporation	NVDA	BUY	$178.07
Oracle Corporation	ORCL	NEUTRAL	$179.92