

Will AI Costs, Supply Constraints, and Competition Prevent Attractive Long-Term Returns for Hyperscalers?

By Adam Bernstein, Investment Analyst

In [our last article](#), we concluded that AI monetization is becoming increasingly tangible and that revenues could close the gap with infrastructure investment in the next 5-8 years. That conclusion, however, assumes the companies currently making the investments: the hyperscalers,¹ can convert those revenues into durable profits. In this piece, we examine whether costs, supply chain challenges, and competition could prevent revenue growth from translating into attractive long-term returns for investors.

The context

While revenues are materializing, and more rapidly than with other technological revolutions² such that they could justify infrastructure investment, revenue alone does not determine shareholder returns. The remaining question is whether they can grow at attractive enough margins to keep pace with expanding capital commitments by the hyperscalers.

David Cahn of Sequoia Capital highlighted the scale of the monetization challenge as AI investment accelerated. In 2023, he estimated that roughly \$200 billion of annual AI revenue was needed to support that year's infrastructure spending, leaving an annual revenue shortfall of approximately \$125 billion.³ As infrastructure investment expanded, Cahn updated that estimate in 2024, concluding that approximately \$600 billion of annual AI revenue would ultimately be required.⁴ Approaching the question from a different perspective, JPMorgan estimated in early 2026 that approximately \$650 billion of recurring annual AI revenue would be needed to generate a 10% return on more than \$5 trillion of cumulative AI infrastructure investment expected through 2030.⁵

As we discussed in [our last article](#), Exponential View estimates the global AI ecosystem is already generating approximately \$175 billion of annualized revenue. More importantly, its analysis suggests the industry has reached an important milestone: AI revenues now exceed the ongoing depreciation associated with newly deployed infrastructure, indicating that monetization has progressed far faster than previous technology cycles.⁶ However, revenues remain well below the level required to generate attractive returns on the cumulative capital already invested, particularly as infrastructure spending continues to expand.

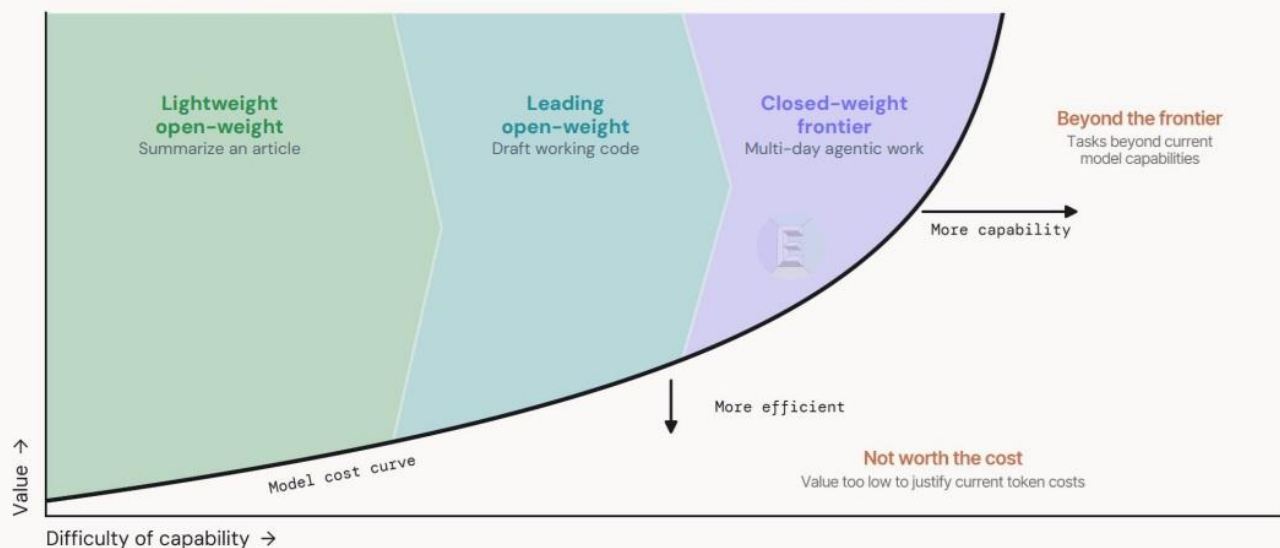
Technology writer and family office investor [Kevin Zhang](#) argues that traditional revenue-gap analyses may understate the monetization challenge because they do not fully account for



continued infrastructure investment, depreciation, margin requirements, and the fact that only a portion of end-user AI revenues ultimately accrues to the companies funding the AI buildout. Together, these factors suggest the revenue required to generate attractive long-term returns could be materially higher than simpler capex-versus-revenue comparisons imply.⁷

This gap could close in the coming years depending on how fast spending grows. Thus far, planned infrastructure investment has expanded almost as rapidly as revenue growth, implying that monetization must continue accelerating to justify today's capital commitments, although some reported AI-related revenues may not fully reflect incremental end-user demand. In pursuit of this, companies may have to evolve their business models. "Many leading models lose money on each interaction, and while that is not unusual for a new service, it is unclear how long it can persist. Subscription revenue has dominated so far, but advertising and alternative pricing models are likely to emerge."⁸

Labs must outrun open-weight commoditization to hold margin



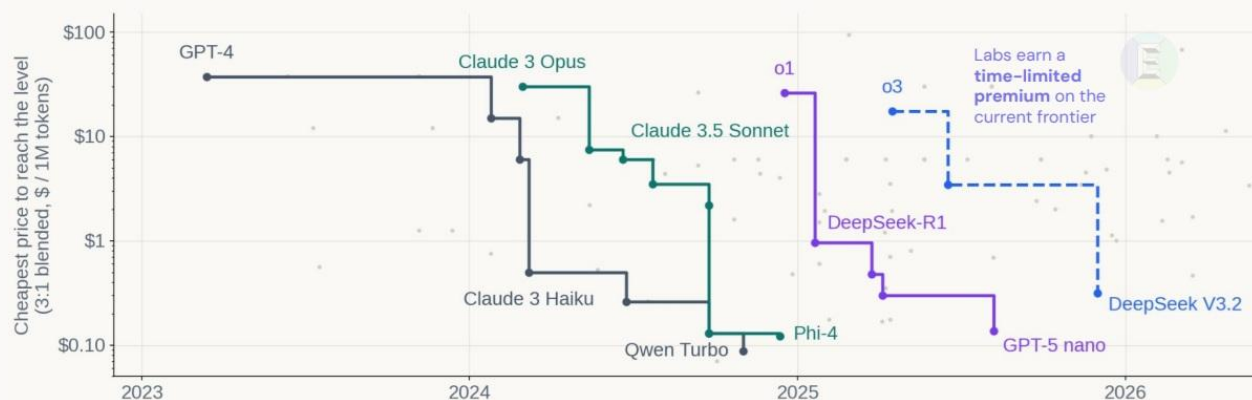
Source: [Exponential View](#)



Last year's frontier is commoditizing fast

Price per capability frontier

Blended price \$, grouped by performance at GPQA Diamond (PhD-level science)



Sources: Exponential View analysis; Epoch AI.

59

Source: [Exponential View](#)

Who captures the economics?

Even if AI creates trillions of dollars of value for the overall economy, it does not necessarily follow that the companies currently funding the infrastructure will capture proportionate returns. Investors must therefore distinguish between the value AI creates for the economy and the value ultimately captured by shareholders, and by which particular companies.

Cloud computing provides a useful framework. Amazon Web Services (AWS) demonstrated that renting compute, storage and networking capacity to enterprises has become one of the most profitable businesses ever built. For context, AWS generated nearly \$130 billion in annual revenue with operating margins exceeding 35% in 2025, with annualized revenue rising to \$150 billion and the operating margin increasing to 37.7% for the first quarter of 2026.⁹ At the same time, many of cloud's highest-value businesses did not own the underlying infrastructure. Companies such as Salesforce, ServiceNow and Snowflake leased compute from cloud providers and differentiated themselves through proprietary software, customer relationships, enterprise workflows, and unique datasets. They monetized what ran on the cloud rather than the cloud itself.

AI may evolve differently. Research from Exponential View estimates that approximately 82% of generative AI revenue currently accrues to hosting services, while foundation model providers¹⁰ account for only 11%.¹¹ Unlike the early cloud era, today's hyperscalers increasingly control



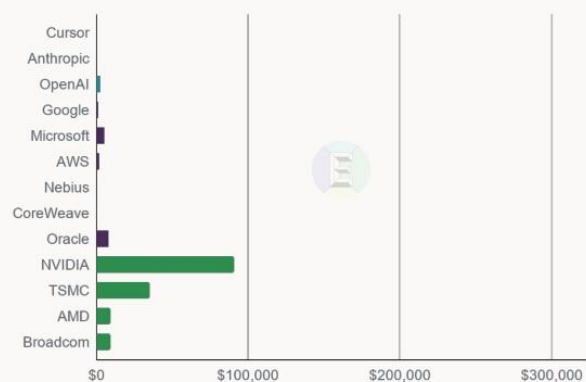
multiple layers of the value chain. They simultaneously own the infrastructure, operate many of the leading foundation models, distribute AI through enterprise software, and maintain direct customer relationships. Vertical integration gives them a stronger claim on future economics than infrastructure providers enjoyed during the cloud transition.

The question for investors is whether that advantage proves durable. Exponential View outlines several plausible industry outcomes, many of which assume that an increasing share of AI's economic value accrues to end users and application providers rather than the hyperscalers. As open-source models improve and the cost of serving AI responses continues to fall, "good enough" intelligence may become sufficient for many enterprise workloads. Maintaining a frontier model could therefore become increasingly expensive relative to the value it captures. In that scenario, greater value could accrue to businesses building differentiated applications, both internally and for external customers. Conversely, if the hyperscalers continue to leverage their vertically integrated ecosystems, combining infrastructure, models, software, and distribution, they may be capable of earning economics similar to those achieved by mature cloud platforms, which have demonstrated durable, high operating margins over time.

Revenue is concentrated today, but the mix is shifting

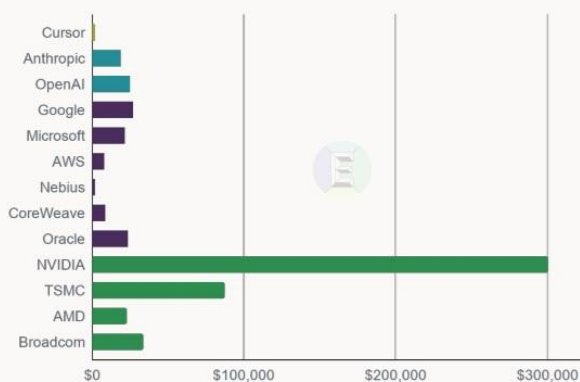
Annualized GenAI revenue

\$million, Q1 2024



Annualized GenAI revenue

\$million, Q1 2026



Apps Foundation Models Hosting Chips

Sources: Exponential View analysis; company filings.
Note: Revenues not subject to deduplication adjustment.



Source: [Exponential View](#)

How constrained are supply chains?

Returning to [Richard Sutton](#), seven decades of research showed “that general methods that leverage computation are ultimately the most effective, and by a large margin.”¹² It is less about better code or integrating more human knowledge. But what happens if adequate compute is not available?

Compute refers to the computational power necessary for training and running AI models and relies on hardware like GPUs and networking equipment as well as the power generation and data centers that house them, and therefore multiple supply chains. Supply chain challenges could affect the entire industry’s ability to convert demand into revenues but could also shake out the winners from the losers based upon which companies have the most robust supply chain management.

There have already been a number of challenges to date, for example in relation to GPUs. In February 2024, Nvidia cautioned investors to “Expect next-generation products to be supply-constrained as demand far exceeds supply.”¹³ In May 2024, the company stated a similar message with reference to its H200 GPU and Blackwell products whereby “demand may exceed supply well into next year.”¹⁴ By June 2026, there is some easing as Jensen Huang stated, “We have supply for very, very robust growth.” However, he noted that the company is “still supply constrained.”¹⁵ Scarcity of inputs creates pricing power, higher margins, and multiple expansion for the firms providing the GPUs and other critical hardware. Therefore, investors focusing on the bottlenecks in the AI ecosystem have benefited.¹⁶

As of December 2025, David Cahn of Sequoia Capital saw risks in the data center space where delays could be caused by industrial companies that are not able to provide the power generation and cooling systems on time. He also notes that skilled labor shortages could lead to construction timelines lengthening.¹⁷

How are hardware costs showing up in financial reports?

In our earlier piece, we looked at earnings quality, focusing on the revenue side. On the cost side, there are concerns that depreciation schedules used in financial accounting may not be reflective of how hardware will actually be used and the long-term expenses. Tech companies are implementing longer depreciation schedules for chips, servers, and other equipment, which leads to an increase in reported earnings given the reduction in annual depreciation. According to the

Wall Street Journal, “Meta said the latest extension reduced its depreciation expense by \$2.3 billion for the first nine months of 2025.”¹⁸

Meanwhile, the actual economic life of AI hardware is uncertain given the speed of this technology cycle. In addition, accelerated depreciation may be more appropriate for certain assets like chips that lose significant value in their initial years of use.¹⁹ Depreciation for traditional data center hardware cannot be extrapolated to AI data center hardware: “the speed at which Nvidia releases new architectures means that hardware performance has been increasing significantly relative to that of CPUs. Since 2020, Nvidia has released a major new architecture every ~2 years.”²⁰ Conversely, if there are instances where the useful lives of semiconductor chips can be extended beyond the standard six-year depreciation schedule they could earn yields well above their depreciation charges.

While the accounting data may be more difficult to interpret, tech companies are not the first (and will not be the last) sector to pick the best ways to present their accounting. While the accounting data may be more difficult to interpret today, we do not believe this will materially affect the long-term investment thesis. As AI revenues become a more meaningful share of overall business performance, investors will increasingly evaluate companies based on reported revenue growth, margins, cash flows and returns on invested capital rather than management's presentation of individual AI-related accounting items.

So, what is the outlook?

In summary, our view is that costs and supply chains remain important risks, but that they represent execution challenges that may result in delays rather than existential threats. Inference costs have declined following Moore’s Law, meanwhile the overall spending on compute is still increasing proving that demand so far appears elastic (as prices fall, usage increases faster). As AI is used more extensively across functions and tasks, aggregate compute requirements, and therefore total costs, are rising even as chips become more powerful and efficient.

Hyperscalers are bearing the majority of the costs of compute and passing through only some to customers. Jim Schneider of Goldman Sachs believes that as agentic AI use grows in the corporate world, “the margin inflection for hyperscalers and model providers is very different from the prevailing market narrative that AI usage will simply drive an increasing and unsustainable cost burden. The evolution is going to be uneven and a bit non-linear; not all the players are at the same level. You are going to start to see differentiation between the hyperscalers, especially when it

comes to their operating cash flows. Every player will get dragged along to the upside but at different rates.”²¹

The remaining debate centers on margins, capital efficiency, and the ultimate distribution of profits throughout the AI ecosystem. It is unclear how much AI use will stay within hyperscaler companies. As end users increase their deployment of open-source or self-hosted models, or other providers, it may limit the revenue that can be earned by those dominating the current capex spend. That said, the sheer size and scale of the hyperscalers and their market reach certainly suggests that they are well-positioned to capture massive revenues into the future.

Overall Conclusion

As an allocator this is a challenging time, there are many uncertainties with many different possible outcomes, especially between companies. For advisors, investing new client assets into strategies with significant exposure to highly valued companies can be scary. The tailwind of liquidity has dissipated, interest rates remain at an elevated level (with some expectations that they could move higher), and the hyperscalers are now pursuing significant debt and equity issuance rather than share buybacks.

However, we see more of the challenges manifesting primarily within the passive S&P 500 ETF landscape than in our own models, which we believe are more insulated. Throughout the rise of passive investing, the markets have not experienced a sharp reduction in share buybacks nor multiple unprofitable companies going to IPO. Investors are beginning to demonstrate more caution, and we have seen a reduction in hyperscaler valuations as a result.

We might see a correction given the shift from share buybacks to debt issuance and given the lack of excitement about hyperscaler earnings and their quality. But we do not believe this is a long-term issue. Infrastructure build out necessitates upfront costs before revenues can be generated. We see revenues growing to fill the gap, supply chain issues being remediated, and innovation being deployed. Therefore, we believe that the earnings gap is a temporary challenge that should resolve over time. As we stated in our first piece, we recommend that investors retain a market-weight allocation to the broader AI infrastructure ecosystem, including hyperscalers, semiconductor companies, networking providers, and power infrastructure beneficiaries, complemented by other growth drivers. This aligns with our own allocation.

As with any major innovation, there will be winners and losers and some bumps in the road. Advisors and their clients should therefore think about alpha diversifiers beyond passive strategies,

and beyond the obvious AI companies. Our thematic SMAs and Climate Alpha UMA can be added alongside existing allocations to help broaden exposure to other parts of the market that can benefit from AI's growth while including other long-term structural themes.

Disclosures

This document does not constitute advice or a recommendation or offer to sell or a solicitation to deal in any security or financial product. It is provided for information purposes only and the understanding that the recipient has sufficient knowledge and experience to be able to understand and make their own evaluation of the proposals and services described herein, any risks associated therewith and any related legal, tax, accounting or other material considerations. To the extent that the reader has any questions regarding the applicability of any specific issue discussed above to their specific portfolio or situation, prospective investors are encouraged to contact Gitterman Asset Management or consult with the professional advisor of their choosing.

Certain information contained herein has been obtained from third party sources and such information has not been independently verified by Gitterman Asset Management. No representation, warranty, or undertaking, expressed or implied, is given to the accuracy or completeness of such information by Gitterman Asset Management or any other person. While such sources are believed to be reliable, Gitterman Asset Management does not assume any responsibility for the accuracy or completeness of such information. Gitterman Asset Management does not undertake any obligation to update the information contained herein as of any future date.

This presentation is confidential and is intended only for the person to whom it has been directly provided.

Certain information contained herein constitutes "forward-looking statements," which can be identified by the use of forward-looking terminology such as "may," "will," "should," "expect," "anticipate," "project," "estimate," "intend," "continue," or "believe," or the negatives thereof or other variations thereon or comparable terminology. Due to various risks and uncertainties, actual events, results, or actual performance may differ materially from those reflected or contemplated in such forward-looking statements. Nothing contained herein may be relied upon as a guarantee, promise, assurance or a representation as to the future.

Gitterman Asset Management is a dba of Perigon Wealth Management, LLC ("Perigon"), a registered investment advisor registered with the U.S. Securities and Exchange Commission.

¹ Defined primarily as Microsoft, Amazon, Alphabet, Meta. Also includes Oracle, Coreweave and Nebius where specifically stated.

² "PART I - Can hyperscalers generate enough revenue to justify their CAPEX investments?" Gitterman Asset Management

³ "AI's \$200B Question", Cahn, David, September 20, 2023, Sequoia Capital, <https://sequoiacap.com/article/follow-the-gpus-perspective/>

⁴ "AI's \$600B Question", Cahn, David, June 20, 2024, Sequoia Capital, <https://sequoiacap.com/article/ais-600b-question/>



-
- ⁵ “How is AI being monetized?”, Aliaga, Stephanie, January 2, 2026, J.P. Morgan Asset Management, <https://am.ipmorgan.com/us/en/asset-management/adv/insights/market-insights/market-updates/on-the-minds-of-investors/how-is-ai-being-monetized/>
- ⁶ “The State of the AI Economy”, June 25, 2026, Exponential View, <https://intelligence.exponentialview.co/>
- ⁷ “Power Overwhelming: dissecting the \$1.5T AI revenue shortfall”, Zhang, Kevin, December 4, 2025, LessWrong, <https://www.lesswrong.com/posts/XZL9TqptvKsDtWHbi/power-overwhelming-dissecting-the-usd1-5t-ai-revenue>
- ⁸ “How is AI being monetized?”, Aliaga, Stephanie, January 2, 2026, JP Morgan, <https://am.ipmorgan.com/us/en/asset-management/adv/insights/market-insights/market-updates/on-the-minds-of-investors/how-is-ai-being-monetized/>
- ⁹ https://www.sec.gov/Archives/edgar/data/1018724/000101872426000002/amzn-20251231xex991.htm?utm_source
- ¹⁰ General-purpose AI models trained for use across many applications. Some general-purpose models can become frontier models based on performance metric.
- ¹¹ “The State of the AI Economy”, June 25, 2026, Exponential View, <https://intelligence.exponentialview.co/>
- ¹² “The Bitter Lesson”, Sutton, Rich, March 13, 2019, Available at: https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf
- ¹³ Investor Presentation: Q4 FY24, Nvidia, February 26, 2024
- ¹⁴ Investor Presentation: Q1 FY 2025, Nvidia, May 28, 2024
- ¹⁵ “Nvidia has capacity to supply robust AI growth despite constraints, says CEO”, Lee, Wen-Yee & Cherney, Max A., June 1, 2026, Reuters, <https://www.reuters.com/world/asia-pacific/nvidia-ceo-says-has-capacity-supply-robust-cpu-gpu-growth-2026-06-02/>
- ¹⁶ <https://am.ipmorgan.com/us/en/asset-management/adv/insights/market-insights/market-updates/on-the-minds-of-investors/how-is-ai-being-monetized/>
- ¹⁷ “AI in 2026: A Tale of Two AIs”, Cah, David, December 3, 2025, Sequoia Capital, <https://sequoiacap.com/article/ai-in-2026-the-tale-of-two-ais/>
- ¹⁸ “The Accounting Uproar Over How Fast an AI Chip Depreciates,” December 8, 2025, Weil, Jonathan, Wall Street Journal, <https://www.wsj.com/finance/investing/the-accounting-uproar-over-how-fast-an-ai-chip-depreciates-6f59785b>
- ¹⁹ “The Accounting Uproar Over How Fast an AI Chip Depreciates”, December 8, 2025, Weil, Jonathan, Wall Street Journal, <https://www.wsj.com/finance/investing/the-accounting-uproar-over-how-fast-an-ai-chip-depreciates-6f59785b>
- ²⁰ “Power Overwhelming: dissecting the \$1.5T AI revenue shortfall”, Zhang, Kevin, December 4, 2025, LessWrong, <https://www.lesswrong.com/posts/XZL9TqptvKsDtWHbi/power-overwhelming-dissecting-the-usd1-5t-ai-revenue>
- ²¹ “AI Agents Forecast to Boost Tech Cash Flow as Usage Soars”, May 20, 2026, Goldman Sachs, <https://www.goldmansachs.com/insights/articles/ai-agents-forecast-to-boost-tech-cash-flow-as-usage-soars>