

Gartner Insights

Prepare for AI Market Consolidation: Capital Write-Downs Could Exceed \$4 Trillion by 2032

Will Sommer, Mark McDonald, John Lovelock



Prepare for AI Market Consolidation: Capital Write-Downs Could Exceed \$4 Trillion by 2032

14 September 2026 - ID G00861169 - 11 min read

By: Will Sommer, Mark McDonald, John Lovelock

Initiatives: Technologies and Markets

Multitrillion-dollar write-downs of AI infrastructure are likely, but not imminent. The provider consolidation, pricing volatility, and changing AI economics likely to follow should be the primary concern to product managers. Design products that remain viable under higher AI costs, fewer provider choices, and increasing pressure to demonstrate economic value.

Insights at a Glance

Key takeaways

- If AI capital spending continues at its current trajectory, Gartner forecasts that multitrillion-dollar write-downs of AI infrastructure assets are more likely than not by 2032.
- Gartner forecasts that AI-optimized IaaS spending will be too low for most AI compute providers to achieve sustainable returns on their capital assets before the current generation of semiconductors is fully depreciated.
- There will not be enough revenue to sustain all providers in the AI-optimized IaaS and foundational model markets, necessitating consolidation.

Recommendations

- **Roadmap products for AI resilience.** Architect critical workflows to operate across multiple model, cloud, and sovereignty environments. Avoid dependencies that require a single provider, model family, or class of frontier hardware. In-house, domain-specific models run on-premises or at the edge will be the most protected from any broader market turmoil.

- **Pressure-test product economics for the future.** Design products assuming AI pricing, provider availability, and infrastructure optionality may change materially. Prioritize product investments that remain economical even if AI costs rise materially or model-provider choice narrows. Diversify AI capabilities where practical, support sovereign and domain-specific models where viable, and maintain fallback processing paths for business-critical workflows.
- **Minimize dependence on leading-edge compute.** Products that rely on scarce, top-tier hardware are at risk of systemic cost overruns and even outright failure in the event of major capital write-downs. Product leaders must architect solutions to run primarily on mainstream infrastructure, reserving frontier chips for no more than the top 20% of highest-value workloads.

Strategic Planning Assumptions

- By 2030, cost pressure and a lack of a differentiating business model will force over 80% of neocloud providers to cease independent operations.
- By 2030, the top two model vendors will control 85% of the foundation model monetization pool in APAC and North America.

Issue

Gartner predicts that total capital spending on **AI infrastructure** will exceed \$1 trillion in 2026. For these investments to be sustainable, returns booked against AI infrastructure must exceed 12% (the typical hurdle rate for invested capital), and overall AI spending must grow to a significant portion of global GDP.

Although enterprise and consumer spending on AI has grown tremendously, it is still nascent and may prove to be insufficient:

- AI technologies are failing to achieve overly high expectations — 74% of surveyed CIOs report their AI initiatives are losing money or only breaking even (see 2026 Customer Service and Support Leaders Priority: Maximize ROI From AI).
- Disillusioned tech leaders and consumers are losing trust in AI providers and pulling back. Only 34% of surveyed IT leaders have high trust in vendors' ability to deliver on AI roadmap promises, and just 21% believe pricing will be fair and predictable (see The Human Compass: Build Trust in an AI-Native Workplace).

- AI commoditization is accelerating, challenging monetization (see [The Inference Paradox: Inference Tiering Is Critical to Protect Margins](#)).
- COGS is escalating, driven by higher compute prices and constrained access to energy. DRAM and server enterprise SSD prices, for example, are up by 231% and 306% respectively, in 2026 (see [Forecast Analysis: Servers, Worldwide, 2026](#)).
- The rate of AI innovation is significantly outpacing the rate of consumer and enterprise adoption (see [Trillions in AI Investments and Slow Enterprise Adoption Are Widening an Already Large Value Chasm](#)).

In our baseline modeling scenario, the enterprise and consumer markets for AI will not generate enough revenue to profitability support all of the planned AI capital spending until 2032, at earliest. We estimate that through 2032 AI data center owners would need to earn over \$14 trillion to cover capital costs, financing costs, and an ROIC of 12%. However, total revenue from IT and consumer sources that can be booked against data centers during this period is forecast at no more than \$8 trillion (see “Implications” and “Evidence” sections for methodology).

This imbalance indicates that at forecasted trends massive AI capital asset write-downs are more likely than not (see [The Potential \\$3 Trillion Dollar ROIC AI Gap](#)).

Product leaders should not view AI capital write-downs as an infrastructure problem alone. Even if a product organization owns no AI infrastructure, provider consolidation, price volatility, model discontinuations, service rationalization, and sovereign AI requirements can materially affect product economics, roadmap execution, and customer commitments. Product leaders should therefore design AI-dependent products assuming that today’s AI supply landscape will become materially more concentrated through the end of the decade.

Impact

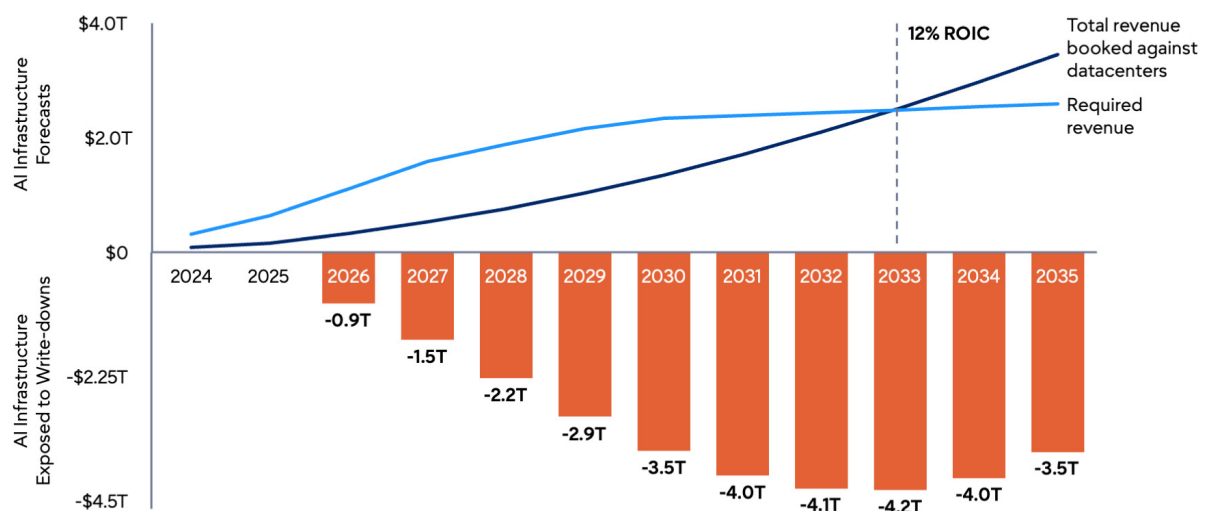
- The overall cost of AI products and services will rise significantly, even as per-token cost economics become more efficient. We forecast that through 2030, AI compute providers will need to generate a profit of over 100% on every AI token that can be monetized for inference to achieve an ROIC of 12%.
- Through at least 2030, providers of AI software and services will struggle to earn sustainable margins because they will not have control over their COGS.
- Agility will differentiate winners from losers. Providers and enterprises will ultimately benefit from access to multiple compute environments where a diverse range of hardware is dynamically aligned with workload requirements. Aligning to a single vendor or offering will link corporate fates.

Implications

AI Capital Investments Will Struggle to Generate Returns Through 2032

Figure 1: AI Capital Investments Will Struggle to Generate Returns Through 2032

AI Capital Investments Will Struggle to Generate Returns Through 2032



Source: Gartner
861169

Gartner®

Cumulative global AI capital spending is forecast to reach \$7.5 trillion between 2025 and 2030. By 2030, annual AI capital spending is forecast to approach \$2 trillion, representing well over 1% of global GDP (see the “Evidence” section for methodology and assumptions). This analysis is an extension of previously published research (see The Potential \$3 Trillion Dollar ROIC AI Gap).

By comparison, total AI revenue from software, services, and consumer products is forecast to reach \$4.4 trillion by 2030. However, this revenue needs to be shared across the value chain, and we estimate that no more than 30% of revenue from AI-powered products and services will be booked against underlying capital investments during this period.

Under these baseline conditions, AI capital investment in aggregate would not achieve ROIC of 12% — the investment hurdle rate — until 2032 at earliest.

This analysis is generalized to the entire AI market. Many individual assets and providers will outperform the mean and many will underperform. Provision of raw AI compute will become a fractional-penny business where incremental and cumulative advantages in hardware, software, AI model configuration, energy access, financing terms, and channel development will be necessary to earn sustainable returns. Value will come from the services and ecosystem built around compute, rather than the raw compute itself.

The neoclouds will be the most exposed to the imbalance between revenue and capital spending. The hyperscalers and other more diversified providers will be less exposed due to their access to other markets and significant cash flow from existing operations.

Profitability will be challenged by market fundamentals. While Gartner forecasts that AI markets will remain compute-constrained through at least the end of the decade, this imbalance between supply and demand will not guarantee ROI (see Tech FutureSight: Token Scarcity Forces Advanced Reasoning Rationing Through 2030). Both providers of compute and AI products and services will struggle to generate sustainable margins through at least the end of the decade. This is because:

- AI markets are crowded and awash with capital, supporting existential competitions for market share (see Market Share Analysis: Generative AI Models, Worldwide, 2025).
- AI markets are volatile and unpredictable. Data center space, hardware and energy have become costly bottlenecks in many domains, and risk is endemic across AI segments (see The AI Risks No Enterprise Leader Can Afford to Ignore).

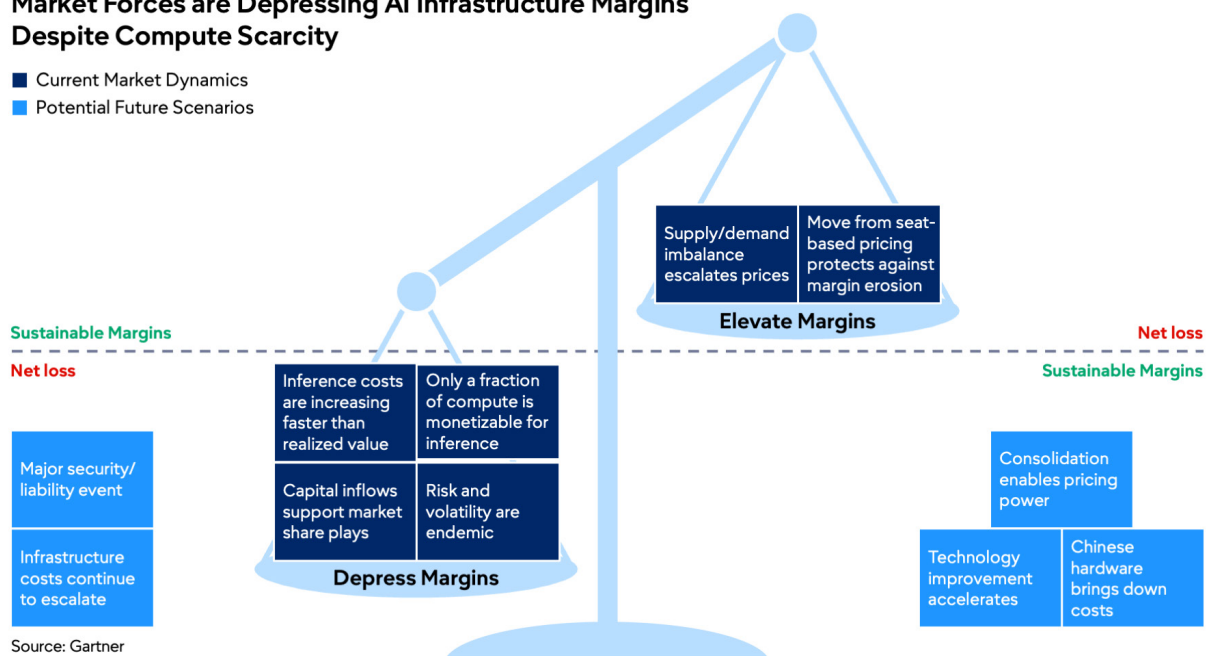
- Many AI compute providers are their own customers, meaning that training and inference costs need to be recouped from unproven AI models. Circular investments have become commonplace (see Assessing the Risks in AI's Circular Investment Cycle).
- Falling token costs are not expanding gross margins; they are subsidizing growing token usage. This means that better unit economics are escalating the overall consumption and cost of AI without providing a clear pathway to commensurate, predictable value (see The Inference Paradox: Inference Tiering Is Critical to Protect Margins).

Providers could ostensibly earn healthy gross margins on each token while still being unable to cover the costs of each subsequent wave of capital and IP investment necessary to remain competitive in their respective markets and fuel the next generation of models.

Figure 2: Market Forces are Depressing AI Infrastructure Margins Despite Compute Scarcity

Market Forces are Depressing AI Infrastructure Margins Despite Compute Scarcity

- Current Market Dynamics
- Potential Future Scenarios



Gartner

Over the long run, market prices that are depressed below provider costs will be unsustainable. However, through at least 2028, it is likely that the massive pool of capital deployed to AI, combined market fragmentation, will fuel intense competition for market share, inclusive of price subsidization. Only once markets consolidate and mature will providers be able to earn sustainable margins (see Tech FutureSight: Token Scarcity Forces Advanced Reasoning Rationing Through 2030).

Low Returns on Invested Capital Will Result in Write-Downs

Publicly traded companies must test assets for capital impairment following significant events or structural changes that signal an investment may not be recoverable.

Circumstances where a test for impairment must be performed include:

- A significant drop in the market price of an asset
- When an asset becomes obsolete
- When a major client is lost
- When an asset fails to generate positive cash flow

The magnitude of a write-down is determined by the following formula:

$$\text{Impairment Loss} = \text{Carrying Value of Asset} - \text{Fair Value of Asset}$$

Failure to achieve sustainable ROIC on rapidly depreciating assets would almost certainly necessitate write-downs. Based on our projections, exposure to asset write-downs on a sustained capital investment trajectory is likely to exceed 30% through 2032, representing over \$4 trillion in total asset value.

This assessment assumes that investments were made with a 12% target ROIC. If investments targeted an ROIC approaching 25%, the rate commonly earned by hyperscalers on capital investments prior to the AI buildout, write-down exposure would exceed \$6 trillion through the forecast period.

To put this disparity in perspective, achieving a 12% ROIC by 2030 — by which time the first big wave of AI-optimized semiconductors will begin to become commercially obsolete — would require a 7x increase in Gartner's current revenue forecast for AI-optimized IaaS, holding all else constant. Annual spending on AI-optimized IaaS would need to grow from \$21.5 billion last year to over \$1.2 trillion in 2030.

Major asset write-downs would reset market expectations and squeeze capital flows. Historically, major asset write-downs have been correlated with significant market corrections. Gartner predicts that a major correction would reduce U.S. economic output by up to \$650 billion relative to trendlines (see Emerging Issue: U.S. Economy Is Forecast to Shrink by \$650 Billion If the AI Market Crashes).

Evidence

This analysis is predominantly based on proprietary Gartner models using Gartner IT spending forecasts benchmarked to announce AI spending targets by major AI providers.

ROIC is calculated as:

$$ROIC = \text{Net Operating Profit After Tax (NOPAT)} / \text{Invested Capital}$$

Gartner IT spending forecasts are through 2030 only. After 2030, IT spending forecasts are extended as projections applying various methodologies:

- Required revenue: Capital spending grows with inflation only.
- Total revenue booked against data centers: Linearly extended line of best fit adjusted with segment-level AI multipliers to account for the changing percentage of AI intensity in enterprise IT and consumer products.

A significant proportion of AI revenue earned against data centers will be pass-through, meaning an AI provider will charge for an AI-enabled service that relies on infrastructure it owns. For example, a hyperscaler may offer deployment support services that fall under Gartner's "direct AI services" IT spending segment, but where revenue is supporting the AI data center buildout. The proportion of each Gartner IT spending forecast segment that we predict will ultimately support the ROIC of owned AI infrastructure assets is detailed in Table 1.

Table 1: Percentage of AI Revenue Booked Directly Against Data Centers by Segment
(Enlarged table in Appendix)

Segment	2025	2035
Direct AI services	8%	15%
Indirect AI services	4%	8%
Securing AI	18%	23%
AI-amplified security	18%	23%
Embedded chatbots and assistants	25%	60%
Embedded agentic AI	25%	60%
VSS	8%	15%
AI models	35%	70%
AI platforms	40%	60%
AI data	35%	53%
AI-optimized IaaS	100%	100%
Consumer	25%	60%

*This table only represents revenues booked directly against data centers. While all consumer AI spending eventually is processed by a data center, much of this attribution is already included as "AI-optimized IaaS." For other categories, revenues booked against data centers represent pass-through revenues.

Source: Gartner (September 2026)

Our modeling assumes that semiconductors depreciate over five years, while other infrastructure (cooling, power shell, etc.) depreciates over 20 years. The modeling does not account for complex financial arrangements such as exchanging compute for equity or using multitiered leasing arrangements. Although extended depreciation schedules and exotic financing structures could modestly affect the timing and allocation of reported costs, they would not materially change the underlying economics. The gap between the revenue required to support infrastructure investment and the revenue that can realistically be generated is far more significant than any realistic accounting or financing adjustments.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

[The Potential \\$3 Trillion Dollar ROIC AI Gap](#)

[Tech FutureSight: Token Scarcity Forces Advanced Reasoning Rationing Through 2030](#)

[Emerging Issue: U.S. Economy Is Forecast to Shrink by \\$650 Billion If the AI Market Crashes](#)

[Agentic Arbitrage Is Making Enterprise SaaS Headless – Putting \\$234 Billion Up for Grabs](#)

© 2026 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates. This publication may not be reproduced or distributed in any form without Gartner's prior written permission. It consists of the opinions of Gartner's Business and Technology Insights Organization, which should not be construed as statements of fact. While the information contained in this publication has been obtained from sources believed to be reliable, Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information. Although Gartner insights may address legal and financial issues, Gartner does not provide legal or investment advice and its insights should not be construed or used as such. Your access and use of this publication are governed by [Gartner's Usage Policy](#). Gartner prides itself on its reputation for independence and objectivity. Its insights is produced independently by its Business and Technology Insights Organization without input or influence from any third party. For further information, see "[Guiding Principles on Independence and Objectivity](#)." Gartner insights may not be used as input into or for the training or development of generative artificial intelligence, machine learning, algorithms, software, or related technologies.

Table 1: Percentage of AI Revenue Booked Directly Against Data Centers by Segment

Segment	2025	2035
Direct AI services	8%	15%
Indirect AI services	4%	8%
Securing AI	18%	23%
AI-amplified security	18%	23%
Embedded chatbots and assistants	25%	60%
Embedded agentic AI	25%	60%
VSS	8%	15%
AI models	35%	70%
AI platforms	40%	60%
AI data	35%	53%
AI-optimized IaaS	100%	100%
Consumer	25%	60%

*This table only represents revenues booked directly against data centers. While all consumer AI spending eventually is processed by a data center, much of this attribution is already included as "AI-optimized IaaS." For other categories, revenues booked against data centers represent pass-through revenues.

Source: Gartner (September 2026)

Actionable, objective insights

Position your organization for success. Explore these additional complimentary resources and tools for CIOs:

eBook



Strategic Cost Optimization for CIOs in the Age of AI

Manage costs strategically to spend smarter
and scale faster.

[Download Now](#)

Hub Page



Gartner AI Hub

Discover practical AI insights for confident
decision making.

[Learn More](#)

Virtual Briefing



IT 2030: Empowerment or Chaos? CIO's Role in AI-Powered Democratization

Learn how CIOs can enable AI innovation while
managing risk.

[Watch Now](#)

Tool



Gartner AI Maturity Assessment

Assess your AI maturity and build a roadmap to
scale impact.

[Explore Tool](#)

Already a client? Get access to even more resources in your client portal.
[Log In ↗](#)

Connect with us

Get actionable, objective business and technology insights that drive smarter decisions and stronger performance on your mission-critical priorities.

Contact us to become a client:

U.S.: 1 855 322 5484

International: +44 (0) 3300 296 946

[Become a Client](#)

Learn more about Gartner for CIOs
gartner.com/en/chief-information-officer

Stay connected to the latest insights



Attend a Gartner conference
[View Conference](#)