signal65

Governing AI Economics: Sovereign Infrastructure and the Token Cost Equation

AUTHORS

Brian Martin
AI Data Center Performance | Signal65

Matthew Goldensohn
Performance Analyst | Signal65

IN PARTNERSHIP WITH

DELL Technologies

JULY 2026

Executive Summary

The industrialization of enterprise AI has reached a critical inflection point. As organizations migrate from experiments to utility-grade production, agentic AI is consolidating around the AI Factory, a purpose-built stack of high-performance compute, low-latency networking, and AI-optimized software whose output is measured not in headcount but in token throughput. Deloitte infrastructure research finds that 73% of surveyed enterprise leaders expect AI Factory deployments at scale by 2028 and 72% expect AI at the edge at scale, roughly double late-2025 adoption. [1]

That scale introduces a new category of financial risk. In the era of "tokenomics," the token has superseded the seat license as the unit of digital cost, and unlike licenses, token spend is usage-driven, nonlinear, and volatile. Advanced reasoning models with "thinking tokens" and agentic workloads materially increase consumption; one healthcare enterprise growing 10% per month reached a trillion tokens within six months and $6 million in unplanned annualized cost before finance could identify the driver. Compounding the volume risk is price risk: retail token rates widely reflect subsidized introductory economics, and enterprises anchored to API pricing carry repricing exposure they do not control. [1,2]

Key Highlights

Governed Demand
Compute-Ready Documents enrich data once at ingestion and retrieve it repeatedly.

Owned Economics
Sustained inference on Dell and Broadcom infrastructure replaces volatile, subsidized API pricing with predictable amortized cost.

Sovereign by Design
PII/PHI and proprietary context stay behind the firewall; auditable retrieval grounds every answer in a verifiable path of truth.

Cost governance reduces to a single equation: AI spend = tokens consumed × cost per token. The first lever governs demand. TopicLake Insights, a data-enrichment platform from Gadget Software, transforms raw documents into Compute-Ready Documents (CRDs): enriched, governed digital artifacts built once at ingestion and retrieved repeatedly thereafter, replacing the pattern of reprocessing the same source material through inference on every query. Structured retrieval delivers grounded context at roughly 2 KB per query, cutting token demand structurally rather than behaviorally. [10]

The second lever governs price. Moving sustained, predictable inference onto an on-premises AI Factory — Dell PowerEdge XE7745 compute interconnected with Broadcom-based Dell PowerSwitch Z9864F-ON switches — replaces volatile usage-based pricing with amortized, owned capacity, with a modeled cost advantage of up to 18x at steady-state utilization. Data sovereignty binds the two levers together: proprietary context and personally identifiable information / protected health information (PII/PHI) stay behind the firewall, where retrieval is auditable and every answer is grounded in a verifiable path of truth. Signal65 benchmarking confirms on-premises solutions can deliver governed demand, owned economics, and sovereign data, converting AI from an unpredictable expense into an economically governed asset. [9,10]

The 2028 Enterprise AI Outlook

Utility-grade production infrastructure requirements are shifting toward a centralized model. A true AI Factory rests on three foundational pillars:

1. **Purpose-Built High-Performance Infrastructure.** A cohesive stack of compute, low-latency networking, and high-throughput storage.
2. **AI-Optimized Software.** End-to-end lifecycle management, from training and fine-tuning to high-velocity inference.
3. **Token-Based Output.** Intelligence delivered as measurable token throughput, the new raw material of digital business.

The urgency of this transition is evident in the “scaling gap.” While about 50% of leaders already manage 31 or more pilots, only 44% have reached that volume in production; by 2028, 67% expect to manage 31 or more production-ready use cases. The same survey quantifies the trajectory across both maturity and volume. [1]

Metric	2025 Status	2028 Projection
Respondents with 31+ AI pilots	50%	70%
Respondents with 31+ production-ready use cases	44%	67%
High-volume token consumption (>10B tokens/month)	30%	61%

Table 1: *2028 Scaling Trajectory [1]*

Bridging this gap requires a structural shift in how compute is consumed and governed, beginning with the equation that determines AI spend.

AI Economics: The Token Cost Equation

In the AI-native enterprise, the token has superseded the user license as the fundamental unit of digital cost — a shift the industry has begun calling "tokenomics." The economics reduce to a single equation: AI spend = tokens consumed × cost per token. Unlike seat licenses, both factors resist traditional budgeting. Consumption is usage-driven and nonlinear, rising with query volume, context size, and model reasoning depth rather than headcount. Price, meanwhile, is set by a market the enterprise does not control.

Demand Risk: Nonlinear Consumption and Thinking Tokens

Unmanaged reliance on third-party cloud APIs introduces significant forecast volatility. Advanced reasoning models generate "thinking tokens", internal processing outputs produced before a final answer, materially increasing consumption compared to standard models. The scale is already visible at the frontier: at Google I/O 2025, Google reported processing roughly 480 trillion tokens per month, a 50x year-over-year increase, while industry reporting indicates AT&T scaled its orchestration from 8 billion to 27 billion tokens per day. The financial consequences are stark. One large healthcare enterprise saw token consumption grow 10% per month; within six months it reached a trillion tokens, producing $6 million in unplanned annualized cost before the finance team could identify the driver. [1,2] [3,4]

Price Risk: Repricing Exposure and Subsidized Rates

The second factor carries a quieter risk. Retail token pricing today widely reflects market-entry economics: providers competing for enterprise workloads have priced inference aggressively, in many cases below the fully loaded cost of the underlying compute. For the enterprise, this means the per-token rate in the budget model is not a stable input, it is a promotional price set by parties with every incentive to raise it once workloads are embedded and switching costs are high. An organization can exercise perfect discipline over its token demand and still watch AI spend climb, because it controls only one factor of the multiplication. Volume discounts and committed-use contracts soften but do not remove this exposure; they fix a rate for a term against a price curve the enterprise cannot see. Price risk is therefore structural to the rented-inference model itself.

Four Immediate Risks for the CFO

Demand risk and price risk compound into four exposures that arrive on the CFO's desk in recognizable forms:

1. **Forecast Volatility.** AI spend behaves like a variable input cost. Mitigation: model token consumption explicitly (average tokens/user × user volume × cost/token) and set firm budgeting guidelines for scaling.

2. **Margin Leakage.** Costs are often hidden inside SaaS contracts or cloud bills. Mitigation: demand visibility into metered usage and stand up a FinOps function to scrutinize solution design and user behavior.

3. **Capital Timing Risk.** Reactive CapEx decisions arrive after token volumes are already committed. Mitigation: align infrastructure planning with business revenue drivers to secure proactive approval before infrastructure becomes the bottleneck.

4. **Earnings Narrative Risk.** Failure to tie AI spend to measurable gains weakens the investor story. Mitigation: govern AI with a dedicated function pairing technical telemetry with value-based KPIs such as revenue growth and cost-to-serve reduction.

These risks share a common root: both factors of the token cost equation currently float. Governing AI spend therefore requires pulling a lever on each — reducing the tokens an enterprise consumes, and stabilizing the price it pays for the tokens that remain. Unit economics vary sharply by venue, and at sustained volume the cost curves cross; but before the placement decision comes the demand decision. The next section begins where the largest and most durable savings live: the data itself.

Lever 1 — Governing Token Demand

Demand is where cost governance begins, because token demand is set by architecture, not by usage policy. In a traditional retrieval-augmented generation (RAG) pattern, the enterprise pays to reprocess the same source material on every query: raw text is retrieved in bulk, prompts bloat, and inference re-derives at query time what could have been established once at ingestion. Shifting from raw-text prompts to structured retrieval reverses the pattern; targeted context arrives at roughly 2 KB per query, cutting unnecessary token burn at the source. [10] And because the enrichment investment is made once while the savings recur on every query the enterprise will ever run, the reduction is structural rather than behavioral: pay once to structure the data, and every downstream retrieval inherits the discount. The leverage grows with agentic adoption: an agent that consults the document estate at every step of a workflow inherits the CRD discount, multiplying the savings by workflow depth.

The primary bottleneck in enterprise AI is the legacy document estate. Traditional RAG often starts by searching raw text, which can miss the semantic context required for mission-critical applications, searching strings rather than retrieving knowledge. TopicLake Insights from Gadget Software uses the CRD architecture to shift work from repeated inference-time reasoning to governed ingestion-time structure. Built once at ingestion, each CRD is a semantic twin, an enriched digital artifact in which security designation, PII masking, and lineage are embedded directly into the data object, so downstream AI systems retrieve governed knowledge rather than repeatedly reprocessing raw documents at query time. [10]

Three technical layers make this possible:

- **The Ontology Layer:** the cognitive map of the enterprise domain. It defines entities and their relationships, ensuring the AI understands the distinction between a concept (a “hot start”) and a literal string of text.
- **The Lookup Table (HITL Junction):** the human-in-the-loop (HITL) junction where subject-matter experts curate connections. Rather than managing raw text, SMEs manage a SQL-based lookup table, grounding the AI in an auditable path of truth rather than statistical probability.
- **Q2Q (Question-to-Question) Matching:** semantic intent captured by aligning user queries with pre-distilled FAQs. Because it matches intent to intent rather than query to passage, it delivers higher precision at lower compute overhead.

Transformative Impact

The result is shared semantic consistency across AI chatbots, BI dashboards, and mobile apps. In FAA turbine maintenance, for example, a query on “gas generator limits” joins the ontology to specific CRD topic IDs and returns the precise RPM limit, 38,100 RPM, to each enterprise endpoint from the same governed source, with citation lineage intact. Figure 1 traces that path from query to grounded answer. [10]

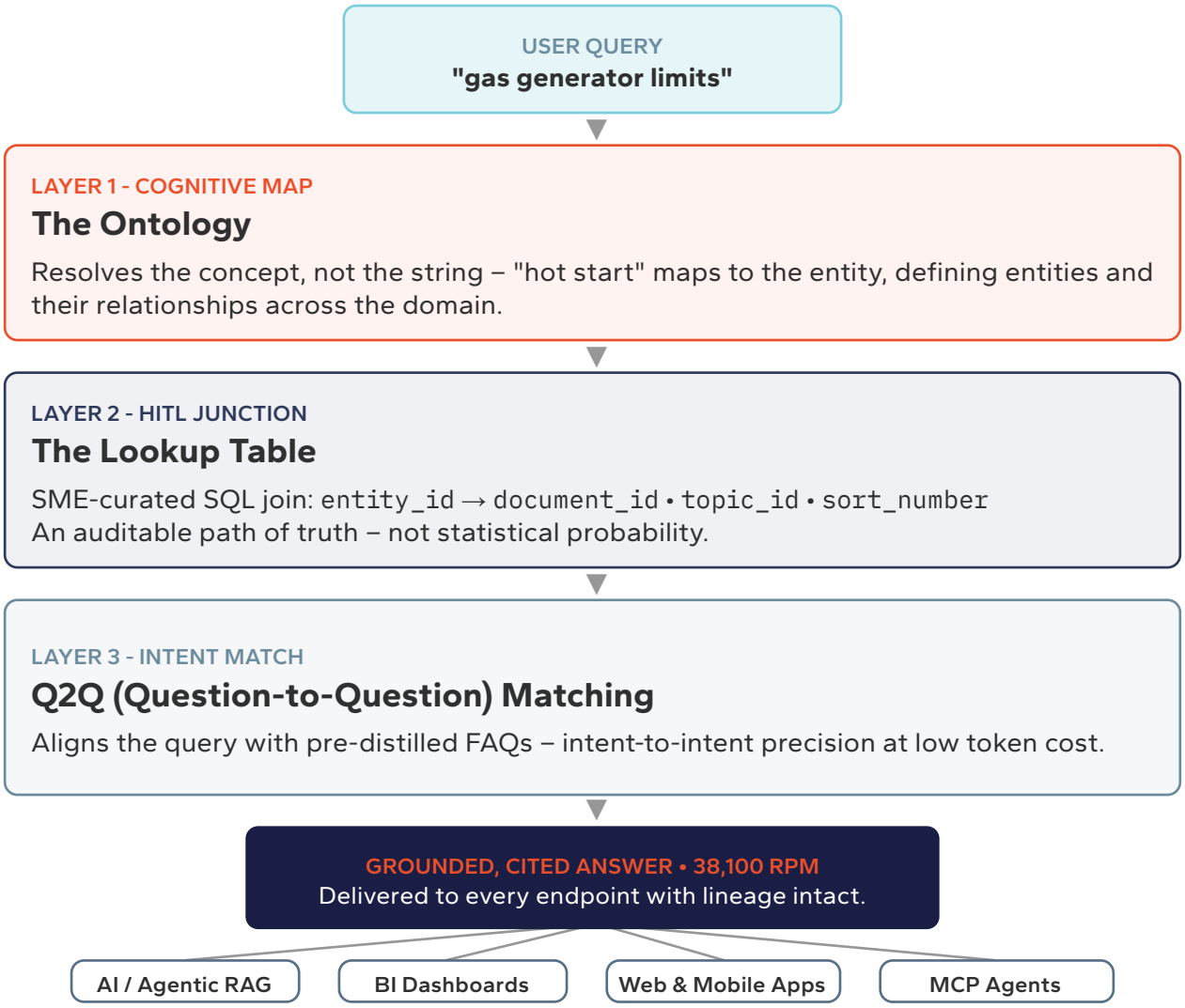


Figure 1: *The Compute-Ready Document Tri-Layer Architecture*

Universal Utility: One Substrate, Every Endpoint

Because intelligence is embedded at ingestion, the CRD substrate functions as a universal building block, consumed through the same governed source of truth across the enterprise:

1. **AI RAG / Agentic AI.** "Type boosts" prioritize technical keywords and FAQs, delivering the richest retrieval at minimal token cost (≈ 2 KB per query).
2. **BI Dashboards.** AI and dashboards read from the same materialized Postgres rows, keeping chatbot responses and analytics aligned.
3. **Web & Mobile Apps.** The sort_number field auto-attaches paragraphs split by page breaks, ensuring safety warnings and procedures are never fragmented.
4. **MCP (Model Context Protocol).** Semantic twins enable interoperable tool-use across heterogeneous AI agents.

Lever 2 — Stabilizing Token Price

Governed demand addresses how many tokens an enterprise consumes. The second lever addresses what each of those tokens costs — and that is a question of venue. Margin control depends on placing each workload where its unit economics are best, and as sustained volume rises, the optimal venue shifts from flexible cloud toward owned infrastructure.

Volume Tier	Recommended Placement	Rationale
Low volume	Cloud APIs	Flexibility at high unit cost; ideal for bursty demand
Midscale volume	"Neocloud" / GPU-as-a-Service	Improved unit economics for growing, semi-predictable load
Sustained high volume	Dell on-premises factory	Lowest unit cost and maximum margin control

Table 2: *Strategic Inflection Points for Workload Placement*

The tipping point is measurable. In the modeled scenario used here, workloads sustaining utilization above 20% can reach breakeven against cloud providers in as little as four months, with up to an 18x cost advantage for steady-state operation and zero egress fees. [10]

Economic Metric	Cloud Inference (OpEx)	On-Premises Factory (CapEx)
Unit costs	High retail token rates (2–3x wholesale)	Amortized hardware (3–5 year schedule)
Data transfer	Potential egress fees and data-transfer overhead	Zero egress fees
Cost advantage	Ideal for bursty, <20% utilization	Modeled up to 18x advantage for steady-state (>20%)
Forecasting	Volatile usage-based pricing	Predictable long-term TCO

Table 3: *Cloud Inference vs. On-Premises Factory [10]*

The economics of ownership run deeper than the unit-cost comparison. An owned factory converts the per-token rate from a market-set price into an internally derived one: amortized hardware, power, cooling, and support divided by delivered throughput. Every term in that calculation is visible to the enterprise and stable across a depreciation schedule — the direct answer to the repricing exposure described earlier. Price risk does not taper with volume discounts; it ends where the meter does.

The Platform: Dell Compute on a Broadcom Fabric

Capturing owned-infrastructure economics requires a platform that can sustain production token throughput without introducing new bottlenecks — because every idle GPU-hour raises the effective cost of every token produced. Utilization is the denominator of on-premises economics, and the platform is engineered to keep it high.

The Dell PowerEdge XE7745 provides the compute foundation: a high-memory, multi-GPU platform equipped with dual AMD EPYC processors, up to 3 TB of DDR5 memory, support for eight NVIDIA L40S, H100, H200, or RTX Pro 6000 GPUs, and eight Broadcom BCM57608 400 GbE network controllers per node, delivering the bandwidth and parallelism required for large-context inference and high-volume retrieval. [8,9]

At the fabric layer, Dell PowerSwitch Z9864F-ON switches — built on Broadcom silicon and running enterprise SONiC — interconnect the cluster over RoCEv2, delivering deterministic, low-jitter throughput across workloads. In an AI Factory, the network is not plumbing; it is an economic component. Congestion, retransmission, and jitter surface as idle accelerator time, and idle accelerators inflate cost per token. A lossless, low-latency Ethernet fabric keeps expensive compute saturated — which is precisely what the amortization math assumes. This balance of memory capacity, compute density, and network efficiency enables enterprises to deploy Agentic AI systems for the most demanding reasoning workloads. [9,10]

The same fabric choice protects the capital roadmap. Standards-based Ethernet scales from a two-node cluster to a multi-rack deployment without re-platforming, letting infrastructure grow in step with governed demand rather than ahead of it — the proactive capital posture that resolves the CFO's timing risk.

The Token Cost Equation: A Modeled Scenario

Cost governance reduces to a single equation: **AI spend = tokens consumed × cost per token.** The two levers examined in this paper act on different factors — Compute-Ready Documents reduce token demand, and the Dell and Broadcom on-premises factory stabilizes token price. To make the combined effect concrete, Signal65 modeled an illustrative enterprise document-intelligence workload. Figures are modeled, not measured, and are indexed to the baseline; customer-specific results depend on configuration, utilization, pricing, and workload mix. [10]

Modeled workload assumptions. One million retrieval-augmented queries per month against a governed document estate. In the baseline raw-text RAG pattern, each query retrieves roughly 40 KB of raw passage context (approximately 10,000 tokens at ~4 characters per token), plus ~300 tokens of instruction overhead and ~500 output tokens. In the CRD pattern, structured retrieval delivers the same grounding at roughly 2 KB (~500 tokens) per query, with instruction and output tokens unchanged.

Lever 1 — demand. Per-query input falls from ~10,300 tokens to ~800; total per-query consumption falls from ~10,800 tokens to ~1,300 — a demand reduction of roughly 8x. At workload scale, monthly consumption drops from ~10.8 billion tokens, squarely in the high-volume tier that 61% of enterprises expect to reach by 2028, to ~1.3 billion. [1] The critical property of this lever: the reduction is structural, not behavioral. Enrichment happens once at ingestion; every subsequent query inherits the savings.

Lever 2 — price. Applied to the tokens still consumed, venue economics compound the effect. At sustained utilization above the ~20% breakeven threshold, the modeled on-premises factory delivers up to an 18x unit-cost advantage over retail cloud rates, with zero egress fees. [10]

Scenario	Monthly token demand (modeled)	Cost index
Raw-text RAG on cloud APIs (baseline)	~10.8B	100
CRD retrieval on cloud APIs (Lever 1)	~1.3B	~12
CRD retrieval on Dell/Broadcom factory (Levers 1 + 2)	~1.3B	~0.7–2.4

Table 4: *Modeled Token Cost Equation, indexed to baseline = 100. The combined range reflects a conservative 5x venue advantage at the high end and the modeled 18x steady-state ceiling at the low end. [10]*

A sequencing note for the CFO. The two levers interact. Governing demand first shrinks the compute footprint that the factory must be sized against — and in some cases may move a workload below the on-premises breakeven threshold entirely. The disciplined order is therefore: govern token demand, measure the stabilized baseline, then size owned infrastructure against governed demand rather than ungoverned burn. Enterprises that repatriate first and optimize second risk purchasing capacity for token waste they were about to eliminate.

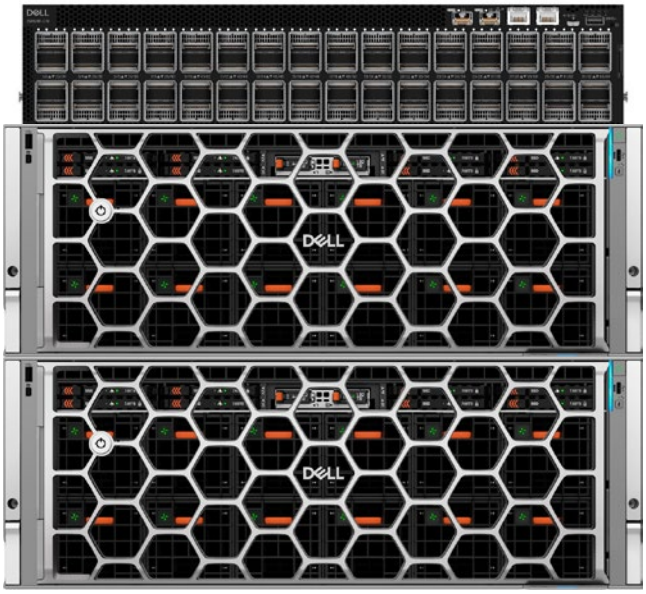
Performance Benchmarks: GPU Throughput and Pipeline Efficiency

Sustaining an AI Factory requires hardware that can absorb massive extraction pipelines without creating a throughput ceiling. Signal65 benchmarked document processing on a 1,385-page FAA technical document set, dense with tables and technical diagrams, using gpt-oss-20b served with vLLM, comparing NVIDIA RTX Pro 6000 Blackwell against the L40S across the full extraction-and-enrichment pipeline. Runtime and reliability results reflect this benchmark configuration and observed run conditions. [5,6,10]

Testing Configuration

Model: gpt-oss-20b **Serving:** vLLM **Document Set:** 1,385-page FAA handbook (tables + technical diagrams) **Hardware:** NVIDIA RTX Pro 6000 Blackwell vs. NVIDIA L40S on Dell PowerEdge

Test Environment



- Dell PowerSwitch Z9864F-ON
- SONiC 4.4.0

Cluster:

- 2 Dell PowerEdge XE7745
- Dual AMD EPYC 9555
- 2.3 TB DDR5 Memory
- 8 NVIDIA RTX Pro 6000 Blackwell Server
- 8 Broadcom BCM57608 400GbE RoCEv2 Network Controller
- Ubuntu 24.04 LTS
- CUDA 13.0 / Driver 580.95.05
- vLLM v0.10.2

RTX Blackwell vs. L40S: Optimized Pipeline

With vLLM deployment optimization, the Blackwell architecture posts strong gains in OCR-heavy extraction and a meaningful overall pipeline advantage. [6,10]

Metric	RTX Blackwell (Optimized)	L40S	Advantage
OCR extraction time	22m 22s	43m 18s	**1.94x**
Overall pipeline time (Document Set)	117m 51s	143m 8s	**1.21x**

Table 5: *Optimized Pipeline, 1,385-Page FAA Document Set [10]*

Across the full document set, that translates to roughly 11.75 pages per minute for optimized Blackwell versus 9.67 pages per minute for the L40S. [10]

Per-Document Stage Performance

The advantage is consistent at the per-document level across both pipeline stages.

Pipeline Stage	RTX Blackwell	L40S	Speedup
Extraction (OCR)	90.12s	144.25s	**1.60x**
Enrichment (LLM)	51.71s	84.31s	**1.63x**
Total time	141.83s	228.56s	**1.61x**

Table 6: *Per-Document Processing Time [10]*

The LLM Bottleneck

Despite the 1.94x OCR gain, the overall pipeline speedup is capped at 1.21x because corpus-level timing shows that non-OCR/artifact-generation work accounts for roughly 81% of optimized Blackwell runtime. Reasoning intensity, not extraction alone, now governs system performance, making artifact generation the primary target for future optimization. [10]

Energy Efficiency and Operational Reliability

Throughput vs. Energy

Scaling a factory means balancing raw performance against energy draw, and adding hardware does not yield linear efficiency.

Configuration	Wall-Clock Time	Energy Consumption
2 GPU	241 min	6.75 kWh
4 GPU	215 min	**6.59 kWh**
8 GPU	205 min	7.28 kWh

Table 7: *GPU Configuration, Throughput vs. Energy [10]*

Efficiency Sweet Spot

The 4 GPU configuration is the efficiency sweet spot: 11% faster than the 2 GPU baseline while reducing the absolute energy footprint from 6.75 kWh to 6.59 kWh. More hardware does not equal linear efficiency, the 8 GPU setup is fastest but spikes energy to 7.28 kWh. Organizations prioritizing raw speed can select 8 GPU; those optimizing for tokens-per-watt should target 4 GPU.

The Reliability Gap

For utility-grade AI, repeated timeouts, retries, rate limits, and service dependencies can become structural risks to safety, compliance, and predictable operations.

Environment	Observed / Potential Failure Mode	Operational Implication
Self-Hosted (Dell AI Factory)	**No observed failures in benchmark run**	Controlled / repeatable
Cloud API	Timeout, retry, rate-limit, or service risk	Variable / externally dependent

Table 8: *Extraction Reliability by Environment [10]*

A self-hosted architecture reduces external service dependencies and gives operators direct control over retry policy, capacity, and audit evidence, helping keep mission-critical extraction consistent, repeatable, and auditable. [10]

Data Sovereignty and the Dell AI Factory Blueprint

The principle of data gravity dictates that compute must move to the data, not the reverse. This "AI at the edge" posture keeps sensitive enterprise data sovereign while minimizing latency. Building on Dell Technologies lets organizations navigate the "Decision Stack", Compliance, Sovereignty, Latency, and Gravity, to place each workload correctly, with PII and PHI masked locally before any metadata is used and the cloud relegated to an acceleration lane for non-sensitive burst loads. [9,10]

From Risk to Resolution

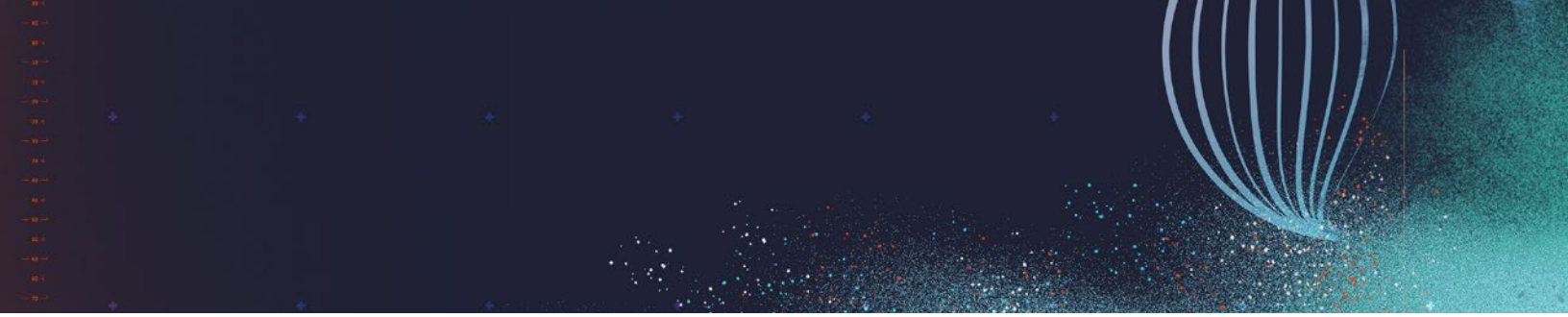
Dell Technologies with Broadcom networking and Gadget Software directly address the four CFO risks introduced earlier. [9,10]

CFO Risk	How the Sovereign Factory Resolves It
Forecast Volatility	Converts AI spend from an unpredictable variable cost into a manageable capital asset.
Margin Leakage	Reduces uncontrolled escalation of token costs embedded in SaaS and cloud-API contracts.
Capital Timing Risk	Replaces reactive purchasing with a predictable roadmap for GPU/CPU scaling.
Earnings Narrative Risk	Provides an auditable link between infrastructure investment and measurable KPIs.

Table 9: *From Risk to Resolution [9,10]*

Physical Infrastructure: Engineering the Cost per Token

The cost per token is ultimately engineered at the hardware layer, where Dell Technologies AI factory is designed to manage the 'Memory Wall' — the point where GPU memory capacity, memory bandwidth, interconnect, storage, and cooling become constraints on model reasoning and retrieval performance. Dell AI Factory deployments can be configured with air- or liquid-cooled server designs and high-speed Ethernet or InfiniBand fabrics, depending on workload, rack density, and enterprise standards. The executive point is not a single topology; it is selecting infrastructure that keeps model, retrieval, and artifact-generation pipelines close to governed data with enough memory and network headroom to avoid becoming the next bottleneck. [7,8,9]



Conclusion: Turning Volatility into Advantage

The path to 2028 requires treating AI as an economic system, not a technical project. Organizations that move from experimental OpEx to a durable, sovereign, Dell-powered AI Factory will convert AI's volatility into structural advantage: predictable economics, benchmarked throughput, observed operational reliability, and proprietary context that never leaves the firewall. Combining the processing power of Dell PowerEdge servers with the network efficiency of Broadcom connectivity, enterprises can create the foundation for adopting manageable AI capabilities as they mature. [9,10]

The discipline reduces to a single mandate: govern AI cost with the same rigor applied to capital allocation.

Cost Governance: The Hybrid-by-Design Roadmap

Audit token consumption, baseline monthly token burn across every SaaS and cloud-API contract, including "thinking token" usage.

Repatriate predictive workloads, move predictable, high-volume inference to Dell on-premises hardware to stabilize unit economics.

Implement hybrid-by-design, balance GPU-intensive OCR with optimized LLM artifact generation to relieve the gpt-oss-20b bottleneck.

Modernize the document estate, replace legacy PDF silos with a structured CRD architecture for reliable, audit-ready outputs.

The next three years will determine whether AI becomes a durable operating advantage or a structural cost burden. Those who govern tokens with capital-allocation discipline, and who own the substrate their intelligence is built on, will secure a lasting competitive edge.

Acknowledgments

The authors thank Gadget Software, Dell Technologies, Broadcom engineering teams, and the open-source vLLM community for their technical support and contributions.

Sources and Notes

External citations validate market, platform, and infrastructure claims. Signal65 benchmark and modeled TCO figures are noted separately so readers can distinguish third-party sources from Signal65 analysis.

[1] Deloitte, "Deloitte's enterprise AI infrastructure survey: A 2028 outlook," Deloitte Insights, March 30, 2026. Used for AI Factory, AI at the edge, pilot/production use-case, and token-consumption survey figures. URL: https://www.deloitte.com/us/en/insights/topics/technology-management/ai-infrastructure-survey.html

[2] OpenAI Help Center, "What are tokens and how to count them?" Used for the description of reasoning tokens and their use in billing and usage tracking. URL: https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

[3] Google, "Google I/O 2025: Sundar Pichai's opening keynote," May 20, 2025. Used for the 480 trillion monthly-token figure. URL: https://blog.google/innovation-and-ai/technology/ai/io-2025-keynote/

[4] VentureBeat, "8 billion tokens a day forced AT&T to rethink AI orchestration — and cut costs by 90%," February 26, 2026. Used for the AT&T token-volume example. URL: https://venturebeat.com/orchestration/8-billion-tokens-a-day-forced-at-and-t-to-rethink-ai-orchestration-and-cut

[5] OpenAI, "gpt-oss," GitHub repository. Used for gpt-oss-20b and vLLM serving references. URL: https://github.com/openai/gpt-oss

[6] vLLM Project, "GPT OSS — vLLM Recipes." Used for GPT-OSS serving and Blackwell configuration references. URL: https://docs.vllm.ai/projects/recipes/en/latest/OpenAI/GPT-OSS.html

[7] NVIDIA, "NVIDIA RTX PRO 6000 Blackwell Server Edition." Used for RTX PRO 6000 Blackwell Server Edition architecture, 96 GB GDDR7 memory, power, and air/liquid form-factor references. URL: https://www.nvidia.com/en-us/data-center/rtx-pro-6000-blackwell-server-edition/

[8] Dell Technologies, "Dell Technologies and NVIDIA unveil next-generation enterprise AI solutions," May 19, 2025. Used for Dell PowerEdge, air/liquid server-design, and RTX PRO 6000 platform references. URL: https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2025~05~dell-technologies-and-nvidia-unveil-next-generation-enterprise-ai-solutions.htm

[9] Dell Technologies, "Dell AI Factory with NVIDIA RTX PRO 6000 on Dell PowerEdge Servers," datasheet. Used for validated Dell AI Factory architecture, Dell PowerEdge + RTX PRO 6000 configurations, high-performance networking, and security/compliance positioning. URL: https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/dell-ai-factory-with-nvidia-rtx-pro-6000-on-dell-poweredge-servers-datasheet.pdf

[10] Signal65 internal benchmark results and modeled TCO scenario, May 2026, as summarized in this report's tables and benchmark narrative. Used for pages-per-minute, OCR and pipeline speedup, energy-consumption, observed reliability, CRD retrieval-size, and workload-placement claims. Customer-specific TCO should be validated against current configuration, utilization, pricing, power, cooling, staffing, and support assumptions.

Important Information About this Report

CONTRIBUTORS

Brian Martin
AI Data Center Performance | Signal65

Matthew Goldensohn
Performance Analyst | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65

Signal65 is a leading research organization specializing in enterprise AI infrastructure optimization and deployment strategies. Our lab focuses on evaluating and optimizing AI hardware and software solutions for real-world enterprise applications, with particular expertise in large language models, retrieval-augmented generation systems, and distributed AI architectures.

For more information, visit signal65.com
or contact research@signal65.com

IN PARTNERSHIP WITH

signal65

CONTACT INFORMATION
Signal65 I signal65.com