Deloitte.

Together makes progress

The pivot to tokenomics: Navigating AI's new spend dynamics

A note from the authors:

AI economics affect most organizations and the C-suite uniquely. ***This paper guides those familiar with AI tokens in making strategic choices.*** If you're just beginning your exploration of tokenomics, look for additional research soon.

Traditional total-cost-of-ownership frameworks miss *the reality of AI*

Volatile workloads, new infrastructure demands, and tokens as the practical unit of cost

Across industries, Generative AI (GenAI) has become the fastest-growing line item in most corporate technology budgets—already consuming up to half of IT spend in some firms.[1] Cloud bills are rising nearly 20% year over year, driven by AI workloads.[2] At the same time, geopolitical uncertainties are intensifying calls for data sovereignty and technology infrastructure independence, making many enterprises think about AI sovereignty and gaining greater control over their infrastructure.[3] This is no longer a CIO operational issue; it is a CFO-and-board capital question about how to responsibly manage an investment of this scale and volatility.

Unlike prior technology waves governed by licenses or virtual machines, AI spend often scales in nonlinear and unpredictable ways. AI capabilities run on **tokens**: small chunks of data—text, image or audio—that AI systems process in training, inference, and reasoning. Every AI interaction consumes tokens, and every token carries a cost.

The complexity of AI's economics hides within these tokens. Costs rise not only with user adoption but with workload design, algorithmic complexity, and infrastructure intensity. What exactly are the thresholds to move across different consumption choices? It depends on the organization. Roughly a quarter of respondents in a Deloitte 2025 survey[4] of data center and power executives say they or their clients are ready to make the move off of cloud to alternatives as soon as costs reach just 26% to 50% of those alternatives, showing high sensitivity to even modest price changes, while others plan to wait until cloud costs exceed 150% of the cost of alternatives. The decision point remains unclear given the high variability patterns of AI technologies. For example, advanced reasoning models that keep context across multiple steps can consume much more compute than basic one-shot responses. As NVIDIA projects a billion-fold surge in AI computing and Google now processes 1.3 quadrillion tokens a month[5]—a 130-fold leap in just a year—the capital and energy implications are profound.

Traditional total cost of ownership (TCO) approaches are no longer the best way to manage AI economics. Leaders may be better served by precision economics—the ability to track, predict, and optimize spend at the token level. Tokens translate opaque infrastructure choices into tangible financial terms: the true cost of generating a dollar of revenue, margin, or productivity.

The competitive divide will not likely hinge on who adopts AI first, but on who manages its cost structure with discipline. AI spend will likely separate value creators from value eroders. The former convert tokens into measurable enterprise output; the latter accumulate ungoverned cost that compounds quietly across the stack.

The elusive AI ROI

Despite rising investment, many leaders appear to still be chasing measurable return on investment (ROI) from AI initiatives.

- Nearly half (45%) of 500 leaders surveyed in Deloitte's 2025 US Tech Value survey expect it will take up to three years to see return on investment from basic AI automation.[6]
- Six in 10 of those completing Deloitte's 2025 Tech Value survey believe more advanced AI automation will take even longer to reach ROI.
- Of the 1,326 global finance leaders surveyed for Deloitte Global's inaugural Finance Trends report, fielded May 2025, 28% said AI investments are delivering clear, measurable value.[7]

But the issue isn't whether AI will deliver value—it's how to measure and manage that value in a way ROI frameworks cannot. For many organizations, adopting AI is no longer optional; it's a strategic response to competitive or existential pressure.

That makes understanding the ***economics of AI***—how costs, workloads and returns flow through tokens—the new imperative for leaders.

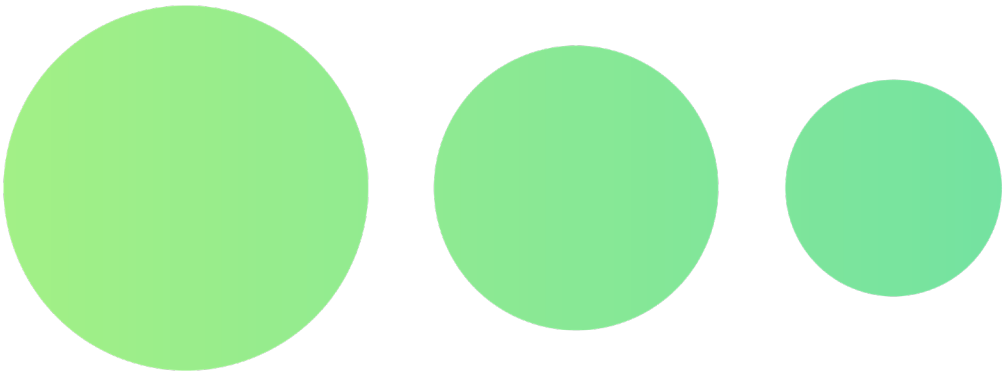
Tokens: The new currency of AI

Unlike traditional pricing based on compute time—which is relatively static—token-based pricing ties cost directly to the actual work AI performs. **Each token represents both a unit of computation and a unit of cost.** In that sense, tokens are the **true currency of AI economics**—as indispensable to machine intelligence as kilowatt hours are to electricity. The difference is that token demand is far harder to predict or control, making AI spend inherently volatile.

- **Nonlinear demand:** Complex reasoning models improve performance but can consume more tokens than simple inference tasks.
- **Fluctuating token use:** Token use fluctuates with experimentation levels, workload design, model choice and even prompt engineering.
- **Varying pricing:** Token price keeps changing based on AI model capabilities and the efficiency of the underlying infrastructure.[8]

While this volatility appears to stem from usage patterns, its roots are in the tech stack. The compute, storage, and networking decisions that power AI models determine how efficiently tokens are processed—and how costly each one becomes.

A token is not just a technical measure—it is an economic signal. Each token carries the compound effect of GPU design, storage, throughput, network latency, and facility economics. The discipline lies in tracing lineage—from infrastructure to the AI model to outcome—and aligning those decisions so token costs stay proportional to business value.



How tokens are bought

AI spending is not a single market; it fractures into different economic realities depending on how organizations consume intelligence. Some leaders experience AI costs only as a software-as-a-service (SaaS) line item, others as metered application programming interface (API) calls, and a growing group/cohort manage it directly through infrastructure ownership—balancing GPUs, storage, networking, and energy.

Buying patterns

- **Generating through packaged software** abstracts tokens almost entirely. Leaders see a predictable subscription or per-seat fee, but little transparency into token consumption efficiency. The risk is less control for more simplicity.
- **Consuming through APIs** makes tokens explicit. Every query is metered, billed, and exposed. This brings transparency, but also volatility: Costs rise based on workload design, prompt length, and hidden choices of infrastructure providers. Costs go up due to a token meter running in real time.
- **Running on owned infrastructure** brings token economics fully in-house. Tokens become the outcome of decisions about GPUs, storage tiers, networking, and energy contracts. This approach demands high capital and technical capability but offers the greatest control over long-term cost structure and data sovereignty. The emerging shorthand for this strategy: **the AI factory**.

Each of these choices is grounded in existing and future technical and operating decisions given system cost, latency, security, and other needs, which change how tokens flow into enterprise profit and loss (P&L).[9]

What is an AI factory, and when does one make sense?

Deloitte defines an AI factory as a specialized infrastructure (compute, network, and storage) along with optimized software and services that enables the entire AI life cycle at high performance scale. The primary product is intelligence, measured by token throughput, which drives decisions, automation, and new AI solutions.

One of the hardest decisions enterprises face is whether to continue paying for tokens off premises (off-prem)—through APIs or traditional SaaS companies—or to build an AI factory and self-manage the infrastructure. The economics vary sharply depending on scale, sensitivity, and predictability of demand:

- **Off-prem (API or traditional SaaS companies):** May be most efficient for early pilots, spiky or seasonal workloads, or use cases with low data sensitivity. Costs are typically higher per token but predictable and flexible, with no up-front capital expense (capex).
- **AI factory:** Can become attractive when workloads are large, predictable, latency-sensitive, and cross a threshold where building and operating infrastructure delivers lower effective token economics than continuing to rent them. Although capex investment may be needed, per-token costs fall as infrastructure is fully utilized, and sovereignty risks are controlled. Beyond the traditional on-premises (on-prem) or colocation (co-lo) providers, **an AI factory can also be stood up using fast-growing cloud alternatives (neoclouds) to manage workload redistribution trends,** as detailed in a recent Deloitte survey.[10]

The decision is not binary. For most global enterprises, the reality is hybrid. Smaller, less predictable and exploratory workloads may stay in API form, while scaled, high value workloads may run on an AI factory as applications scale and economics stabilize. AI model preference and selection may also drive enterprise decision making.

How tokens are priced

Once leaders understand their buyer type (generate, consume, run), the next challenge is to see how tokens are priced. The same AI model could be billed as a seat license, or a token meter or GPU-hours, depending on how it is consumed. There are three major constituents to token pricing:

1. The underlying tech stack
2. How it is hosted and consumed
3. What type of AI model and level of customization is required to power the solution

The AI tech stack

Every token processed by an AI model reflects a cascade of infrastructure decisions.

For packaged buyers, in most cases and at least for now, these costs are hidden. Costs are abstracted, bundled into familiar enterprise contracts and vendor managed across every layer of the tech stack, which makes unpacking TCO challenging.

For API consumers, every element of the AI tech stack shows up indirectly as per-token fees or throughput charges. Price varies by AI model accessed, with different input and output rates, usually reported in token per million. Discounted pricing options such as reserved token capacity, prompt caching, or batch execution rates are usually offered, while in some cases enterprise customers may also get user-based pricing. Additionally, storage or egress charges may further add to TCO.

For self-hosted solutions, tokens are not purchased at all; they emerge from explicit capex and operating expense (opex) decisions related to infrastructure choices (figures 1 and 2).

What changes across buyer types is not whether these costs exist—they always do—but who sees them, controls them, and pays for them.

Figure 1. **How technical decisions can drive token costs and implications for an AI factory**

STACK COMPONENT	TOKEN IMPLICATIONS	SELF-HOSTED AI FACTORY
Compute **Graphical processing units (GPUs) and accelerators**	Modern GPUs and **high-bandwidth memory** shorten time per token but come with higher acquisition or rental cost.	Largest direct cost Direct infrastructure spend Rapid release cycles
Storage **High-speed data access**	AI workloads stream terabytes using **nonvolatile memory** and parallel file systems to sustain performance and manage cost. Legacy storage inflates per-token costs by adding latency as GPUs wait for data.	Nonvolatile memory, parallel file systems, vector databases Heavy investment
Networking **GPU Interconnects (InfiniBand, NVLink, PCIe Gen 5)**	Training across thousands of GPUs requires ultra-low-latency interconnects to cut idle cycles and lower cost per token, while traditional approaches often drive token costs higher.	Direct spend
Power and cooling **Energy intensity of AI racks**	A single next-generation GPU rack can draw between **250–300 kW**, compared with **10–15 kW** for non-AI servers. Whether billed directly (on-prem) or embedded in cloud pricing, this power use shows up in every token consumed.	High opex (250–300 kW racks) Liquid cooling requirements
Facilities **Physical infrastructure requirements**	Heavier racks (up to **3,000 lb**,[11] nearly 40% more than traditional), may need reinforced flooring and advanced cooling to be embedded in the cost of every token.	Direct capex (reinforced floors, racks)
Operational costs	Related to staffing and operations: • IT ops and management • Software and licensing • Application development and integration • Data management and governance • Inference and serving • Security and compliance • User training and change management	Full machine learning operations (MLOps) costs Full center of excellence (COE) and upskilling Orchestration frameworks and MLOps tools (data, orchestration, security) Direct compliance spend, etc.

Source: Deloitte analysis based on project experience

Hosting models

How tokens are priced also depends on where and how AI models are hosted. The same large language model (LLM) can be deployed via on-prem, colocation, hyperscalers, or API access, with radically different economics. For a package buyer, this decision is again invisible and resides with the vendor. For the API consumer, it can vary based on which of the many models on the market is being consumed, and this explains why the same task may cost more depending on the provider. For self-hosted AI infrastructure users, all hosting types are possible, and it is often the most important determinant of unit economics.

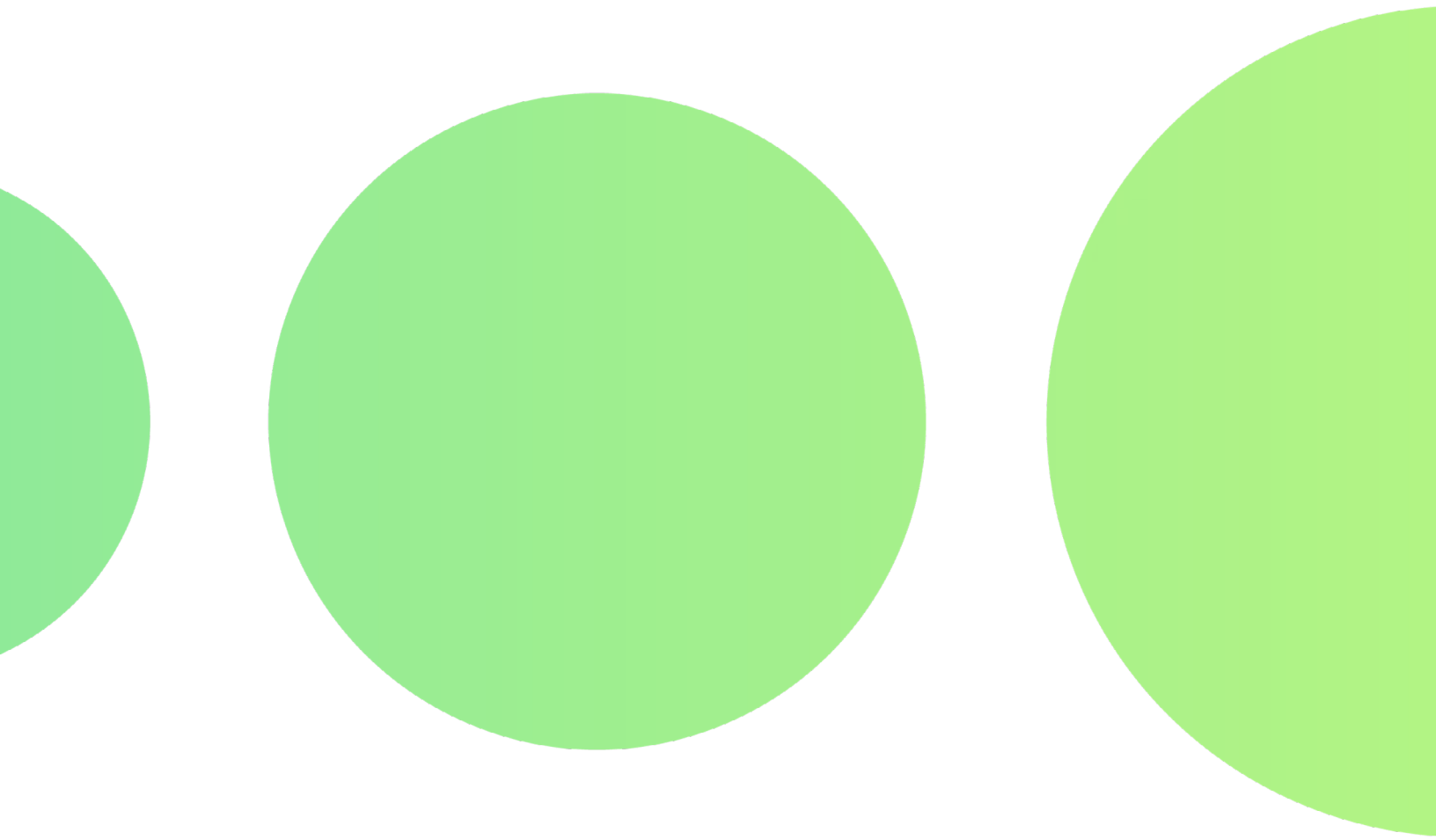


Figure 2. **GPU consumption models and cost structure**

	ON-PREM	NEOCLOUD PROVIDERS	HYPERSCALER	API ACCESS
Capex vs. opex	**High capex/low opex**	**Pure opex**	**Pure opex**	**Pure opex**
Unit cost of compute (GPU/hour)	**Lowest** ~$1–$2	**Medium** ~$1–$4 average, but high variability, on demand	**High** ~$3–$7, region/model dependent	**Very high** $0.40–$100 or more per million output tokens
Scalability	**Medium** Slow due to procurement, power, and setup	**High** Dynamic resource provisioning	**Medium/high** Dynamic scaling with near-infinite top-end	**Very high** 100% managed by the provider
Latency	**Lowest** Full control over hardware stack	**Low** Purpose-built for AI, but physical layout not controllable; with neoclouds, low physical proximity is manageable	**Medium** Near-zero control over physical layer and workload placement	**Medium/high** No control over provider infrastructure/network, with long-distance communication
Control and customization	**Full**	**Medium** No control over physical layer or maintenance; high control over what's hosted	**Medium** Treated identically to neocloud providers	**Very low** No control over infrastructure layer and limited control over AI model tuning, format of response
Security and data sovereignty	**Highest** Complete control over data encryption, transit, storage	**High** Treated identically to co-lo; neoclouds offer higher data encryption	**Medium** Data leakage risk and low control over exact hosting location	**Low** No control over provider architecture or governance practices
Deployment time	**Long** Multi-month procurement, delivery, and setup	**Instant**	**Instant**	**Instant**
Maintenance responsibility	**Customer** Managed services and shared responsibility model (e.g., facilities, energy, etc.)	**Shared** Physical infrastructure: provider; all other layers: customer	**Shared** Physical infrastructure: provider; all other layers: customer	**AI model provider**
Best use cases	**Stable, high-throughput workloads**	**Elastic compute, proofs of concept (POCs), cost-sensitive workloads; neoclouds may bring added functionality for data-sensitive workloads**	**Elastic compute, POCs**	**Fast experimentation, agents, retrieval-augmented storage (RAG)**

Source: Deloitte analysis based on public and proprietary estimations, including publicly available GPU pricing data, API pricing benchmarks, and hyperscaler cost calculator references. Indicative references include public GPU cost analysis and total-cost-of-ownership models (e.g., semi-analysis AI TCO framework); public API pricing benchmarks for Generative AI models (e.g., representative GPT-5 family rates); hyperscaler compute pricing estimates derived from standard cloud cost calculators

Ultimately, the cost structure follows the architecture. Compute density, network proximity, and storage throughput each influence how efficiently tokens are processed—and therefore, where a model should live. The decision isn't about speed or preference; it's about matching workload physics to business economics. In our experience, we've found hybrid architectures sustain performance without inflating token costs.

AI model selection

AI model strategy is a second decision point: open-source or closed AI models (proprietary). Package buyers inherit whatever the vendor builds. API users can choose providers but not the models' economics. Only self-hosted AI factory users control the full trade-off across cost, flexibility, and sovereignty.[12]

Open-source AI models

Open-source models are generally free and typically run in self-hosted environments, giving enterprises greater control, customization, and data sovereignty. They are well suited for fine-tuning on proprietary or sensitive data, minimizing vendor lock-in, and lowering token costs over time.

Examples include Meta Llama, Mistral, and others. Emerging frameworks such as NVIDIA NIM Microservices illustrate how vendors are packaging open-source models into standardized, secure deployment units—bringing operational discipline to what was once bespoke integration work.

Proprietary (closed) AI models

These are consume-as-you-go, typically billed per token and allow users to quickly hit the ground with no up-front investment, are pretrained, have strong out-of-the-box functionality, and enable access to vendor support for operational support. Examples of such AI models include Anthropic Claude, Google Gemini, OpenAI GPTs, xAI Grok, and others. However, this typically comes with higher per-token cost, lower cost predictability due to fluctuating token usage, lack of customization, open concern around data storage, and risk of vendor lock-in.

Decoding the AI *cost curve*

AI economics follow Jevons' paradox: As efficiency improves, total consumption rises.[13] Token prices are falling fast—what once cost dollars per thousand now costs pennies per million—and Deloitte projects the average inference cost will drop from $0.04 per million tokens in 2025 to about $0.01 by 2030.[14]

Yet enterprise spending continues to surge.[15] As agentic systems and multiagent workflows proliferate, token demand grows exponentially—often faster than infrastructure efficiency gains can offset. The paradox isn't that AI is becoming cheaper; it's that efficiency itself is driving expansion. Without disciplined cost governance, total costs grow.

Who pays the bill?

The cost curve doesn't affect every participant the same way. As token consumption accelerates, the question becomes who ultimately absorbs that spend—the enterprise, the vendor, or the end user—and how those dynamics evolve as workloads scale and grow more complex. Deloitte's TCO analysis examines exactly where and when those costs shift.

The token TCO estimation and *scenario analysis*

To quantify these dynamics, Deloitte conducted a detailed token TCO analysis designed to capture how AI's underlying economics shift across the full tech stack. The analysis tested how total cost of ownership evolves along three critical dimensions that shape token pricing:

1. **Technology stack:** The GPUs, AI models, and architectures powering AI workloads.
2. **Hosting approach:** Comparisons as usage and complexity scale over time.
3. **Usage scaling:** Increase in the overall token consumption driven by increase in user count or the complexity/depth of reasoning each use case demands.

The objective was to understand how these factors interact to redefine organizational strategy based on what the key drivers of AI TCO are, how costs evolve as usage scales, and where the inflection points emerge in cost per token. Before presenting the outcomes, the next section outlines the key assumptions and configurations underpinning the model used in our tests.

Model assumptions

The model was built to test realistic, enterprise-scale conditions rather than idealized lab settings.[16] While it can accommodate a wide range of configurations, the version summarized here reflects a representative scenario across common enterprise workloads. The baseline configuration included:

- **Compute stack:** NVIDIA HGX B200 GPU Server (NVLink/NVSwitch Enabled) | CPU – AMD EPYC 9654.
- **LLM:** Llama 3.3 70B FP8 TP2, GPT-4o selected because a variety of common configurations were being tested.
- **Hosting models:** On-prem, API access, specialized neocloud providers (NCPs). NCPs offer hourly rates as well as reserved contracting for different periods. In this model, we assumed hourly and not reserved pricing.

This setup enabled Deloitte to isolate how hosting choices, AI model selection, and usage maturity interact to drive token consumption and total cost. The following analysis highlights the resulting cost curves and inflection points that emerge as usage scales. The analysis simulates growth scaling in increments of 8 GPUs (figure 3).

Figure 3. **Scenario complexity and token assumptions driving four-year TCO dynamics**

TOKEN SCENARIOS	EXAMPLE SCENARIO DESCRIPTION/USE CASE
YEAR 1 Pilot stage	**Initial deployment of simple use cases such as chatbot or FAQ assistant:** A lightweight conversational AI used for customer service, HR inquiries, or basic IT help desk support. Handles short, structured Q&A with minimal context retention.
YEAR 2 POC/lightweight adoption	**Scaling to include knowledge-driven use cases such as document summarization and knowledge search:** Internal enterprise assistant that retrieves and summarizes policy documents, proposals, or contracts. Includes semantic search and multiturn conversations.
YEAR 3 Inferencing at scale	**Maturing to drive decision-support use cases such as an analytics co-pilot:** Assists consultants, analysts, or auditors in generating insights, drafting reports, or performing data analysis across multiple data sources. Includes reasoning, structured output, and integration with enterprise systems.

Source: Deloitte analysis

Navigating the economics of an accelerating technology environment

The rapid pace of AI hardware advancement has created obsolescence cycles that far outpace traditional depreciation schedules, with GPU generations now refreshing rapidly. For example, recent model releases quickly outgrew the capabilities of previously leading GPUs to unlock features, while legacy support for older hardware diminishes. Newer GPUs that switch to an annual release cycle further accelerates these refresh demands, challenging enterprises to continually balance the benefits of faster upgrades with the risk of falling behind.

Such recent advances in GPU technology have enabled AI applications requiring larger context lengths, such as reasoning models, summarizing extensive text corpora, and high-fidelity multimodal tasks like analyzing hour-long videos. These use cases, including agentic reasoning, demand substantial GPU memory and the latest hardware to accurately process such complex or large-scale data. However, adoption of multimodality, and agentic reasoning at the enterprise level is in its early stages, and inference tasks often run well on older GPUs especially for midsize models.

As token pricing for AI models declines and the economics of "build vs. buy" shift rapidly, enterprises cannot rely on static assumptions and should develop forward-looking infrastructure strategies—carefully planning upgrades, assessing costs, and ensuring investments remain viable as the market stabilizes over time.

Analysis outcome

The TCO simulation incorporated real-world parameters across the full AI value chain—from hardware utilization and energy costs to facilities expenses. Each variable was calibrated to reflect current market conditions and operational realities rather than theoretical efficiency.

This approach ensured a holistic view of cost behavior: how GPU utilization rates, power efficiency, and AI model complexity combine to shape effective cost per token. The resulting analysis surfaced *the underlying mechanics of a new AI economy—one where technical decisions directly dictate financial outcomes.*

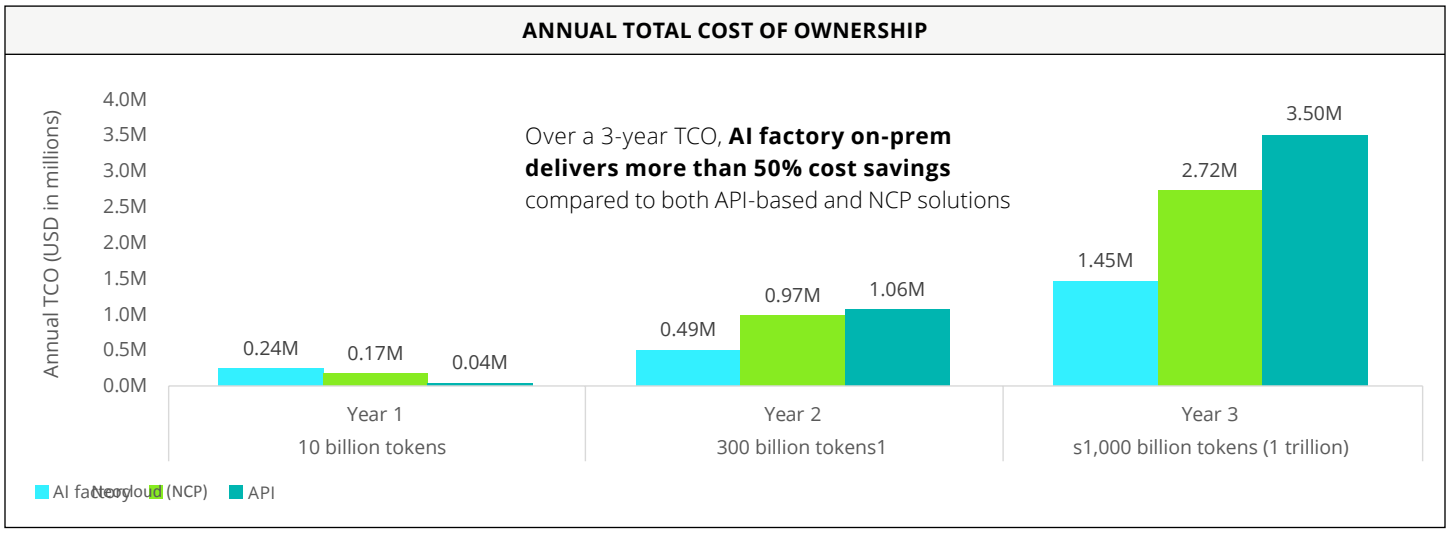
1. Usage scaling and complexity drives hosting advantage.

In our TCO modeling, the first year at 10 billion tokens, workloads favor the API access approach—pay-as-you-go approaches minimize idle capacity costs. As the number of tokens rises in year two, the economics flip. At higher reasoning loads more tokens are consumed, and self-hosted AI factories outperform APIs as fixed infrastructure costs are absorbed and utilization increases. After four years, the simulation projected cumulative TCO is twice the cost for API hosting as it would be for an AI factory, given the same configuration and token scaling (figure 4).

Figure 4. **Over 3 years, an AI factory is *~2.1x more* cost-effective than API-based solutions**

AI factory averages ~150% annual TCO growth vs. >1,000% (API) and >800% (NCP), ensuring more stable, predictable, and manageable costs

AI factory sees > 90% drop in $/B tokens from Y1 to Y3 ($24K to $1.45K) vs. 64% (API) and 84% (NCP), becoming most cost-efficient at high scale



Source: Deloitte simulation

Pay-as-you-go APIs and NCP are more suited to simple, low-volume workloads, while AI factory (self-hosted) is cost-effective for complex, high-usage, long-term needs.

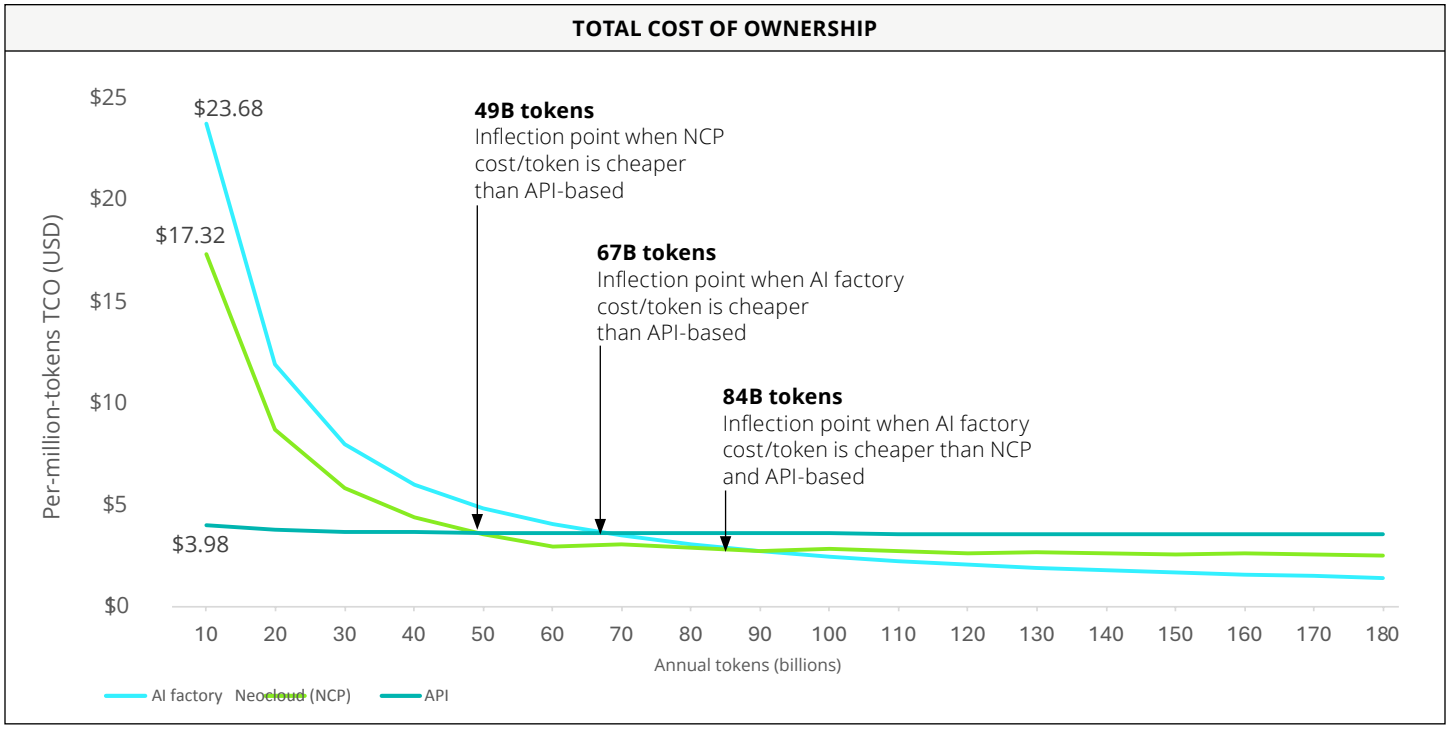
2. Scale changes the TCO equation—the inflection point is ~7 billion tokens per month.

For lower token volumes in the modeling, API costs scale linearly with use. As the number of active workloads grows, those same variable costs outpace fixed infrastructure. At scale, **AI factory** and specialized high-performance NCPs deliver stronger unit economics, especially for inference-heavy tasks that increase token consumption (figure 5).

Figure 5. AI factory becomes most cost-effective at ***84B tokens per year***

>90% drop in TCO per token for AI factory as overall usage scales vs. NCP (85%) and API (11%), indicating significant cost-savings opportunities at scale

Once annual usage exceeds 84B tokens, AI factory consistently reflects lowest TCO, with API costs scaling linearly with usage and NCP heavily dependent on GPU scaling and average usage



Source: Deloitte simulation

AI factory's upfront investment delivers greater cost advantages as usage scales, outperforming API and NCP options.

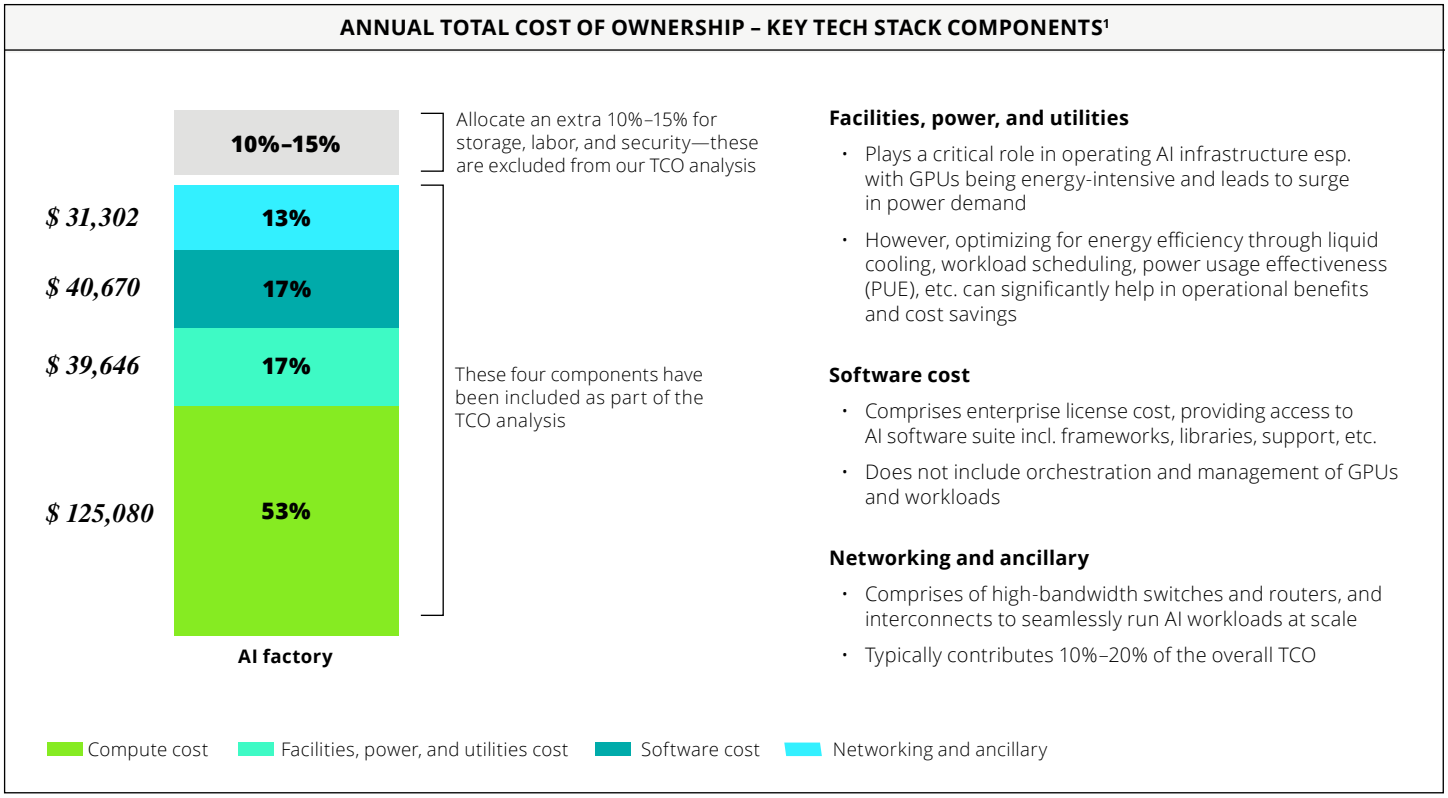
The 67 billion tokens were derived by taking the total capital cost of the systems required to service the annual token load for each consumption type (e.g., model API, NCP, on-prem) plus the annual operating expenses and dividing that combined cost by the total token load to determine per-token TCO. Token growth was then simulated, with 67 billion tokens observed as the inflection point.

Below this volume, proprietary API AI models remain cost-efficient in Deloitte's TCO modeling. Beyond it, self-hosted AI factories become structurally cheaper, with total costs crossing over at nearly **84 billion tokens annually**. This threshold holds across workload types, illustrating where ownership overtakes consumption in cost advantage.

3. The majority of AI factory TCO at 10 billion-token levels (53%) is compute costs.

How does the $1.45 million annual TCO modeled earlier break out for those considering running an AI factory at a 10 billion-token scale? Our modeling shows that compute costs comprise the majority of TCO with an almost equal distribution across facilities, software, and networking costs thereafter (figure 6).

Figure 6. **~50% of the AI factory cost is attributed to factors other than GPUs**



Source: Deloitte simulation

AI factory on-premises demands notable up-front costs, but careful architecture design and optimization strategies across the stack can drive significant long-term benefits and cost-savings. This modeling assumed air-cooled systems for on-prem, which is a likely scenario for organizations retrofitting existing on-prem data centers for AI. Longer-term, power and utilities costs should also account for one-time liquid cooling setup (e.g., cage build, liquid cooling retrofitting, piping, chilling and distribution units [CDUs] implementation) and purchase based on need.

The takeaways

Across the analysis, the implications seem clear: At a small scale, API access approaches are likely priced to compete. As workloads grow in size and complexity, token dynamics shift—which appears to make self-hosted options more desirable depending on workload needs and business priorities.

- **Know your AI workloads:** Understanding, measuring, and prioritizing workloads determines the right infrastructure decisions—ultimately shaping whether investment should be capex or opex, hosted or consumed.
- **Understand your AI consumption scale:** Both current and future demand directly influence hosting strategy and AI model selection. As AI consumption scales, the effective price per token and total cost of ownership can change dramatically.
- **Don't get locked in:** Strategies evolve—design for flexibility. Consider building modular, hybrid architectures and refresh your AI strategy regularly so that business and financial decisions, not technical constraints, guide hosting choices.

AI economics are volatile, but not ungovernable. The strategic task is to understand how costs flow through the value chain—and to act now to manage cost dynamics.

Optimizing AI cost structures

While token prices may fall, aggregate AI spend will likely still rise as adoption expands and workloads become more complex. Leaders cannot flatten this curve, but they can manage it—by optimizing what is within their control, preventing overspending, and protecting the enterprise from runaway bills and lock-in.

Best of breed versus fit for purpose

One common driver of overspending is using AI that is far larger than the problem requires.[17] Frontier-scale AI models deliver versatility, but their token consumption is high. For domain-specific tasks, smaller or fine-tuned AI models can achieve comparable accuracy while consuming only a fraction of the tokens.

Infrastructure choices play a part. High-end GPUs are indispensable for large-scale training but may be excessive for lighter workloads such as anomaly detection or classification. Midrange GPUs or CPUs can provide cost-effective alternatives. Multimodel strategies—reserving high-capacity AI models for tasks that truly require them while routing other workloads to smaller or open-source models—offer another layer of protection.

Driving efficiency into tokens to prevent unnecessary token burn

Even when the right model is chosen, tokens can be wasted through poor design. Many AI agents overconsume by running long reasoning chains where simpler logic would suffice. Streamlined design—using decision trees, rule engines, or capped context windows—can help ensure tokens are spent only where they add value.

Algorithmic techniques reinforce this efficiency: Early stopping halts processing when accuracy thresholds are met, prompt truncation reduces context length, and compressive transformers preserve capability while limiting irrelevant token use. Multiple model optimization techniques can also be used to provide additional safeguards: Quantization reduces weight precision to shrink compute needs, pruning eliminates redundant parameters, knowledge distillation allows smaller AI models to replicate larger ones, and transfer learning enables efficient adaptation without full retraining.

Inference design is equally critical. Retrieval-augmented generation (RAG) can reduce the need for bloated context windows. Prompt engineering discipline can ensure inputs are concise. Response caching can avoid paying for duplicate queries. Batching improves throughput. Finally, using traditional deterministic models in conjunction with probabilistic reasoning models can keep token use proportional to task complexity.

Embedding operational discipline

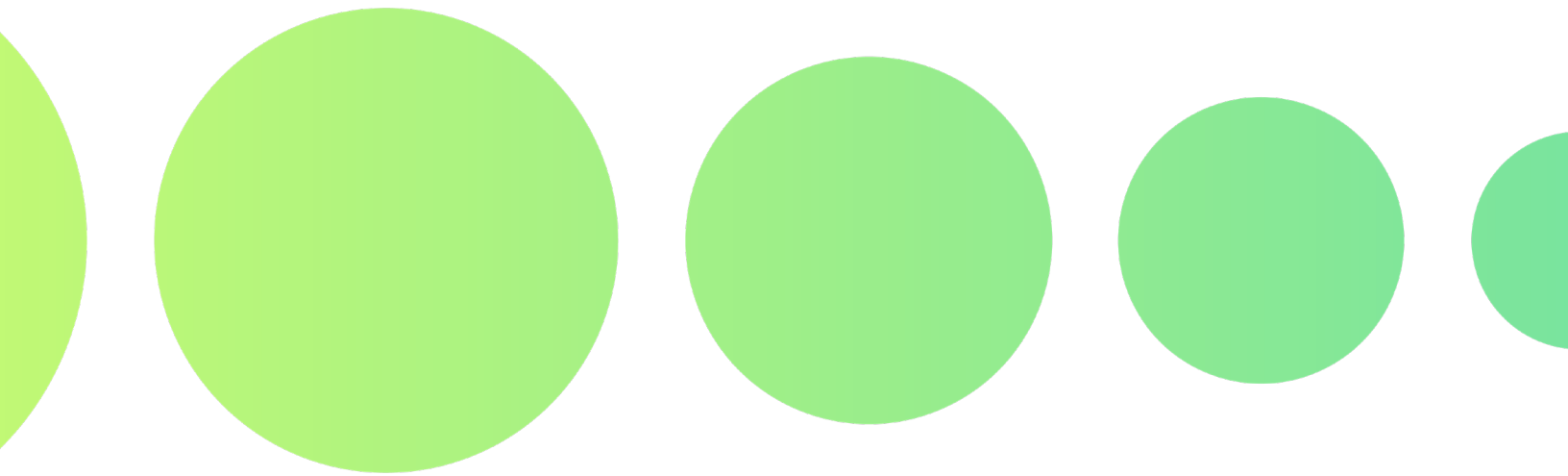
Token optimization is not enough without governance. AI spend can balloon silently across business units if not monitored and controlled. Leaders should bring the same financial rigor to AI that they brought to cloud computing.

Workload orchestration is one lever. Targeting ~85% GPU utilization confirms infrastructure is well used without running idle. Unified monitoring across GPU hours, storage, egress, and token use gives leaders real-time visibility into spend. Tagging enables chargebacks to business units, helping to prevent "shadow AI" costs from building unchecked, unsanctioned solutions.

Guardrails reinforce discipline. Budget alerts, context window limits, and API usage caps can help prevent runaway consumption. FinOps practices complete the loop—forecasting token demand, enforcing ROI thresholds, and approving only those projects that meet defined economic standards. The economics of AI will likely remain dynamic and complex. Enterprises cannot control market pricing or eliminate growth in token demand. But they can:

- Optimize what they use;
- Avoid overspending on oversized models or infrastructure;
- Protect against runaway token costs; and
- Guard against vendor contracts or technical decisions that limit flexibility.

By treating AI economics with the same rigor as energy or capital allocation, leaders can capture the benefits of AI adoption without losing control of spend. Those who fail to do so are not only likely to overpay but also risk being trapped in models and vendors that limit their strategic flexibility.



The C-suite imperative: Governing AI as a new economic system

AI should not be governed with the same cost models that guided prior enterprise technology waves. Traditional frameworks—total cost of ownership, per-user licensing, or static virtual machine pricing—were designed for predictable workloads and stable consumption patterns. AI is different. Workloads scale nonlinearly and consume resources at unpredictable rates, and costs are measured not in licenses or cores but in tokens.

Understanding tokens is not optional. They are the true unit of AI economics, the common denominator that reveals what organizations are paying for, how efficiently they are consuming it, and where value is (or isn't) being created. For many organizations this may require a paradigm shift: *CIOs should think like CFOs, and CFOs should think like CIOs*. The potential implications for C-suite leaders are profound. Without discipline, AI spend can drift upward quietly, hidden in traditional SaaS renewals, spiking in unpredictable API bills, or locked into infrastructure commitments that cannot be unwound. And unlike prior technology cycles, where budget overruns were frustrating but manageable, AI overspend can directly erode competitiveness.

This creates a new imperative for leadership: AI should be managed as an economic system, with tokens at its core. It means building a strategy that aligns consumption to value and adoption to discipline.

Insights that define this new system:

- Costs do not vanish—they migrate. Volatility will surface somewhere in the value chain, often through vendor pricing or licensing structures.
- Scale and buyer type shape economics. Smaller workloads may be a better fit for traditional SaaS companies or APIs; larger, predictable ones appear to favor owned infrastructure. Vendors face the same trade-offs—pass on cost or absorb it. The curve can be managed. Infrastructure optimization, workload design, and consumption discipline all bend the cost trajectory.

Token costs may outpace labor offsets. Simple use cases remain cost-sensitive; as agentic complexity grows, human-in-the-loop models can determine efficiency. This is where hybrid infrastructure and FinOps can come together as critical enablers.

Hybrid architectures provide the flexibility to run workloads where they make the most economic and strategic sense: sensitive data on-premises, elastic experimentation on cloud, and latency-critical inference at the edge. Building fit-for-purpose solutions, leveraging frontier models where truly required, and building smaller fine-tuned models/domain-specific agents can deliver equal outcomes at a fraction of the cost. Attacking inefficiency at its source by eliminating poorly designed agents or bloated prompts can eliminate wasteful tokens usage. Every token wasted is enterprise value burned.

The discipline cloud adoption forced through FinOps should be applied to AI. FinOps disciplines can provide the transparency to see, measure, and control token economics in real time. Together they do more than control costs; they can create the structural conditions for AI to scale responsibly and predictably. As adoption continues to scale, left unchecked, AI projects could proliferate across business units, workloads could scale beyond their original scope, and costs could balloon invisibly. Real-time monitoring of token use, budget alerts, chargebacks to business units, and ROI thresholds that projects must clear are not back-office exercises; they are financial guardrails that can help keep AI adoption sustainable. Without them, token consumption could grow faster than value realization—a formula for strategic failure.

The path forward is clear. Enterprises that do this well could be better positioned to scale AI with confidence, turning token consumption into measurable enterprise value. Those that do not could see costs spiral, contracts tighten, and flexibility vanish—just as AI becomes central to competitive advantage.

Token economics isn't a detail of AI strategy—it is the ***operating model***. Hybrid infrastructure and FinOps can help make it sustainable. The broader imperative is to govern AI with the same rigor applied to any other enterprise resource—capital, energy, or talent—with ***tokens as the new currency*** of value.

Authors

Nicholas Merizzi
Principal, Silicon2Service & AI Infrastructure US Offering Lead
Deloitte Consulting LLP
nmerizzi@deloitte.com

Tim Smith
Principal, Monitor Deloitte Strategy US Offering Lead
Deloitte Consulting LLP
timsmith6@deloitte.com

Diana Kearns-Manolatos
Senior Manager, Deloitte Center for Integrated Research
Deloitte Services LP
dkearnsmanolatos@deloitte.com

Nitin Mittal
Principal, Global AI Leader
Deloitte Consulting LLP
nmittal@deloitte.com

Gaurav Churiwala
Managing Director, Monitor Deloitte Strategy
Deloitte Consulting LLP
gchuriwala@deloitte.com

Special thanks to **Jason Chmiel** and **Ahmad Osman** for their contributions in writing this report and **Ahmed Alibage, PhD**, for his support throughout the process.

Explore more:

As cloud costs rise, hybrid solutions are redefining the path to scaling AI

Deloitte research reveals how rising cloud costs are prompting organizations to blend legacy systems with emerging solutions to scale AI more efficiently.

Continued reading

Deloitte's AI scenario modeling analysis and TCO

AI solutions, especially those leveraging Generative AI models, demand robust financial planning and scenario analysis to ensure sustainable deployment and operation. Our scenario modeling encompassed a comprehensive approach to model AI total cost of ownership (TCO) grounded in real-world data and based on critical analysis dimensions. The result is a structured approach for evaluating the economic impact of various AI deployment strategies and workloads.

Analysis dimensions

Understanding that AI has cost implications across the full tech stack, our AI TCO and scenario modeling tool factored three analysis dimensions:

- The LLM scenario—the type of model (open vs. proprietary).
- Hosting approach—inferencing location (on-prem, neoclouds, and API access).
- Workload scaling—increase in workload complexity, reasoning depth, or user count.

Modeling variability

The modeling is grounded in real-world data assumptions that are used to estimate the monthly and annual TCO across these scenarios. These include:

- **Token sizing:** Token estimation based on projected monthly input/output tokens, queries per month, user count, and capacity factors.
- **Workload and data sizing:** With an estimation engine that models token costs based on projected monthly input/output tokens, queries per month, user count, and capacity factors as well as sizing the data footprint based on data storage and inbound/outbound data volumes.

Additionally, the model is grounded in a set of assumptions based on secondary research, client engagements, and vendor pricing. Those general assumptions relate to costs such as electricity price in USD/kWh, PUE, colocation, inferencing costs, and GPU sizing (workloads, price per thousand tokens) based on different LLM models and GPU options. All the costs assumed in the analysis are for the US market.[16]

GPU and CPU footprint projections

Based on these inputs, our scenario estimates the GPU and CPU footprints—relative to the workload and configuration. The GPU footprint sizing is based on the initial GPU utilization as a percent, the number of tokens processed per second, and a calculation of the number of GPUs required to address the use cases being analyzed. Similarly, the CPU footprint is based on the number of CPUs needed to run the application and the estimated number of cores used for the workload.

AI TCO calculations

The analysis dimensions, initial configuration inputs, and CPU and GPU estimates come together to form the foundation for the AI TCO analysis. These values all contextualize and contribute to the AI TCO model that includes the following parameters per month in US dollars.

- **Compute costs:** Including the total cost of GPUs and CPUs. NCPs offer hourly rates as well as reserved contracting for different periods. In this model, we assumed hourly and not reserved pricing.
- **Inference costs:** Including input/output token costs per month.
- **Network and ancillary costs:** Including costs associated with networking and other ancillary spend such as on accessories.
- **Software costs:** Including monthly AI software and application costs.
- **Facilities and maintenance costs:** Including maintenance, energy and cooling, and space utilization. This modeling assumed air-cooled systems for on-prem, which is a likely scenario for organizations retrofitting existing on-prem data centers for AI. Longer term, power and utilities costs should also account for one-time liquid cooling setup (e.g., cage build, liquid cooling retrofitting, piping, chilling and distribution units [CDUs] implementation) and purchase based on need.

The following tech stack components were built into the model but excluded from the TCO analysis:

- **Storage and data egress costs:** Including costs associated with data storage and throughput (i.e., egress for scenarios with data leaving cloud).
- **Security costs:** Including monthly security and compliance costs, which were considered but removed from the model for simplicity.
- **One-time and staffing costs:** Associated with app innovation and integration to support AI enablement, GPU integration/solution activation (i.e., coding in hardware acceleration), and IT support staff costs to deploy solution.

While the report reflects average AI TCO relative to common hosting options, this tool ***analyzes many different configurations*** based on model approach and inferencing location to understand the pricing breakdown for current workload levels and to project those costs.

Endnotes

1. Tim Smith et al., "AI is capturing the digital dollar. What's left for the rest of the tech estate?," Deloitte, October 16, 2025.
2. George Fitzmaurice, "Cloud spending projected to grow 19% this year on back of strong 2024," ItPro, February 21, 2025.
3. Bram De Schouwer, Thomas Kessler, and Aurélien Descamps, "Cloud sovereignty: Succeeding in the evolving landscape," Deloitte, March 12, 2025.
4. Chris Thomas, Ganesh Seetharaman, and Diana Kearns-Manalatos, "AI workloads are surging. What does that mean for computing?," Deloitte, August 21, 2025. Survey of 60 data center executives and 60 power company executives, fielded March–April 2025.
5. Sundar Pichai, "Q3 earnings call: Remarks from our CEO," Google's The Keyword, October 29, 2025.
6. Smith et al., "AI is capturing the digital dollar. What's left for the rest of the tech estate?"
7. Steve Gallucci et al., "Finance Trends 2026: Navigating the expanded scope of finance," Deloitte Insights, October 6, 2025.
8. Steven Rosenbush, "AI economics are brutal. Demand is the variable to watch," Wall Street Journal, October 14, 2025.
9. Chris Thomas et al., "Is your organization's infrastructure ready for the new hybrid cloud?," Deloitte, June 30, 2025.
10. Thomas et al., "AI workloads are surging. What does that mean for computing?"
11. Eka Linwood, "Design parameters for data center facilities," STRUCTURE, January 1, 2023.
12. Frank Nagle, "Revealing the hidden economics of open models in the AI era," The Linux Foundation, November 19, 2025.
13. Greg Rosalsky, "Why the AI world is suddenly obsessed with a 160-year-old economics paradox," NPR's Planet Money, February 4, 2025.
14. LLM Token Cost, NVIDIA Token Calculator, accessed November 18, 2025.
15. Smith et al., "AI is capturing the digital dollar. What's left for the rest of the tech estate?"
16. Lambda, "AI cloud pricing," accessed December 12, 2025; CoreWeave, "GPU cloud pricing," accessed December 12, 2025.
17. Thomas et al., "Is your organization's infrastructure ready for the new hybrid cloud?"

Deloitte.