THE AI COMPUTE FLYWHEEL

Why AI Infrastructure Demand May Be Larger and More Durable Than Traditional Financial Models Assume

A Research Report on Compute Demand, Model Capability, Usage Growth, and Nvidia's Strategic Position

Prepared May 27, 2026

Core thesis: AI compute demand is driven by a two-pronged flywheel - capability growth and usage growth. Better hardware enables better models; better models enable broader and heavier usage; broader usage creates more inference demand; and that demand reinforces the need for more compute.

Table of Contents

Executive Summary

This report argues that AI compute demand should not be analyzed only through the lens of a traditional semiconductor cycle. Traditional models tend to assume that explosive growth is followed by digestion, normalization, margin compression, and a return toward ordinary growth rates. That pattern is often sensible in semiconductors. However, AI compute has a distinctive property: the infrastructure does not merely serve a fixed pool of existing demand. It can improve the product that creates the next wave of demand.

The central framework is a two-pronged compute flywheel. The first prong is capability growth: better hardware enables better models, stronger reasoning, better coding, longer context, better multimodal systems, better agents, and eventually better physical AI. The second prong is usage growth: better models produce more usage across chat, search, coding, enterprise workflows, agents, video, robotics, and other inference-heavy products. Those two prongs reinforce each other. More compute can make models better; better models can increase usage; greater usage requires more inference capacity; and larger inference demand funds and justifies the next wave of compute investment.

This matters because the psychology of AI labs and hyperscalers is different from ordinary IT budgeting. If leading labs see that additional compute continues to produce meaningfully better models, they are unlikely to voluntarily slow down simply because financial models say growth should normalize. The strategic risk of underinvesting is too high. In a frontier race, compute is not merely a cost; it is research velocity, product capability, competitive positioning, user experience, and platform leverage.

The current data supports the argument that demand remains aggressive. Nvidia reported Q1 fiscal 2027 revenue of $81.6 billion, up 85% year over year, with Data Center revenue of $75.2 billion, up 92% year over year, and guided Q2 fiscal 2027 revenue to $91 billion [1]. Microsoft reported Q3 FY2026 capital expenditures of $31.9 billion, with roughly two-thirds directed toward short-lived assets primarily GPUs and CPUs [4]. Amazon projected 2026 capital expenditures of roughly $200 billion [6]. Alphabet guided 2026 capex to $175 billion to $185 billion [7]. Meta raised its 2026 capex forecast to $125 billion to $145 billion [8]. CoreWeave reported revenue backlog of $99.4 billion as of March 31, 2026 [9]. These figures suggest that large buyers are not behaving as though AI infrastructure demand is exhausted.

The usage side is also scaling rapidly. Google reported that the Gemini app surpassed 900 million monthly active users, more than doubling in a year, while daily requests grew more than seven times [15]. OpenAI said ChatGPT reached 900 million weekly active users and 50 million paying subscribers [16]. These are not niche-adoption numbers. They indicate that consumer AI usage is already extremely large, even before fully mature enterprise agents, autonomous workflows, physical AI, and broader embedded AI become mainstream.

The main limitation to the thesis is not likely to be voluntary restraint from AI labs. The real brakes are physical and economic: power, energized data-center capacity, HBM supply, advanced packaging, networking, cooling, construction timelines, custom silicon competition, margin pressure, enterprise ROI, export controls, and the possibility that model capability gains eventually

flatten. The International Energy Agency projects global data-center electricity consumption roughly doubling from 485 TWh in 2025 to 950 TWh in 2030, with AI-focused data centers growing much faster than overall data-center demand [10]. That points to the central constraint: the world may want to build AI factories faster than the physical infrastructure can support.

The investment implication is that Nvidia's valuation cannot be evaluated only by looking at a current or trailing multiple. If earnings compound rapidly, a seemingly high P/E can collapse quickly on a cost-basis basis. A starting P/E of 20x becomes 6.8x after two years if EPS grows 95% and then 50%. It becomes 5.3x after three years if the third-year growth rate is 30%. At a 32x starting P/E, the same three-year path brings the effective multiple to 8.4x. The point is not that these growth rates are guaranteed. The point is that the earnings path matters more than the optical starting multiple.

The conclusion is that the strongest AI compute thesis is technical and behavioral before it is financial. If better hardware keeps producing better models, and better models keep producing heavier usage, the leading buyers are likely to keep pressing. The financial model comes after that. The central question is not whether Nvidia's growth decelerates mathematically. It will. The central question is whether deceleration reflects saturation or merely a larger denominator while the compute flywheel continues to turn.

Exhibit 1: The Two-Pronged Compute Demand Framework

Prong	Demand Driver	Why it matters for compute
1. Capability growth	Labs buy compute to train and improve models	More compute can improve reasoning, coding, agents, multimodal capability, and product quality.
2. Usage growth	Users and enterprises consume more AI through inference	More chat, search, agents, workflows, video, robotics, and physical AI create recurring inference demand.
Flywheel effect	Capability and usage reinforce each other	Better models drive more usage; more usage funds and justifies more compute.

1. The Central Thesis: AI Compute Demand Is Reflexive

A conventional technology hardware cycle begins with an upgrade wave, moves into aggressive capital spending, and eventually slows as customers digest the new capacity. That pattern explains many semiconductor cycles. Customers pull forward demand, supply catches up, margins peak, and growth rates normalize.

AI compute may not follow that pattern cleanly because the infrastructure has a reflexive quality. Compute does not simply serve current workloads. Compute can improve models, and improved models can create new workloads. That distinction is central.

A traditional data-center buildout can be summarized as demand first, then capacity. A company forecasts its existing or expected workloads and builds enough capacity to serve them. Once capacity is sufficient, spending slows. In AI, the sequence can be different: build capacity, train better models,

unlock better products, increase usage, and then require even more capacity. The capacity can help create the demand it later serves.

This is why the AI infrastructure cycle may be more durable than a standard cyclical chip cycle. If additional compute keeps producing better capability, then the customer does not simply ask how much capacity is needed for today's usage. The customer asks how much compute is needed to avoid losing the next model generation.

The distinction is especially important because frontier AI is competitive. Model quality matters. Latency matters. Inference cost matters. Developer adoption matters. Enterprise adoption matters. The company with better models and better infrastructure can attract more users, more developers, more workflows, more data, and more capital. This makes the downside of underinvesting unusually severe.

The compute buyer's decision is therefore not just a spreadsheet exercise. It is a strategic survival decision. If a lab believes that more compute still produces material capability gains, and that those gains lead to better products and broader usage, then cutting capex too early can be a self-inflicted wound.

This report refers to that dynamic as the AI compute flywheel. It is a loop in which capability and usage reinforce each other:

- More compute enables better models.
- Better models enable more useful products.
- More useful products drive more usage.
- More usage creates more inference demand.
- More inference demand justifies more compute.
- More compute funds the next model generation.

The bull case for Nvidia and AI infrastructure is not merely that current demand is high. It is that the flywheel may continue to turn for several years because both prongs - capability and usage - remain early.

2. The Two-Pronged Framework

The core framework has two prongs.

Prong one is capability demand. AI labs need compute to create better models. If larger and better training runs, better hardware, improved memory bandwidth, faster networking, and better software continue to translate into improved model performance, then labs have a strong incentive to keep buying compute. The compute purchase is not just infrastructure spending; it is a direct input into model capability.

Prong two is usage demand. Once models improve, users consume more AI. That includes normal chatbot usage, search, code generation, enterprise workflows, document analysis, Excel and finance agents, research agents, video generation, robotics, voice assistants, autonomous workflows, and physical AI. As usage expands, inference demand increases. Inference is not a one-time training expense. It is the recurring cost of running the model every time a user or system calls it.

The two prongs reinforce each other. Better models make the products more useful. More useful products generate more usage. More usage makes the infrastructure more valuable. Higher usage and revenue justify more compute investment, which then enables the next model generation.

This two-pronged framework is useful because it avoids a common analytical mistake: treating AI demand as if it is only training demand. Training demand is crucial, but inference demand may become larger and more recurring over time. If AI becomes embedded into daily workflows, the demand for inference capacity can be massive.

It also avoids the opposite mistake: treating usage demand as if it is independent of model capability. Users do not adopt AI simply because it exists. They adopt it because the models become useful enough. Capability drives usefulness, and usefulness drives usage.

Figure 1: The Compute Flywheel

Compute supply -> model capability -> product usefulness -> user and enterprise usage -> inference demand -> revenue and strategic pressure -> compute supply.

The key feature of the loop is that demand is not fixed. Demand expands as capability expands. This is what can make traditional deceleration models too conservative.

Exhibit 2: How AI Demand Differs From a Traditional Hardware Cycle

Traditional hardware cycle	AI compute flywheel
Capacity is built to serve known demand.	Capacity can create better models, which create new demand.
Growth slows once customers digest capacity.	Growth may continue if capability and usage keep scaling.
Infrastructure is mostly a cost center.	Compute is a strategic input into product capability.
Efficiency may reduce replacement demand.	Efficiency may expand AI usage and total inference demand.

3. Prong One: Capability Demand

Prong one is the demand for compute created by the pursuit of better models. The technical basis for this prong is that model performance has historically improved with scale. The original neural language-model scaling-laws literature found that model loss scales in a power-law relationship with model size, data, and compute across many orders of magnitude [13]. DeepMind's Chinchilla work later showed that training compute should be allocated more efficiently between model size and data, and that a smaller model trained on more tokens could outperform larger undertrained models at the same compute budget [14].

The practical takeaway is not that scaling is simple or infinite. It is that compute has been a reliable input into capability when combined with enough data, good architecture, and good training methods. That is why labs keep competing on clusters, chips, memory, networking, and training infrastructure.

Epoch AI estimates that training compute for frontier language models has grown roughly 5x per year since 2020, doubling about every 5.2 months. It also estimates pre-training compute efficiency has improved about 3x per year [12]. This combination is important. Labs are using more compute, but they are also getting more efficient at turning compute into performance.

Efficiency does not automatically reduce total demand. In many technology markets, efficiency expands the market. Cheaper compute enabled more software. Cheaper storage enabled more data. Faster broadband enabled richer internet applications. Lower inference costs can make AI usable in more products, more workflows, and more markets. Therefore, better efficiency can reduce cost per task while increasing total tasks.

For frontier labs, capability gains are strategic. Better models can improve reasoning, code generation, tool use, multimodal understanding, scientific analysis, document comprehension, video generation, and agents. Each improvement potentially unlocks new product categories. This means that compute is not a passive infrastructure input. It is a lever for product creation.

The psychology is aggressive because the frontier is competitive. A lab that slows compute investment while competitors continue scaling risks falling behind. If a competitor ships a model that is materially better at coding, research, enterprise automation, or multimodal work, users and developers may shift quickly. The history of consumer and developer platforms suggests that quality gaps can become distribution gaps, and distribution gaps can become durable market-position gaps.

This is why capability demand can remain intense even if current usage were temporarily less profitable than investors want. Labs are not only optimizing current-period margins. They are trying to win the next platform layer. Compute is one of the main inputs into that race.

The hardware cadence reinforces the dynamic. Nvidia's platform roadmap from Hopper to Blackwell to Rubin gives labs new performance, efficiency, memory, and system-level capabilities. Nvidia management has stated confidence in $1 trillion of Blackwell and Rubin revenue from 2025 through calendar 2027 [3]. That statement should not be treated as a guarantee, but it indicates that Nvidia believes demand visibility extends beyond a single-quarter spike.

Capability demand will weaken if scaling weakens. That is one of the most important risks. If additional compute stops producing meaningful improvements, the psychology changes. The strategic urgency of buying more compute falls. But as of the latest public evidence, the leading labs and hyperscalers are not acting as though scaling has stopped. They are acting as though compute remains a scarce strategic input.

4. Prong Two: Usage and Inference Demand

Prong two is the demand created when better models are actually used. This is where the thesis becomes larger than training clusters.

Training receives enormous attention because it creates the frontier model. However, inference is the recurring activity that happens each time the model is used. A user asks a question, an agent reads a file, a coding tool writes and tests code, a search product generates an answer, a legal agent analyzes documents, a finance agent processes a workbook, a video model generates clips, or a robot interprets sensor inputs. Each of those actions consumes inference compute.

Usage is already very large. OpenAI said ChatGPT reached 900 million weekly active users and 50 million paying subscribers [16]. Google reported that the Gemini app surpassed 900 million monthly active users, more than doubling in a year, and that daily requests grew more than seven times [15]. These figures suggest that AI usage has moved far beyond early adopters.

But the more important point is that usage intensity can rise. The average AI interaction today may be relatively light compared with future workflows. A short chatbot question is one type of usage. A coding agent that works for twenty minutes, writes code, runs tests, reads logs, fixes bugs, and produces a finished pull request is much heavier. A finance agent that reviews a workbook, writes Python, creates charts, cross-checks totals, generates commentary, and updates a file is also much heavier.

Agents are the key accelerant. A chatbot usually responds once. An agent may plan, act, check, recover, call tools, and retry. One user request can become many model calls. That means agents can increase compute per user even if the number of users remains constant.

Enterprise workflows amplify the effect. In a consumer chatbot, usage is discretionary. In an enterprise workflow, usage can become embedded into the operating process. A model that classifies transactions, reviews contracts, creates financial commentary, answers accounting research questions, summarizes tickets, or drafts code can become part of daily production. Once that happens, inference is not a novelty cost. It is operating infrastructure.

Physical AI expands the usage side further. Robotics, autonomous vehicles, drones, industrial automation, warehouse systems, and humanoids require training, simulation, and real-time inference. The physical world contains far more potential tasks than office software. If AI begins to control machines in the real world, compute demand expands beyond chat, search, and enterprise SaaS.

The key point is that usage demand can grow both horizontally and vertically. Horizontal growth means more users and more organizations. Vertical growth means more compute per user or per workflow. Agents and physical AI increase vertical intensity because they require longer reasoning, more tool use, more context, more retries, and more continuous inference.

This is why the usage prong is so important to Nvidia. If AI remains mostly a chatbot layer, demand is still large. If AI becomes a workflow layer and then a physical-action layer, demand is much larger. The most powerful version of the thesis is that AI becomes embedded in software, work, search, coding, media, science, and machines. In that world, inference capacity becomes a core economic input.

Exhibit 3: Selected Current Demand Indicators

Indicator	Evidence	Source
Nvidia Q1 FY2027 revenue	$81.6B, up 85% YoY	[1]
Nvidia Q1 FY2027 Data Center revenue	$75.2B, up 92% YoY	[1]
Nvidia Q2 FY2027 guide	$91B revenue, plus or minus 2%	[1]
Microsoft Q3 FY2026 capex	$31.9B; roughly two-thirds GPUs/CPUs	[4]

Amazon 2026 capex projection	About $200B	[6]
Alphabet 2026 capex guide	$175B-$185B	[7]
Meta 2026 capex guide	$125B-$145B	[8]
CoreWeave backlog	$99.4B as of Mar. 31, 2026	[9]
Gemini app usage	900M+ monthly active users; daily requests up 7x	[15]
ChatGPT usage	900M weekly active users; 50M paying subscribers	[16]

5. Why the Buyer Psychology Is Different

The most important difference between AI infrastructure and ordinary IT infrastructure is the psychology of the buyer. Ordinary IT buyers typically want enough capacity to support expected demand at acceptable cost. AI labs and hyperscalers are also sensitive to cost, but they are operating under a strategic race dynamic.

If a lab believes more compute will produce better models, the cost of underinvesting is severe. Falling behind does not merely mean missing one quarter of growth. It can mean losing developer attention, enterprise credibility, consumer mindshare, distribution partnerships, and technical talent. The frontier lab that ships the strongest model can reset the market's expectations overnight.

This creates a competitive fear loop. Each lab sees other labs raising capital, securing data centers, contracting for GPUs, building custom silicon, and hiring talent. If the internal experiments show that scaling continues to improve capability, the rational response is not restraint. The rational response is to secure more compute before competitors do.

This does not mean buyers ignore economics. They still care about inference cost, utilization, gross margin, customer ROI, and capital efficiency. But those considerations are balanced against strategic survival. In a market that may define the next platform layer of computing, a company can rationally spend ahead of near-term profit.

This is why the phrase capex discipline can be misleading. Discipline does not necessarily mean lower spending. It can mean spending aggressively where the strategic return is highest. If management believes AI is the most important platform transition in decades, then very high capex can be disciplined if it builds scarce capacity, model advantage, and future revenue.

The psychology is similar to an arms race, but not purely irrational. It is grounded in observed capability gains, user adoption, and product opportunity. The labs are not buying compute because compute itself is valuable. They are buying compute because it may create better models, and better models may create better products, and better products may capture enormous value.

This is also why analysts can underestimate demand if they rely too heavily on ordinary cyclical mean reversion. A financial model can see a large base and fade growth. A lab sees a competitor that may use the next cluster to train a better model. Those are different perspectives.

The real turning point would occur if labs stop seeing meaningful returns to additional compute. If model improvements flatten, if agents fail, if user demand disappoints, or if inference economics do not work, the psychology will change. But absent those conditions, voluntary restraint is unlikely to dominate.

6. Current Evidence: Nvidia, Hyperscalers, AI Clouds, and Users

The public evidence currently supports continued aggressive demand. Nvidia's Q1 fiscal 2027 results were extraordinary: revenue of $81.6 billion, up 85% year over year; Data Center revenue of $75.2 billion, up 92% year over year; GAAP and non-GAAP gross margins of 74.9% and 75.0%; and non-GAAP diluted EPS of $1.87 [1]. Nvidia also guided Q2 fiscal 2027 revenue to $91 billion, plus or minus 2% [1].

For fiscal 2026, Nvidia reported revenue of $215.9 billion, up 65% year over year, with fiscal-year non-GAAP gross margin of 71.3% [2]. This means the company entered FY2027 from an already enormous base and still produced a major step-up in quarterly revenue. That pattern is consistent with a market where demand remains extremely strong.

Hyperscaler capex confirms that customers are still building. Microsoft reported Q3 FY2026 capital expenditures of $31.9 billion, with roughly two-thirds spent on short-lived assets primarily GPUs and CPUs [4]. In Q2 FY2026, Microsoft had said customer demand continued to exceed supply and again noted that roughly two-thirds of capex was on GPUs and CPUs [5]. Amazon projected 2026 capital expenditures of about $200 billion, up from $131 billion in 2025, tied to AI infrastructure and AWS demand [6]. Alphabet guided 2026 capex to $175 billion to $185 billion [7]. Meta raised its 2026 capex forecast to $125 billion to $145 billion [8].

AI cloud demand is also visible. CoreWeave reported revenue backlog of $99.4 billion as of March 31, 2026 [9]. Backlog is not the same as realized revenue, and profitability still matters, but it shows that demand for AI cloud capacity extends beyond the largest hyperscalers.

Consumer usage is already massive. ChatGPT's 900 million weekly active users and 50 million paying subscribers indicate broad adoption [16]. Google's Gemini app surpassing 900 million monthly active users, with daily requests up more than seven times year over year, shows that usage growth is not isolated to one company [15].

AI infrastructure mega-projects also support the thesis. OpenAI's Stargate Project was announced as a plan to invest $500 billion over four years in AI infrastructure for OpenAI in the United States, with $100 billion intended to begin deployment immediately [17]. Reuters later reported OpenAI, Oracle, and SoftBank planning five new AI data-center sites tied to the $500 billion Stargate initiative [18]. These projects reflect the same strategic psychology: secure large-scale compute capacity because the frontier race requires it.

None of this guarantees that spending will generate adequate returns. But it demonstrates that the buyers are not currently behaving as though AI compute is a one-quarter anomaly. They are behaving as if compute is central to their product roadmaps and strategic positioning.

Exhibit 4: Major AI Infrastructure Constraints

Constraint	Why it matters
Power and grid interconnection	AI clusters require large amounts of reliable electricity; energized capacity may become the limiting input.
HBM and advanced memory	High-bandwidth memory is critical for accelerator performance and system supply.
Advanced packaging	System-level accelerator production depends on complex packaging capacity.
Networking and optics	Large training and inference clusters require high-bandwidth, low-latency connections.
Cooling and data-center shells	Dense AI racks need specialized cooling and physical deployment capacity.
Software and utilization	Hardware value depends on scheduling, serving, libraries, and workload optimization.

7. Bottlenecks: Power, Memory, Packaging, Data Centers, and Deployment

The strongest practical constraint on the AI compute thesis may not be customer desire. It may be physical infrastructure.

AI data centers require enormous amounts of power, land, cooling, electrical equipment, transformers, high-voltage interconnections, construction labor, networking, memory, advanced packaging, and operational expertise. Even if customers want more compute, revenue can only be recognized when systems are built, delivered, installed, powered, and used.

The International Energy Agency projects global electricity consumption from data centers roughly doubling from 485 TWh in 2025 to 950 TWh in 2030, with AI-focused data centers growing much faster than overall data-center electricity consumption [10]. In another IEA analysis, global data-center electricity consumption is projected to grow about 15% per year from 2024 to 2030, more than four times faster than electricity consumption from all other sectors [11].

This is why AI infrastructure has become an industrial buildout, not just a software trend. Data centers need grid interconnections. They need backup power. They need cooling. They need construction schedules. They need permits. A GPU cluster is only valuable if the facility can power and operate it.

Memory and packaging are also major bottlenecks. AI accelerators rely on high-bandwidth memory, advanced packaging, and dense system integration. Nvidia's advantage is partly that it provides a full system-level architecture, not only a chip. However, full-stack systems also require coordination across a complex supply chain.

Networking matters because large-scale training and inference require clusters to operate as integrated systems. The challenge is not merely acquiring chips. It is building systems where thousands or tens of thousands of accelerators communicate efficiently. Latency, bandwidth, software orchestration, scheduling, and reliability all matter.

These bottlenecks can produce uneven revenue patterns. Demand may be real but delayed. A data-center project may slip. HBM supply may constrain shipment volume. Power availability may limit

deployment. Networking components may create system-level shortages. These constraints can make quarterly growth volatile even if underlying demand remains strong.

The bottlenecks also support pricing. If demand exceeds supply, Nvidia and other infrastructure suppliers retain bargaining power. That is one reason Nvidia's gross margins remain high. But bottlenecks can also create risk. If customers cannot deploy fast enough, if power becomes too expensive, or if the economics of new data centers deteriorate, growth could slow.

The key analytical distinction is that infrastructure constraints are different from demand exhaustion. A slowdown caused by lack of power or supply-chain tightness is not the same as a slowdown caused by customers no longer wanting compute. Investors need to separate those causes.

8. Why FY2028 Deceleration Estimates May Be Too Neat

Analysts are rational to model deceleration. Nvidia's base is already enormous. Growth rates cannot remain near 90% indefinitely. As revenue gets larger, each incremental percentage point represents a larger absolute-dollar amount. Forecast visibility also declines further out in time, so analysts typically fade growth unless they have firm evidence.

The issue is not that deceleration is wrong. Deceleration is inevitable. The issue is whether the modeled deceleration reflects actual demand saturation or simply cautious modeling on a larger base.

A company growing from $200 billion to $300 billion in revenue has added $100 billion. A company growing from $300 billion to $420 billion has slowed from 50% growth to 40% growth but added $120 billion. Growth rate deceleration can coexist with larger absolute-dollar growth. This matters for Nvidia because the base is now so large that even lower percentage growth can produce enormous earnings expansion.

The two-pronged framework suggests that analysts may understate FY2028 if they treat FY2027 as a one-time surge followed by normal digestion. If capability continues improving and usage continues expanding, then demand may remain stronger than a standard fade model assumes.

There are several reasons analysts may still fade FY2028 growth. First, they may have higher visibility into FY2027 orders than FY2028 orders. Second, they may be cautious about product transitions from Blackwell to Rubin. S&P Global noted that estimates for certain Nvidia product lines showed wide ranges, reflecting uncertainty around ramp timing and adoption [19]. Third, analysts may assume hyperscalers eventually digest capacity. Fourth, they may assume margins normalize as competition rises. Fifth, they may haircut China and export-control uncertainty. Sixth, they may not want to underwrite multiple years of extreme growth after a historic run.

Those are reasonable modeling choices. But they are different from saying demand is ending.

The best interpretation is that FY2028 consensus may embed a conservative view of visibility and normalization. It may not fully capture a scenario where agentic AI, inference scaling, enterprise automation, AI clouds, sovereign AI, and physical AI continue to increase demand.

A useful question is not whether FY2028 growth slows. It almost certainly will compared with peak growth rates. The better question is whether FY2028 remains a very high-growth year in absolute

revenue and EPS terms. If it does, today's valuation could still compress quickly on realized earnings.

9. Nvidia-Specific Implications

Nvidia's strategic position is stronger than a narrow GPU-vendor description implies. The company is increasingly selling the architecture of AI factories: accelerators, networking, CPUs, systems, software, libraries, and a developer ecosystem. At frontier scale, performance is not determined by a single chip specification. It depends on the entire system.

This system-level position matters because AI workloads are cluster-scale. Large training runs require thousands of accelerators to operate together. Large inference services require throughput, low latency, batching, scheduling, memory management, and reliability. Agents add further complexity because one user workflow may involve many model calls and tool interactions. Physical AI can require simulation, sensor processing, edge inference, and real-time response.

Nvidia's Q1 FY2027 results show the magnitude of this positioning. Data Center revenue was $75.2 billion in a single quarter [1]. Nvidia management also stated confidence in $1 trillion of Blackwell and Rubin revenue from 2025 through calendar 2027 [3]. Again, this is not a guarantee; it is management commentary. But it is significant because it frames demand as a multi-year platform ramp, not merely a single product spike.

The company's margins show pricing power. Nvidia's Q1 FY2027 non-GAAP gross margin was 75.0% [1]. For fiscal 2026, non-GAAP gross margin was 71.3% [2]. If those margins hold while revenue scales, EPS can grow very quickly. If margins normalize sharply, the EPS path becomes less powerful. Therefore, gross margin is one of the most important indicators to watch.

Competition is real. Google TPUs, Amazon Trainium and Inferentia, Microsoft Maia, AMD Instinct, Broadcom custom silicon, and other accelerators all matter. Some hyperscalers will shift certain workloads to internal silicon, especially stable high-volume inference workloads where cost optimization is critical. That could pressure Nvidia's share and margins over time.

However, custom silicon does not necessarily destroy the Nvidia thesis. The total market can expand fast enough for Nvidia to grow despite share loss in some workloads. Nvidia may remain preferred for frontier training, rapid deployment, AI cloud customers, enterprises, and workloads where software maturity and time-to-market matter. A hyperscaler using custom chips internally does not eliminate the need for Nvidia across the broader market.

The biggest Nvidia-specific risk is not simply competition. It is competition plus margin pressure plus weaker-than-expected demand. If custom silicon takes high-volume inference, if customers push pricing down, if supply constraints ease, and if usage growth disappoints, Nvidia's earnings power would be materially lower than the bullish case implies.

But as of the current evidence, Nvidia is not showing signs of weak demand. It is showing signs of extraordinary demand, high margins, and large customer capex budgets.

Exhibit 5: Bull, Base, and Bear Interpretations

Scenario	Interpretation	What would support it

Bull case	Compute flywheel remains strong; AI labs keep scaling and inference demand explodes.	Frontier model jumps, agent adoption, high hyperscaler capex, strong Nvidia margins.
Base case	Growth decelerates but remains high; demand broadens beyond training.	Strong revenue growth but lower percentage growth; continued data-center investment.
Bear case	Demand was overbuilt or capability gains flatten; margins and capex roll over.	Weak model improvement, capex cuts, margin compression, failed enterprise ROI.

10. Counterarguments and Failure Modes

The bull case is strong, but it must be stress-tested. The following risks are the most important.

Scaling could weaken

If additional compute stops producing meaningful model improvements, the capability prong weakens. This is the most important technical risk. The entire arms-race psychology depends on the belief that more compute still improves capability. If that belief changes, the urgency of buying accelerators falls.

Agents could disappoint

The usage prong depends heavily on agents and workflow automation becoming reliable enough for real work. If agents remain slow, expensive, brittle, or hallucination-prone, usage intensity may not increase as much as expected. Chatbot usage alone is large, but the strongest demand case requires heavier workloads.

Enterprise ROI could lag

Enterprises may pilot AI tools but fail to deploy them broadly because of security, compliance, reliability, integration, governance, or cultural issues. If the technology does not produce measurable ROI, customers may slow spending.

Infrastructure could bottleneck growth

Power, grid connections, data-center construction, cooling, HBM, packaging, networking, and deployment speed can all cap the pace of growth. These constraints may delay revenue even if end demand remains strong.

Custom silicon could take more share

Hyperscalers have strong incentives to reduce dependence on Nvidia where possible. If internal accelerators become good enough for large categories of training and inference, Nvidia may lose share or face pricing pressure.

Margins could normalize

Nvidia's gross margins are exceptionally high. If competition rises, supply catches up, or customers gain bargaining power, margins could decline. EPS growth could slow even if revenue continues to rise.

Export controls and China risk remain material

China is both a potential upside market and a geopolitical risk. Restrictions can limit revenue, force product redesigns, and incentivize domestic alternatives. The long-term effect could be reduced access to one of the world's largest AI markets.

Financial-market psychology could change

Even if Nvidia executes, the stock can fall if investors decide growth has peaked, capex is excessive, or the multiple should compress. A great business can be a poor short-term stock if expectations are too high.

Circular financing concerns could rise

AI infrastructure involves complex partnerships, cloud commitments, equity investments, debt-funded projects, and long-term capacity contracts. If investors believe demand is being artificially supported by circular arrangements rather than end-customer economics, the market could apply a lower multiple.

These risks are real. The correct conclusion is not that Nvidia is risk-free. The correct conclusion is that the strongest version of the thesis survives only if both prongs continue: models improve and usage expands.

11. Valuation Lens: P/E at Cost and EPS Compounding

The P/E-at-cost framework is useful for a fast-growing company because it focuses on the relationship between the purchase price and future earnings power. A stock that looks expensive on current earnings can become cheap quickly if EPS compounds rapidly.

Assume a starting P/E of 20x. If EPS grows 95% in year one and 50% in year two, EPS increases by 2.925x. The effective P/E at the original purchase price falls to 6.84x. If EPS then grows 30% in year three, the total EPS multiplier becomes 3.8025x, and the effective P/E falls to 5.26x. At a starting P/E of 32x, the same three-year path produces an effective P/E of 8.42x.

This does not mean those growth rates are guaranteed. It means the valuation debate must be connected to the earnings path. For Nvidia, the stock's apparent expensiveness depends heavily on how quickly the earnings base resets upward.

The market often asks: What is the current multiple? The better long-term question is: What is the multiple on earnings that can reasonably be generated over the next two to three years?

This is especially important when a business is going through a step-change in demand. Nvidia is not merely growing from a small base. It is scaling from a very large base. That makes the absolute EPS impact unusually powerful.

However, P/E-at-cost can also create false comfort if the growth path fails. If EPS growth slows much faster than expected, or if margins compress, the effective multiple does not fall as quickly. If the market multiple also compresses, the stock can decline despite revenue growth. Therefore, the P/E-at-cost framework is useful but should not replace operational monitoring.

The key drivers of the EPS path are revenue growth, gross margins, operating expense growth, tax rate, share count, and buybacks. Nvidia's Q1 FY2027 results included an $80 billion additional share repurchase authorization and an increase in its quarterly dividend [1]. Buybacks can support EPS, but the main driver remains operating earnings.

For a long-term investor, the most important point is that a 20x cost-basis multiple is not necessarily demanding if the company can double or triple EPS over a few years. The thesis depends on whether the compute flywheel continues to translate into revenue and earnings.

Exhibit 6: P/E at Cost Under Selected EPS Growth Scenarios

EPS growth path	Cumulative EPS multiplier	P/E at cost if starting P/E = 20x	P/E at cost if starting P/E = 32x
95%, 50%	2.925x	6.84x	10.94x
95%, 50%, 20%	3.510x	5.70x	9.12x
95%, 50%, 30%	3.803x	5.26x	8.42x
90%, 60%, 60%	4.864x	4.11x	6.58x

12. What to Watch Over the Next 12 to 24 Months

The thesis should be monitored with technical, behavioral, operational, and financial indicators.

Frontier model progress

Watch whether the next generation of models shows clear improvements in reasoning, coding, tool use, multimodal understanding, long-context performance, and agentic behavior. If model releases begin to feel incremental rather than step-change, the capability prong weakens.

Agent reliability and adoption

The strongest usage case depends on agents moving from demos to production. Watch whether agents can complete real workflows, recover from errors, use tools safely, and deliver value without constant human correction.

Usage intensity

Do not watch only user counts. Watch usage per user, tokens processed, agent sessions, tool calls, multimodal requests, and enterprise workflow volume. Usage intensity is what drives inference demand.

Hyperscaler capex

Microsoft, Amazon, Alphabet, Meta, Oracle, and other large buyers are key demand indicators. Continued high capex supports the thesis. A broad and sustained capex pullback would be a warning.

Nvidia data-center revenue and guidance

Nvidia's Data Center revenue, sequential growth, and guidance are the most direct financial indicators. Watch not just annual growth but sequential momentum.

Gross margin

Margins reveal pricing power and supply-demand balance. Stable mid-70s gross margins suggest continued strength. Sharp margin compression would change the EPS story.

Supply-chain bottlenecks

Track HBM, advanced packaging, networking, optics, cooling, and system-level capacity. Bottlenecks can delay growth but also indicate demand is larger than supply.

Power and data-center deployment

Power availability may become the limiting reagent. Watch grid interconnections, energy deals, gas turbines, nuclear agreements, battery storage, permitting, and data-center construction timelines.

Custom silicon

Watch actual workload migration, not announcements. The key question is whether custom chips take material high-value share, especially in inference.

Enterprise ROI

The AI infrastructure thesis is strongest if enterprises move from pilots to production. Watch whether companies deploy AI into recurring workflows, not just demos.

China and export controls

China risk remains meaningful. Changes in export restrictions or Chinese regulatory approvals can affect revenue and strategic positioning.

The central monitoring question is simple: Are both prongs still working? If models keep improving and usage keeps expanding, compute demand should remain strong. If either prong weakens, the thesis needs to be revised.

13. Conclusion

The strongest AI compute thesis is not that Nvidia is simply a great chip company. It is that intelligence is becoming a compute-scaled product.

If that is true, then demand for AI infrastructure is driven by two reinforcing forces. The first is the need for more compute to create better models. The second is the need for more compute to serve the growing usage created by those better models. This is the two-pronged framework: capability growth and usage growth.

Traditional financial models may capture some of this, but they can miss the psychological and technical mechanism. Analysts can model growth deceleration because the base is large. That is mathematically reasonable. But the labs are not making decisions from a static growth-rate spreadsheet. They are making decisions in a frontier race where compute can determine whether they produce the next superior model.

The current evidence suggests the flywheel is still active. Nvidia's revenue and Data Center growth are enormous [1]. Hyperscalers are spending at unprecedented levels [4][6][7][8]. AI cloud backlog is large [9]. ChatGPT and Gemini usage is already massive [15][16]. Data-center energy demand is projected to rise sharply [10][11].

The risks are real. Scaling could weaken. Agents could disappoint. Enterprise ROI could lag. Power and data-center constraints could limit deployment. Custom silicon could pressure Nvidia's share and margins. Export controls could restrict markets. The stock can be volatile even if the business performs.

But the key question is not whether growth eventually slows. It will. The key question is whether the slowdown reflects saturation or simply a larger denominator while demand remains strong. If the compute flywheel continues, Nvidia's earnings power can keep compounding rapidly.

The concise conclusion is this: AI labs and hyperscalers are unlikely to voluntarily step off the gas while they continue to see meaningful model improvements and exploding usage. The real brakes are physical constraints, economics, and technical saturation - not a simple desire to spend less.

That is why the two-pronged framework is powerful. It explains demand in human terms as well as technical terms. Better hardware makes better models. Better models create more usage. More usage requires more compute. As long as that loop remains intact, AI infrastructure demand can remain larger and more durable than traditional financial models assume.

The purpose of the two-pronged framework is to keep the analysis grounded. It avoids reducing the Nvidia debate to a single P/E ratio or a single annual growth estimate. The core issue is whether capability and usage continue to compound. If they do, the infrastructure demand curve may remain larger and more durable than traditional models expect.

The most important qualitative question remains: are the people closest to the frontier still behaving as though more compute creates better products? If yes, the flywheel remains intact. If no, the thesis changes.

14. Extended Research Analysis: Why the Flywheel Matters

The initial sections set out the core claim: AI compute demand is not merely a one-time training boom. It is a feedback loop between capability and usage. This extended section expands the mechanics behind that loop and explains why the framework is useful for evaluating Nvidia, hyperscaler capex, model-lab behavior, and long-term inference demand.

The reason the framework matters is that it shifts the analysis from a static capacity model to a dynamic capability model. In a static model, customers buy servers to meet known workloads. In a dynamic capability model, customers buy compute partly because compute improves the product, and the improved product creates future workloads. That distinction changes the psychology of the buyer and the duration of the demand curve.

The most important implication is that near-term financial-model deceleration does not necessarily mean the demand curve is breaking. A business can decelerate from extraordinary growth rates while continuing to add extraordinary absolute dollars of revenue and earnings. This is particularly relevant to Nvidia because its revenue base has become so large that even lower percentage growth can still imply historically large absolute-dollar growth.

14.1 Efficiency Is Not Automatically Bearish

One of the most common bearish arguments is that AI models and inference systems will become more efficient, reducing the need for Nvidia hardware. This argument contains a real point but usually reaches the wrong conclusion too quickly. Efficiency can reduce compute per unit of work, but it can also expand the number of economically viable workloads. In many technology markets, efficiency expands total demand rather than shrinking it.

The internet did not use less bandwidth because compression and networking improved. Software did not consume less compute because CPUs became more efficient. Cloud usage did not shrink because virtualization improved utilization. In each case, lower cost and better performance unlocked new applications. AI may follow the same pattern. If inference becomes cheaper and faster, more developers can embed AI into products, more enterprises can deploy agents, and more consumers can use richer features.

DeepMind's Chinchilla work is a useful example of efficiency improving capability and downstream economics at the same time. It showed that better allocation of compute between model size and data could produce stronger performance for the same training compute budget, while facilitating downstream usage [14]. That kind of efficiency does not necessarily reduce the long-run need for compute. It can make AI more useful and more affordable, which can expand total usage.

The same logic applies to smaller models, routing systems, distillation, caching, speculative decoding, and custom inference optimizations. These techniques may lower cost per query. But if lower cost per query makes AI usable in search, customer support, coding, finance, legal, personal assistants, and robotics, the total number of queries and tasks can rise dramatically.

The important variable is elasticity of demand. If lower inference cost leads to much higher usage, total compute can rise. The more useful AI becomes, the more likely this demand elasticity is high. At the current stage of the market, AI is not saturated. Lower cost and better latency are likely to expand usage more than they reduce demand.

14.2 Training and Inference Are Separate Demand Pools

Training and inference should be analyzed separately. Training demand is driven by the frontier race. Inference demand is driven by usage. Both can be enormous, but they behave differently.

Training is lumpy. A lab builds a large cluster, trains a frontier model, runs post-training, evaluates, and then repeats with the next generation. The next training run may require more compute, but training demand often arrives in major waves tied to model cycles and hardware generations.

Inference is recurring. Every AI product call consumes compute. Every chatbot answer, search result, coding-agent step, spreadsheet analysis, document review, video generation, voice session, and robotic decision is inference. If AI becomes embedded in daily work and software, inference demand becomes continuous.

This is why the long-term demand story may shift from training to inference. Training creates the model, but inference monetizes and distributes the model. If a model reaches hundreds of millions of users, or if enterprise workflows begin running continuously, the recurring inference load can become very large.

The usage data from Google and OpenAI points in that direction. Google reported that Gemini app monthly active users surpassed 900 million and daily requests grew more than seven times year over year [15]. OpenAI reported 900 million weekly active ChatGPT users and 50 million paying subscribers [16]. Those numbers are already enormous, but they likely understate future demand if users move from short prompts to long-running agentic tasks.

The key inference question is not only how many users exist. It is how heavy each user's workload becomes. A short prompt and answer may consume little compute. A coding agent or finance agent may perform dozens of steps. A multimodal video workflow may be heavier still. A physical AI system may require continuous inference. Usage intensity matters as much as user count.

14.3 Agents Turn Prompts Into Workflows

Agents are the most important bridge between chatbot usage and industrial-scale inference. A chatbot answers. An agent works. That difference changes the compute profile.

A simple chatbot request might be one model call. An agentic workflow may involve planning, retrieval, file reading, tool calls, code generation, test execution, error handling, retries, summarization, and final output. One user instruction can become dozens or hundreds of model interactions. This means agents can multiply inference demand even without proportional user growth.

The difference is easy to see in enterprise workflows. A user asking an AI to summarize an account variance is one workload. A user asking an agent to pull the general ledger, identify drivers, reconcile supporting files, generate commentary, update a workbook, and flag exceptions is a much heavier workload. If the latter becomes routine, compute demand becomes part of the operating process.

Google's public framing around an agentic Gemini era shows that the largest AI platforms are explicitly moving beyond passive chat toward action-oriented systems [15]. The relevance for Nvidia is direct: agentic systems often need lower latency, higher throughput, more context, more tool use, and repeated model calls. Those are infrastructure demands.

Agents also change user psychology. If an agent saves real time, the user does not view AI as a toy. The user begins delegating work. Once delegation begins, usage can compound. People ask for more. Teams build workflows. Businesses integrate agents into internal systems. The model becomes part of the work fabric.

14.4 Hyperscaler Incentives Are Strategic, Not Merely Financial

The hyperscaler capex cycle should be interpreted through strategic incentives. Microsoft, Amazon, Alphabet, and Meta are not simply buying servers because current demand requires a specific number of units. They are building infrastructure for a platform transition.

Microsoft's disclosure that roughly two-thirds of Q3 FY2026 capex went to short-lived assets primarily GPUs and CPUs is important because it shows AI compute intensity directly [4]. The earlier Q2 comment that customer demand continued to exceed supply reinforces the point that capacity constraints remained relevant [5]. Amazon's $200 billion 2026 capex projection, Alphabet's $175

billion to $185 billion guide, and Meta's $125 billion to $145 billion guide all indicate that the largest technology companies are prioritizing AI infrastructure despite investor scrutiny [6][7][8].

The strategic logic is clear. Cloud platforms need AI capacity for internal products, external customers, and model partners. They also need to avoid losing workloads to competitors. If customers want AI compute and one cloud cannot provide it, the customer can shift to another cloud or specialized AI cloud. Capacity becomes a competitive feature.

The same logic applies to first-party products. Search, office productivity, coding tools, assistants, advertising, social feeds, recommendations, and cloud services can all become AI-enhanced. If AI becomes central to user experience, then infrastructure scarcity constrains product quality and revenue opportunities.

This is why hyperscalers may tolerate near-term pressure on free cash flow. They are not buying optional experimental equipment. They are securing infrastructure for what they believe may be the next major computing platform. In that context, underbuilding can be more dangerous than overbuilding.

14.5 Enterprise Adoption Is Slow Until It Is Not

Enterprise adoption often appears slow because large companies require security review, governance, data controls, integration, training, and workflow redesign. But once a use case proves valuable, adoption can accelerate. The adoption path is often nonlinear.

The pattern is common: a prototype solves a painful workflow, a small team starts using it, the time savings become obvious, adjacent teams ask for access, and the organization begins treating the tool as infrastructure. At that point the AI system is no longer a demo. It is part of operating leverage.

This matters because the most durable AI usage may be boring. Finance agents, accounting research agents, transaction classifiers, invoice pipelines, legal review tools, customer-support triage systems, compliance assistants, and internal data-analysis agents may not produce viral demos. But if they run every day across thousands of companies, they can create enormous recurring inference demand.

Enterprise AI demand also compounds with trust. Early systems are used cautiously. As reliability improves, users delegate more important tasks. As integration improves, agents can access more systems. As security controls improve, more sensitive workflows become eligible. Each improvement increases potential usage intensity.

The two-pronged framework explains this progression. Capability improvements increase reliability and usefulness. Usage follows because the system can now handle more valuable work. More usage then increases inference demand and justifies more infrastructure.

14.6 Physical AI Could Expand the Addressable Compute Market

Physical AI is a longer-duration opportunity, but it may be one of the largest. Digital workflows are already large, but the physical world is larger. Factories, warehouses, vehicles, hospitals, agriculture, logistics, construction, defense, energy, and homes all contain tasks that could eventually be assisted or automated by AI systems.

Physical AI creates compute demand at multiple levels. Training and simulation require large-scale compute to build world models, robotics policies, reinforcement learning environments, and synthetic data. Deployment requires edge inference, sensor processing, planning, and control. A robot or autonomous system does not simply answer one prompt. It continuously perceives and acts.

This is important for Nvidia because the company is positioning itself around AI factories, robotics, edge computing, automotive systems, and industrial AI, not just cloud GPUs [1][20]. If physical AI becomes a major category, the compute market broadens from digital applications into machines and infrastructure.

The timeline is uncertain. Robotics is hard. Safety, hardware cost, reliability, and deployment complexity are significant barriers. But even partial success can add another layer to demand. The market does not require humanoid robots in every home for physical AI to matter. Warehouses, factories, vehicles, and industrial inspection can be meaningful categories by themselves.

14.7 Power Is Not a Side Issue

Power is one of the central variables in the AI compute thesis. Chips, memory, and systems matter, but none of them can operate without reliable electricity and data-center infrastructure.

The IEA's projection that data-center electricity consumption could roughly double from 485 TWh in 2025 to 950 TWh in 2030 shows the scale of the issue [10]. AI-focused data centers are expected to grow faster than the data-center category as a whole. This means AI demand is increasingly tied to energy markets, grid planning, permitting, cooling, and construction.

Power constraints can create a paradox. They can limit near-term revenue because customers cannot deploy as quickly as they want. But they can also indicate that demand is larger than supply. If energized data-center capacity becomes scarce, customers may compete aggressively for it. That can benefit suppliers with scarce systems and trusted deployment architectures.

For investors and analysts, power should be treated as a leading indicator. Watch grid interconnection queues, power purchase agreements, gas turbine orders, nuclear and geothermal deals, battery storage, cooling systems, and data-center construction announcements. These are no longer peripheral infrastructure details. They are part of the AI compute supply chain.

14.8 Custom Silicon Is a Real Risk, But Not a Simple Bear Case

Custom silicon is one of the strongest counterarguments to Nvidia's dominance. Google, Amazon, Microsoft, Meta, and others have strong incentives to develop internal accelerators. If a hyperscaler can serve a large stable workload at lower cost using its own chip, it will try.

However, custom silicon is not automatically fatal to Nvidia. The first reason is market expansion. The total compute market may grow fast enough that Nvidia can lose share in some categories while still growing revenue. The second reason is workload diversity. Frontier training, generalized cloud demand, enterprise adoption, AI clouds, research workloads, and fast-moving model architectures may continue to favor Nvidia's flexible ecosystem.

The third reason is software. Hardware specifications are not enough. Developers need libraries, tools, compatibility, reliability, and operational maturity. Nvidia's CUDA ecosystem and system-level

integration remain major advantages. At cluster scale, the customer is buying a working system, not only a chip.

The real question is not whether custom silicon exists. It does. The real question is how much high-value workload share migrates away from Nvidia, how quickly it happens, and whether it pressures margins. That is what should be monitored.

14.9 The Cleanest Way to Falsify the Thesis

A strong thesis should be falsifiable. The compute flywheel thesis weakens if either prong breaks.

The capability prong breaks if frontier models stop improving materially with more compute. Signs would include model launches that feel incremental, benchmarks that plateau, weak improvements in coding and reasoning, reduced willingness by labs to fund larger training runs, or public comments suggesting that compute scaling is no longer a high-return path.

The usage prong breaks if better models fail to translate into heavier usage. Signs would include stagnant user growth, declining usage intensity, weak enterprise adoption, disappointing agent reliability, poor customer ROI, or price cuts that fail to stimulate demand.

The Nvidia-specific thesis weakens if demand remains high but Nvidia loses too much economics. Signs would include sustained gross margin compression, large hyperscaler migration to custom silicon, pricing concessions, slower Data Center growth, or order cancellations rather than deployment delays.

Until those indicators appear clearly, it is premature to assume that deceleration equals a broken cycle. Deceleration may simply reflect the law of large numbers.

14.10 Practical Research Checklist

A quarterly monitoring checklist should include: Nvidia Data Center revenue and sequential growth; gross margin; customer concentration commentary; supply-chain bottleneck commentary; Blackwell and Rubin ramp updates; hyperscaler capex guides; AI cloud backlog; model-release quality; usage metrics from major AI products; enterprise agent adoption; data-center power availability; HBM supply; custom silicon announcements and actual workload migration; China and export-control developments; and any signs of customer ROI concern.

Appendix A: Source Notes

The following sources are cited in the report. URLs are included for transparency and follow-up review. Access date: May 27, 2026.

[1] Nvidia, "NVIDIA Announces Financial Results for First Quarter Fiscal 2027," May 20, 2026.
https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-first-quarter-fiscal-2027

[2] Nvidia, "NVIDIA Announces Financial Results for Fourth Quarter and Fiscal 2026," Feb. 25, 2026.
https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2026

[3] The Motley Fool / Nvidia transcript, "Nvidia Q1 2027 Earnings Call Transcript," May 20, 2026.
https://www.fool.com/earnings/call-transcripts/2026/05/20/nvidia-nvda-q1-2027-earnings-transcript/

[4] Microsoft, "Fiscal Year 2026 Third Quarter Earnings Conference Call," Apr. 30, 2026. https://www.microsoft.com/en-us/investor/events/fy-2026/earnings-fy-2026-q3

[5] Microsoft, "Fiscal Year 2026 Second Quarter Earnings Conference Call," Jan. 28, 2026.
https://www.microsoft.com/en-us/investor/events/fy-2026/earnings-fy-2026-q2

[6] Reuters, "Amazon projects $200 billion capital spending in 2026," Feb. 5, 2026.
https://www.reuters.com/business/retail-consumer/amazon-projects-200-billion-capital-spending-this-year-2026-02-05/

[7] Reuters, "Alphabet forecasts sharp surge in 2026 capital spending," Feb. 4, 2026.
https://www.reuters.com/business/google-parent-alphabet-forecasts-sharp-surge-2026-capital-spending-2026-02-04/

[8] Reuters, "Meta lifts capital expenditure forecast, doubling down on AI push," Apr. 29, 2026.
https://www.reuters.com/business/meta-lifts-capital-expenditure-forecast-doubling-down-ai-push-2026-04-29/

[9] CoreWeave, "CoreWeave Reports Strong First Quarter 2026 Results," May 7, 2026.
https://investors.coreweave.com/news/news-details/2026/CoreWeave-Reports-Strong-First-Quarter-2026-Results/

[10] IEA, "Key Questions on Energy and AI - Executive Summary," 2026. https://www.iea.org/reports/key-questions-on-energy-and-ai/executive-summary

[11] IEA, "Energy Demand from AI," 2026. https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

[12] Epoch AI, "Trends in Artificial Intelligence," 2026. https://epoch.ai/trends

[13] OpenAI / arXiv, "Scaling Laws for Neural Language Models," 2020. https://arxiv.org/abs/2001.08361

[14] DeepMind / arXiv, "Training Compute-Optimal Large Language Models," 2022. https://arxiv.org/abs/2203.15556

[15] Google, "I/O 2026: Welcome to the agentic Gemini era," May 19, 2026. https://blog.google/innovation-and-ai/sundar-pichai-io-2026/

[16] TechCrunch, "ChatGPT reaches 900M weekly active users," Feb. 27, 2026.
https://techcrunch.com/2026/02/27/chatgpt-reaches-900m-weekly-active-users/

[17] OpenAI, "Announcing The Stargate Project," Jan. 21, 2025. https://openai.com/index/announcing-the-stargate-project/

[18] Reuters, "OpenAI, Oracle, SoftBank plan five new AI data centers for $500 billion Stargate project," Sep. 23, 2025.
https://www.reuters.com/business/media-telecom/openai-oracle-softbank-plan-five-new-ai-data-centers-500-billion-stargate-2025-09-23/

[19] S&P Global Market Intelligence, "Nvidia earnings preview: Q1 2027," May 14, 2026.
https://www.spglobal.com/market-intelligence/en/news-insights/research/2026/05/nvidia-earnings-preview-q1-2027

[20] Reuters, "Nvidia bets on new data center chips for growth as sales outlook tops estimates," May 20, 2026.
https://www.reuters.com/business/retail-consumer/nvidia-forecasts-quarterly-revenue-above-estimates-announces-80-billion-share-2026-05-20/

Appendix B: Scenario Math

The formula for P/E compression from EPS growth is straightforward:

New P/E = Starting P/E / cumulative EPS multiplier.

If EPS grows 95% in year one, the year-one multiplier is 1.95. If EPS grows 50% in year two, the year-two multiplier is 1.50. The two-year multiplier is 1.95 x 1.50 = 2.925. A 20x starting P/E becomes 20 / 2.925 = 6.84x. A 32x starting P/E becomes 32 / 2.925 = 10.94x.

If EPS then grows 30% in year three, the cumulative multiplier is 1.95 x 1.50 x 1.30 = 3.8025. A 20x starting P/E becomes 5.26x. A 32x starting P/E becomes 8.42x.

If EPS grows 90%, 60%, and 60% over three years, the multiplier is 1.90 x 1.60 x 1.60 = 4.864. A 20x starting P/E becomes 4.11x. A 32x starting P/E becomes 6.58x.

These scenarios are not forecasts. They are sensitivity analysis. They show why earnings growth matters so much when evaluating a business growing through a major demand inflection.

Exhibit B1: Scenario Math Detail

Year 1 growth	Year 2 growth	Year 3 growth	EPS multiplier	20x start	32x start
95%	50%	-	2.925x	6.84x	10.94x
95%	50%	20%	3.510x	5.70x	9.12x
95%	50%	30%	3.803x	5.26x	8.42x
90%	60%	60%	4.864x	4.11x	6.58x