



EXECUTIVE SUMMARY

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

DELLTechnologies

JULY 2026

Overview

The industrialization of enterprise AI has reached a critical inflection point. As organizations move from experimental pilots to utility-grade production, the operating model is consolidating around the AI Factory: purpose-built compute, low-latency networking, and AI-optimized software whose output is measured in token throughput. Deloitte research finds 73% of surveyed enterprise leaders expect AI Factory deployments at scale by 2028.

That scale introduces a new category of financial risk. The token has superseded the seat license as the unit of digital cost and token spend is usage-driven, nonlinear, and volatile. Agentic workloads compound this growth, as single requests fan out into chains of inference calls. One Healthcare enterprise reached a trillion tokens 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, leaving enterprises anchored to API pricing exposed to repricing they do not control.

Cost governance reduces to a straightforward equation: **AI spend = tokens consumed × cost per token** and the full report pulls a lever on each factor.

Key Highlights



Governed Demand

Compute-Ready Documents enrich data once at ingestion reducing token consumption.



Owned Economics

Sustained inference on Dell and Broadcom infrastructure replaces volatile, subsidized API pricing.



Sovereign by Design

Proprietary data and context remain internal and auditable retrieval grounds every answer.

Two Levers on the Token Cost Equation

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.

The second lever governs price. Moving sustained, predictable inference onto Dell PowerEdge XE7745 compute with Broadcom-based Dell PowerSwitch Z9864F-ON networking 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 PII/PHI stay behind the firewall, where retrieval is auditable and every answer is grounded in a verifiable path of truth.

Benchmarked at Production Scale

Signal65 benchmarked the platform on thousands of pages in a technical document set, dense with tables and diagrams. Optimized software on local hardware delivered a measurable speedup in extraction over the previous generation, and the self-hosted pipeline completed extraction with no observed workflow failures. In the modeled scenario, the two levers compound: an approximately 8x structural reduction in token demand, multiplied by owned-infrastructure unit economics, compresses the modeled cost index from a cloud-API baseline of 100 to as low as single digits.

From Risk to Resolution

Together, Dell Technologies with Broadcom networking and TopicLake from Gadget Software directly address the four financial exposures that unmanaged token spend places on the CFO's desk.

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.

The Bottom Line

The path to 2028 requires treating AI as an economic system, not a technical project. Organizations that govern token demand at the data layer and token price at the infrastructure layer convert AI's volatility into structural advantage: predictable economics, benchmarked throughput, observed operational reliability, and proprietary context that never leaves the organization. The discipline reduces to a single mandate: govern AI cost with the same rigor applied to capital allocation. Those who do, and who own the substrate their intelligence is built on, will secure a lasting competitive edge.

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

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

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



CONTACT INFORMATION

Signal65 | signal65.com