BostonPartners

SEPTEMBER 2026

AI as an investment cycle: what needs to go right

At a certain point, the current artificial intelligence built out will need to demonstrate a return on investment in order to justify its pace and magnitude. We take a closer look at what needs to go right for the economics behind this transformative technology to work.

The future of artificial intelligence (AI) is often framed in extremes. On one end is a vision of unprecedented productivity, scientific discovery, and economic growth. On the other is disruption: job displacement, misinformation, and systemic risk. The reality, as is often the case with transformative technologies, likely lies somewhere in between.

For investors, however, the question is not philosophical—it's practical. Markets are being asked to underwrite one of the largest capital allocation cycles in modern history, with the expectation that future revenue will justify today's spending. The key question, then, is simple: What needs to be true for the economics of AI to work?

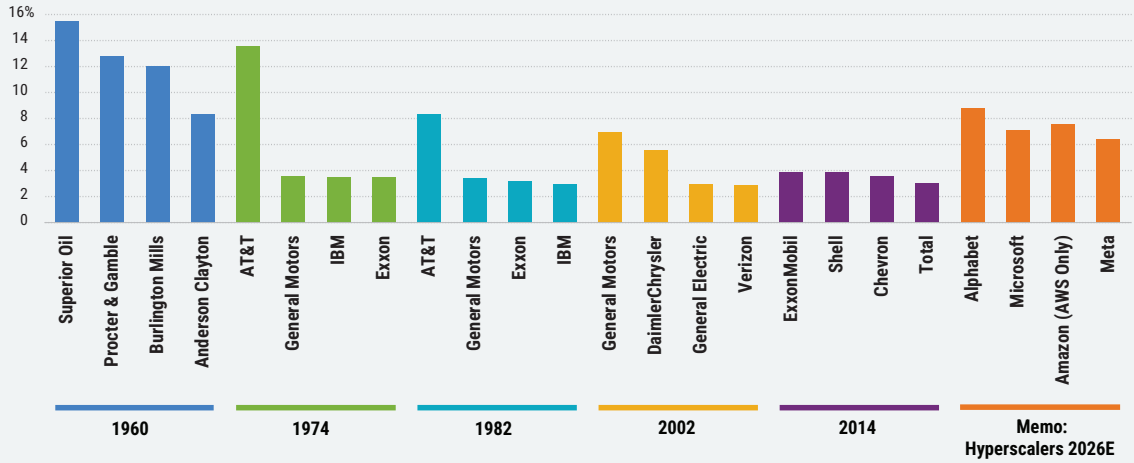
The scale of the bet

AI is not just another technology upgrade; it's a full-scale infrastructure buildout. Hyperscalers are expected to spend more than $5 trillion on capital expenditures (CapEx) between 2024 and 2030, a level of investment that rivals some of the largest industrial expansions in history.

This spending is also unusually concentrated. A handful of companies could account for nearly half of all S&P 500 CapEx by the end of the decade, while roughly two-thirds of global AI compute capacity is already controlled by just five players (Alphabet, Amazon, Meta, Microsoft, and Oracle).

Historical five-year capital spend

Hyperscalers' relative historical spend



Source: Empirical Research Partners LLC. Data as of May 12, 2026. Peaks in Capital Spending Concentration as of 1960, 1974, 1982, 2002, 2014 and 2026E (estimated). Past performance is not an indication of future results.

In certain ways, this resembles the early stages of prior infrastructure cycles—railroads and, more recently, fiber optic cable and cloud computing—where capacity was built ahead of fully realized demand. The difference this time is scale. AI has dramatically compressed the timeline for investment, creating a "build first, monetize later" dynamic that represents an extraordinary leap of faith regarding the pace and trajectory of future AI adoption and monetization.

The economic challenge

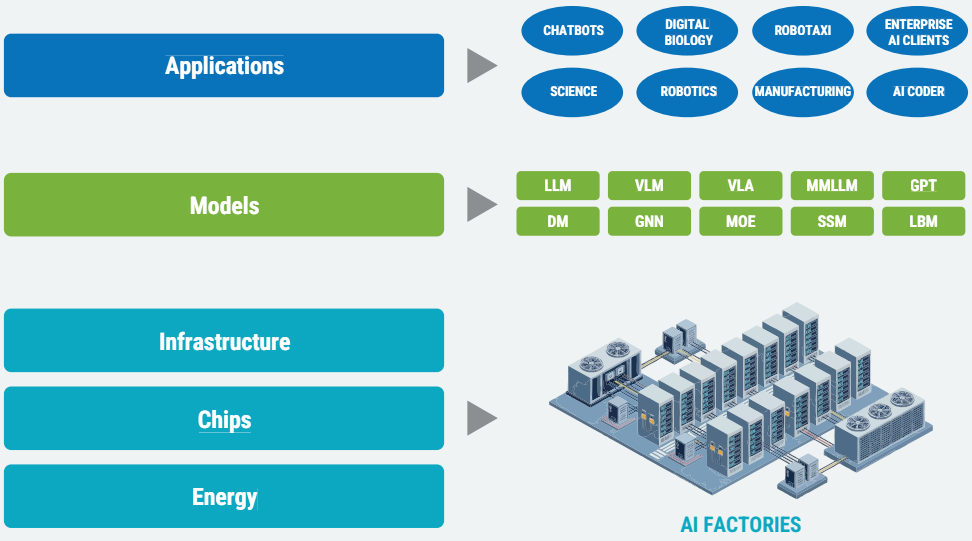
The central risk in AI is not whether the technology works; it clearly does. The question is whether the economics will. The cost of building AI infrastructure is staggering, but fairly well-known: The all-in costs associated with a new one-gigawatt data center—the processor chips themselves, plus the building, cooling, networking, and power requirements—run to at least $50 billion. If we imagine a new center financed with 50% debt and 50% equity, and we stipulate a relatively conservative target return on that equity (ROE) of 10%, such a new data center would need to generate $2.5 billion in net income annually to hit that target; a 20% ROE target, naturally, requires twice the income.

The costs of running this hypothetical one-gigawatt data center—including electricity, maintenance, interest on the debt issued, and depreciation and amortization—total just over $12 billion per year. To achieve a net income of $2.5 billion or $5 billion (which represent those after-tax ROE targets), this data center would need to generate gross revenues of $15.5 billion to $18.6 billion, respectively; that range effectively represents roughly 31% to 37% of total capital deployed.

Again, that's for a single data center.

Some recent estimates call for total AI infrastructure spending (which goes beyond just the hyperscalers) to hit $3 trillion to $4 trillion per year by the end of the decade. If that buildout demands 31% to 37% of the cost to be recaptured in the form of revenue, that translates to a dollar figure of nearly $1.5 trillion in necessary annual revenue accruing just to the "factory layer" (energy, chips, and infrastructure), to say nothing of the income and margin requirements of the model developer and application layers. And that $1.5 trillion is the revenue required from just 2030's CapEx, which would not include the spending and revenue requirements from prior years.

The artificial intelligence ecosystem spans multiple layers, from energy to applications



Source: Boston Partners. For illustrative purposes only.

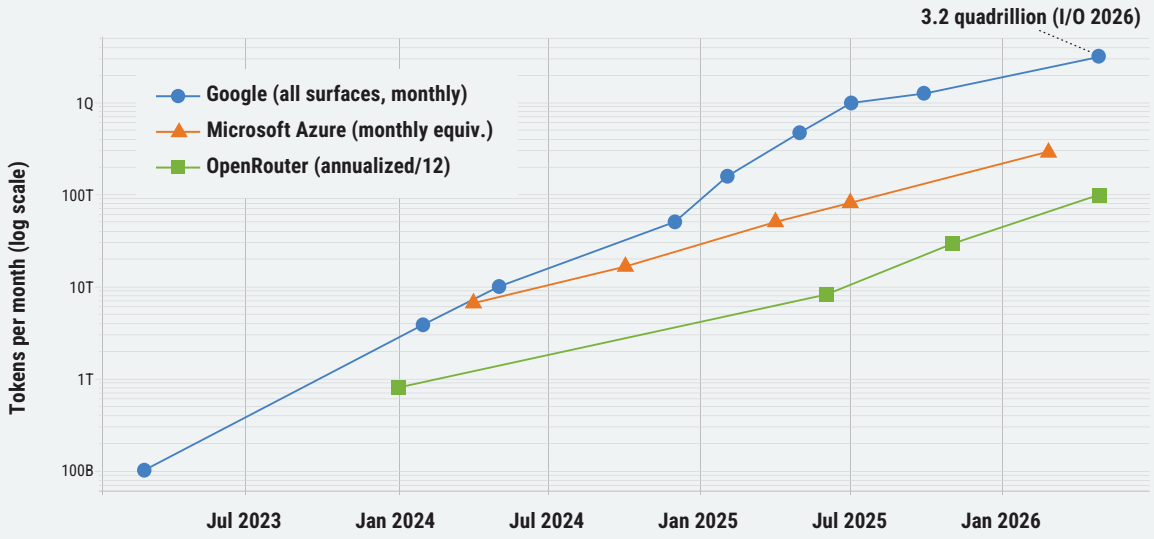
For perspective, $1.5 trillion is nearly $2,000 in spending per person for the entire population of the United States and European Union. The entire ecosystem now being built therefore hinges on a single critical assumption: that compute demand will translate directly into revenue at a scale sufficient to justify the investment.

Demand is real, but the economics remain unproven

There is no question that AI usage is growing rapidly and has, by some measures, increased by several orders of magnitude in just a few years, reflecting significant adoption across both enterprise and consumer applications. At the same time, early revenue signals are encouraging. Leading model providers have scaled revenue quickly, suggesting that users are willing to pay for incremental capabilities.

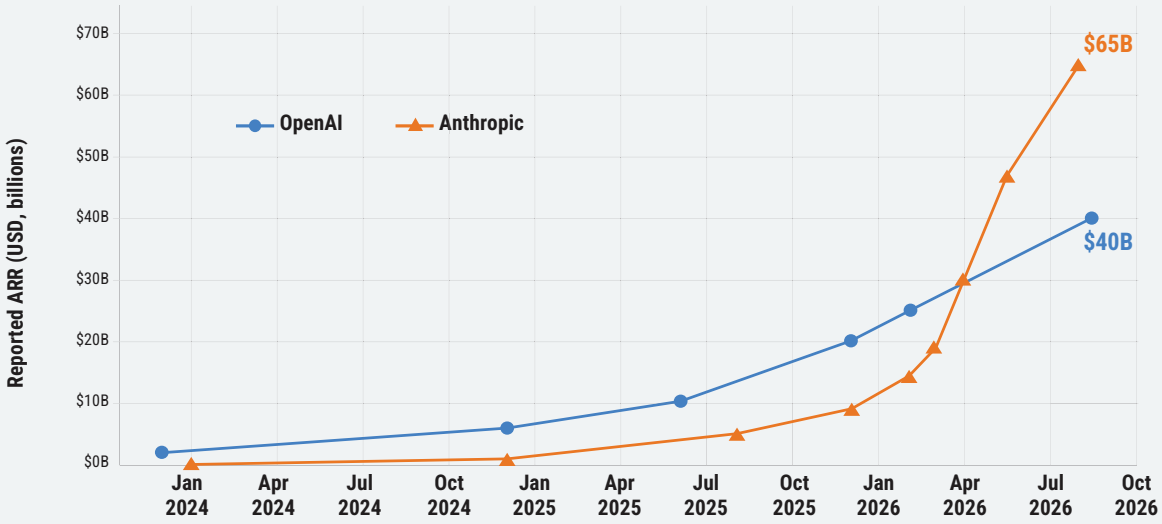
AI token usage

Monthly volume, March 2023 - December 2025



Data through May 2026, compiled August 31, 2026. Sources: Google I/O Developers' Conference 2024, 2025, and 2026; Alphabet Q1/Q2/Q4 2025 and Q1 2026 filings; Microsoft FY25 Q4–FY26 Q4 earnings; OpenRouter State of AI (a16z, December 2025) and 2026 Series B; io-fund, Beth Kindig. Note: Y-axis is logarithmic. Series are not strictly apples-to-apples—Google counts all surfaces; Microsoft 2026 monthly is estimated from partial disclosures.

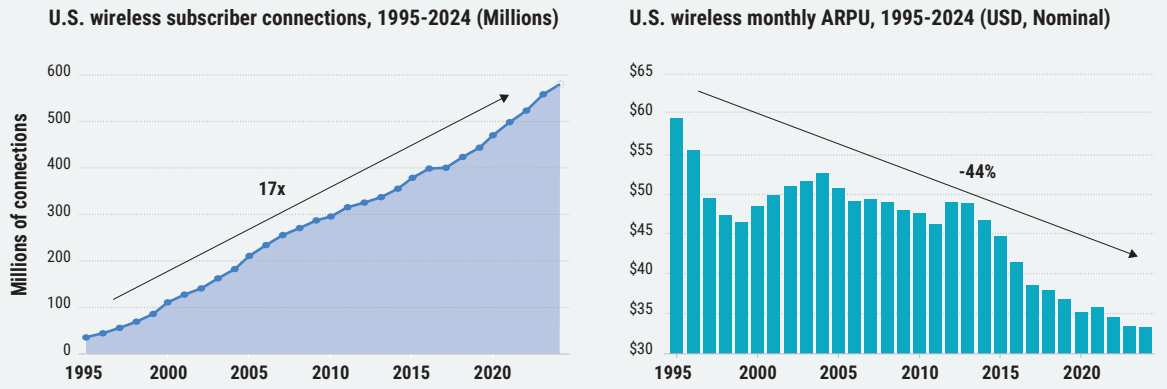
OpenAI and Anthropic annualized recurring revenue



Data from January 2024 - August 2026. Source: OpenAI; The Information; SaaStr; Sacra; Sherwood News; Anthropic. Annualized recurring revenue (ARR) is run-rate (monthly revenue x 12). Anthropic reports cloud-reseller revenue on a gross basis; comparisons are not strictly apples-to-apples. Past performance is not an indication of future results.

However, history suggests caution; the wireless industry provides a useful analogy. Over the past three decades, usage surged—by 17x, according to one estimate—as mobile adoption expanded. Yet pricing power declined significantly over that time, with average revenue per user falling even as total data consumption rose.

U.S. monthly subscriber connections and average revenue per user (ARPU)



Source: CTIA Annualized Wireless Industry Survey Results, year-end 1985-2024 (CTIA Wireless Industry Indices Report), as of December 2024.

The implication is important: Neither demand nor usage alone guarantees attractive economic outcomes. If AI follows a similar path, the industry could see strong usage growth paired with substantial margin compression.

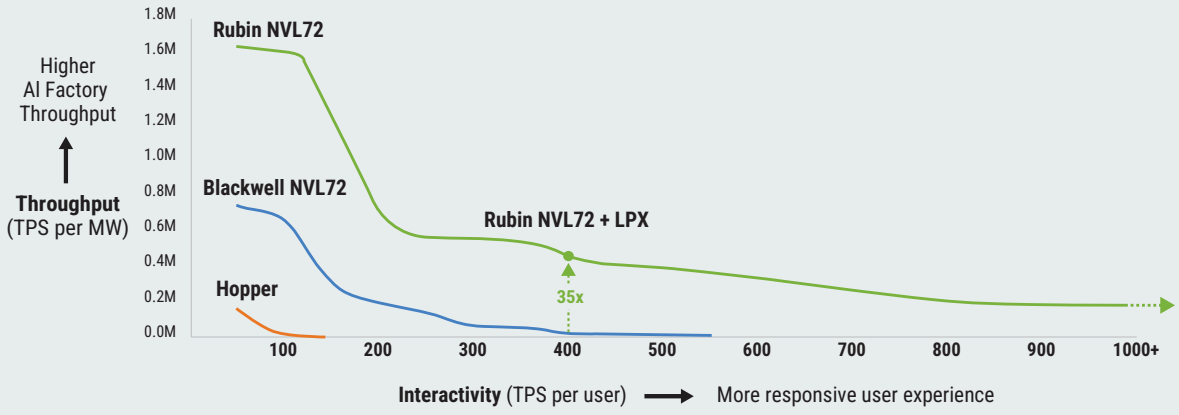
Where will value ultimately reside?

As noted earlier, the modern AI stack can be thought of as a layered system including energy, chips, infrastructure, models, and applications. (The fundamental layer, energy, is a topic worthy of investigation but beyond the scope of what we'll cover here.) Each layer plays a critical role, but not all are equally likely to capture economic value.

At the chip level, the hyperscalers are increasingly turning to homegrown products, while the cost curves for leading chipmakers are being bent down sharply. Nvidia expects its next-generation Rubin chip to deliver 35x the throughput of its current Blackwell offering. That means, all else being equal, a static workload on a Rubin chipset could be performed for a fraction of today's cost.

Next generation chips – GPU, CPU, LPX

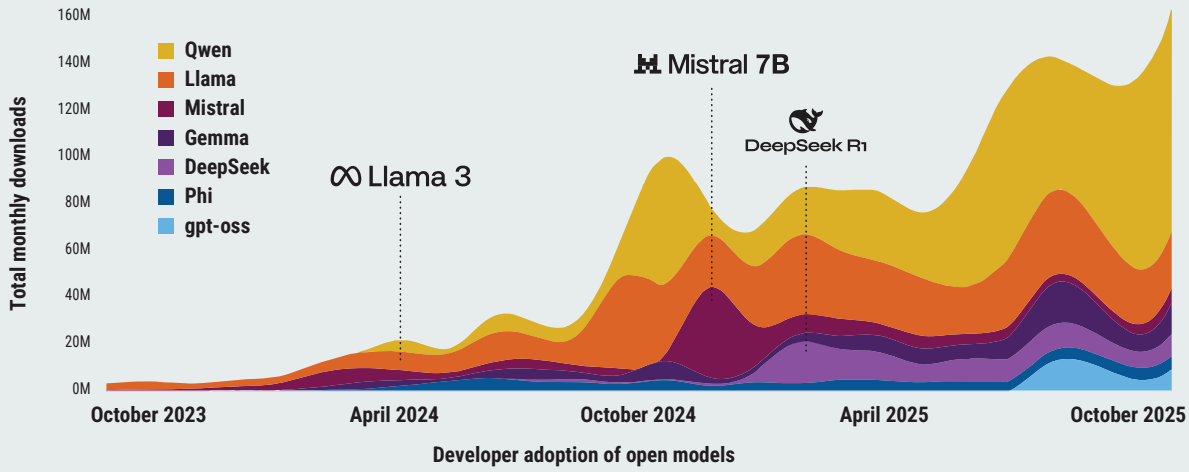
35x "Better"/10x "factory" revenue per gigawatt?



Source: Nvidia, as of March 2026. Semiconductor chips come in several varieties. A central processing unit (CPU) is the primary component of a computer that performs most of the processing functions, including executing instructions and manipulating data. A graphics processing unit (GPU) is a specialized electronic circuit designed to perform complex mathematical calculations rapidly, enabling multiple tasks to be handled simultaneously. LPX refers to a low-profile extended motherboard, which allows expansion cards to be mounted parallel to the board. Rubin, Blackwell, and Hopper are subsequent generations of Nvidia's CPU/GPU architectures.

At the infrastructure level, as we've already seen, the barriers to entry are high, driven by the intense capital and scale requirements. Turning to the model level, differentiation is still possible for the frontier labs given sustained performance and efficiency gains, but competition is only intensifying—particularly with the rise of open-source alternatives. Finally, at the application layer, the question is even more open. Are applications defensible products, or simply interfaces on top of increasingly commoditized models?

Hyperscalers are developing frontier models. What about open-source models?



Source: Nvidia, as of October 2025.

Even business models are evolving. One emerging framework resembles a hybrid of subscription and usage-based pricing, where access comes at a fixed cost, but consumption drives the incremental revenue. The durability of that model, or any model, will depend a great deal on whether providers can maintain pricing power as competition increases. If history is any guide, that outcome is far from certain.

The late 1990s fiber buildout is an instructive example—and a cautionary tale. Companies spent aggressively to build global networks, anticipating future demand. That demand did eventually arrive, but not quickly enough to justify the level of investment. The result was widespread overcapacity, falling prices, and a wave of forced mergers and bankruptcies.

Other cycles tell a similar story. The semiconductor industry, particularly the memory segment, has experienced repeated boom-and-bust dynamics tied to supply/demand imbalances. Likewise, telecommunications investment produced enormous long-term value, but returns were uneven and often delayed.

The lesson is not that AI will fail. Rather, it is that even successful technologies can produce disappointing investment outcomes if capital is misallocated or the timing of expenditures overshoots demand. For investors, the challenge is not identifying which layer of the evolving AI ecosystem will grow fastest, but rather which layer will likely retain the most value as the technology matures.

What needs to go right

For the current AI investment cycle to succeed, several conditions must hold:

- The compute equals revenue assumption must prove valid and durable
- AI must drive measurable productivity gains or create entirely new revenue streams
- The cost of compute must decline sufficiently to expand use cases
- Demand must grow without triggering widespread pricing pressure
- Economic disruption—particularly labor displacement—must not undermine broader demand

Each of these conditions is plausible, but none is guaranteed. It is, of course, still early in the AI cycle. The technology is advancing rapidly, and the potential applications are significant. At the same time, the economic model underpinning the current level of investment is still taking shape.

In that sense, today's environment reflects both measurable fundamentals and a substantial dose of speculation about the future. There is enough evidence to support a constructive view on AI's long-term impact, but also enough uncertainty to warrant caution. For value-minded investors like us, that means maintaining exposure based on disciplined valuation criteria and highly selective positioning. As in prior cycles, the winners may not be those companies closest to the technology itself, but rather those best positioned to capture and benefit from the shifting economics behind it.

About Boston Partners

Boston Partners is a value equity manager with a distinctive approach to investing—one that combines attractive valuation characteristics with strong business fundamentals and positive business momentum in every portfolio. The consistent application of this approach over more than 30 years by an experienced and long-tenured team has created a proven record of performance across economic cycles, market capitalizations, and geographies.

Capital expenditures (CapEx) are investments made by a business to obtain or improve physical assets. **Hyperscalers** are large-scale data centers used for cloud computing and data management, and can refer to the companies that provide such services. **Return on equity (ROE)** measures a company's profitability by revealing how much profit a company generates with the money invested.

BostonPartners

One Beacon Street | Boston, MA 02108 | tel: 617-832-8200 | bostonpartners.com

9117385.1