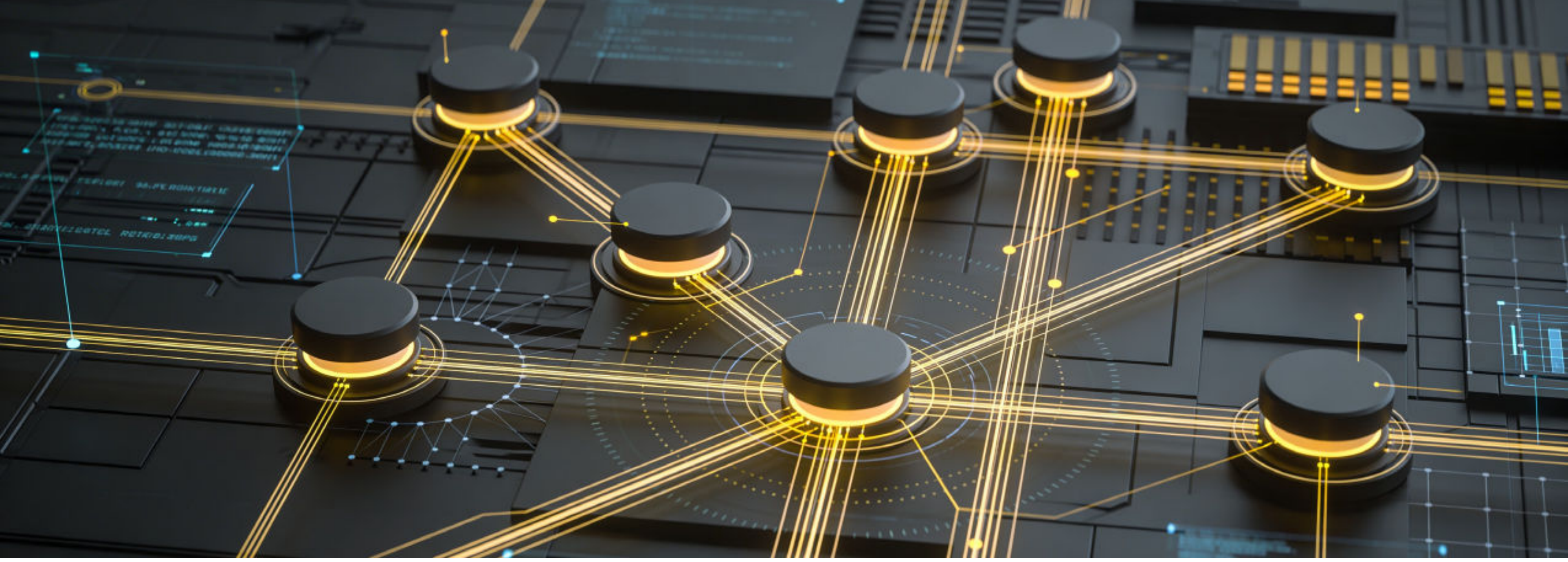
Futurum®

NEBIUS

From Proof of Concept to Inference ROI

Overcoming the Five Failure Modes of Production AI with Nebius Token Factory

MARCH 2026



Executive Summary

2026 marks an inflection point for enterprise AI adoption. The challenge now is operational: turning prototypes into production systems that perform reliably, scale predictably, and deliver sustainable economics.

Most organizations are still early in this transition. Only 13% of enterprises have moved beyond experimentation to fully transformed AI operations. Even AI-native companies often encounter challenges during the pilot phase, particularly as token consumption begins to shape the economics of their products.

Only 13% of enterprises have moved beyond experimentation to fully transformed AI operations

Five operational failure modes stop organizations from progressing from experimentation to production environments: lock-in, governance gaps, unpredictable costs, limited scalability, and quality degradation.

Addressing these issues requires infrastructure designed specifically for production AI workloads. Organizations increasingly need platforms that support both training and inference, provide integrated tooling for developers, and offer detailed visibility into inference usage and cost drivers so AI systems can be managed as production services rather than experimental workloads.

Nebius Token Factory is designed to address these operational requirements. By combining high-performance NVIDIA GPU infrastructure with inference frameworks focused on cost control, performance optimization, and model lifecycle management, the platform enables organizations to move AI workloads from experimentation into scalable production environments with visibility into performance and economics.

The Inference Production Gap

Futurum research shows only 13% of organizations have moved beyond experimentation to fully transformed AI operations, revealing a meaningful divide between pilot enthusiasm and production-grade delivery.[1] When bringing AI applications to production, our surveys highlight five common failure modes that need to be overcome:

- **Proprietary Model API Lock-in:** Model providers used in proofs of concept abstract away the optimization levers that determine token economics. Third-party model access is the model capability with the fewest customers reporting high satisfaction in the entire AI platform vendor relationship, according to Futurum's survey, lower than pricing or customer support.[2] Enterprises want model portability yet find themselves locked into APIs that can degrade in performance and fail in production.

- **Governance Gaps Block Compliance:** Security and compliance have consistently been the leading CIO concerns with AI in each quarter of 2025.[3] As a result, internal legal security teams block AI applications due to concerns over data exfiltration or the lack of Zero Trust architecture within the AI vendor's platform. AI vendors lack the data governance to demonstrate compliance with common standards.

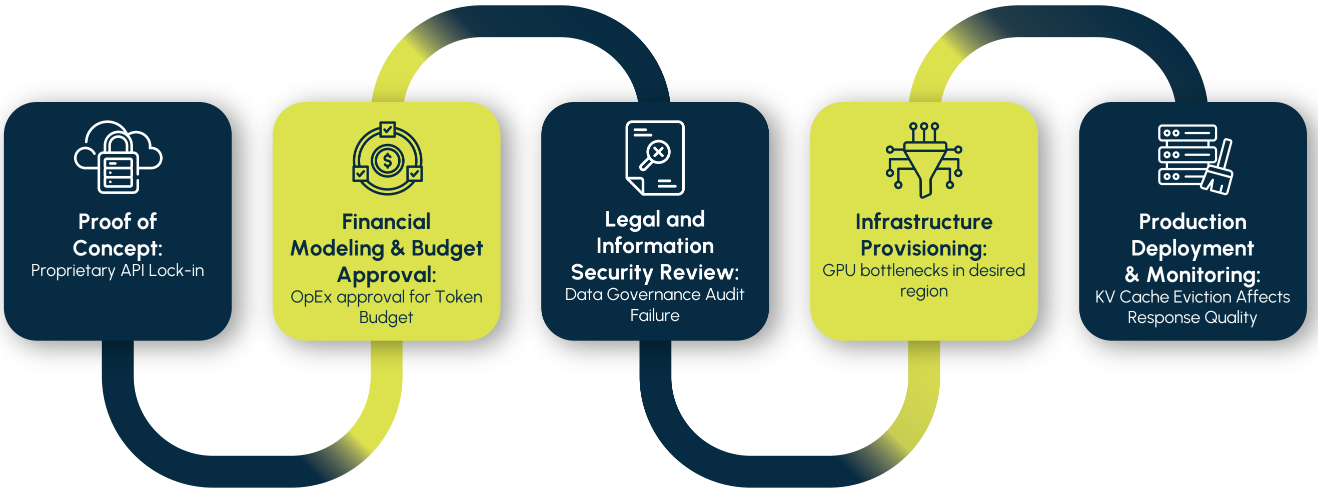
- **Unpredictable Cost Overruns:** Cost is an AI concern for 49% of AI platform decision-makers, making it the third-highest concern behind security and compliance.[3] Unlike training workloads with predictable GPU-hour budgets, inference costs are demand-driven and can compound without cost controls.[4]

- **Limits to Scale:** GPU supply is the leading scaling constraint among AI product and solution developers. Even organizations with optimized inference stacks and sound token economics hit a physical ceiling: there simply aren't enough GPUs available at the right time, in the right configuration.

- **Quality Degradation at Scale:** Production inference quality collapses when concurrent request volume overwhelms GPU memory bandwidth or KV cache management fails under load.

These hurdles can appear along the AI operations lifecycle, threatening to collapse successful proofs of concept (see Figure 1).

Figure 1: Where Failure Modes Occur in AI Inference Deployment



Proof of Concept: Proprietary API Lock-in

Financial Modeling & Budget Approval: OpEx approval for Token Budget

Legal and Information Security Review: Data Governance Audit Failure

Infrastructure Provisioning: GPU bottlenecks in desired region

Production Deployment & Monitoring: KV Cache Eviction Affects Response Quality

Source: Futurum Research

[1] - 1H 2025 AI Platforms Decision Maker Survey Report, Futurum Research, September 2025
[2] - 1H 2025 AI Platforms Decision Maker Survey Report, Futurum Research, September 2025
[3] - CIO Insights Global Survey, Q3 2025, Futurum Research, December 2025
[4] - State of the Market Report: AI Platforms, Q1 2026, Futurum Research, February 2026

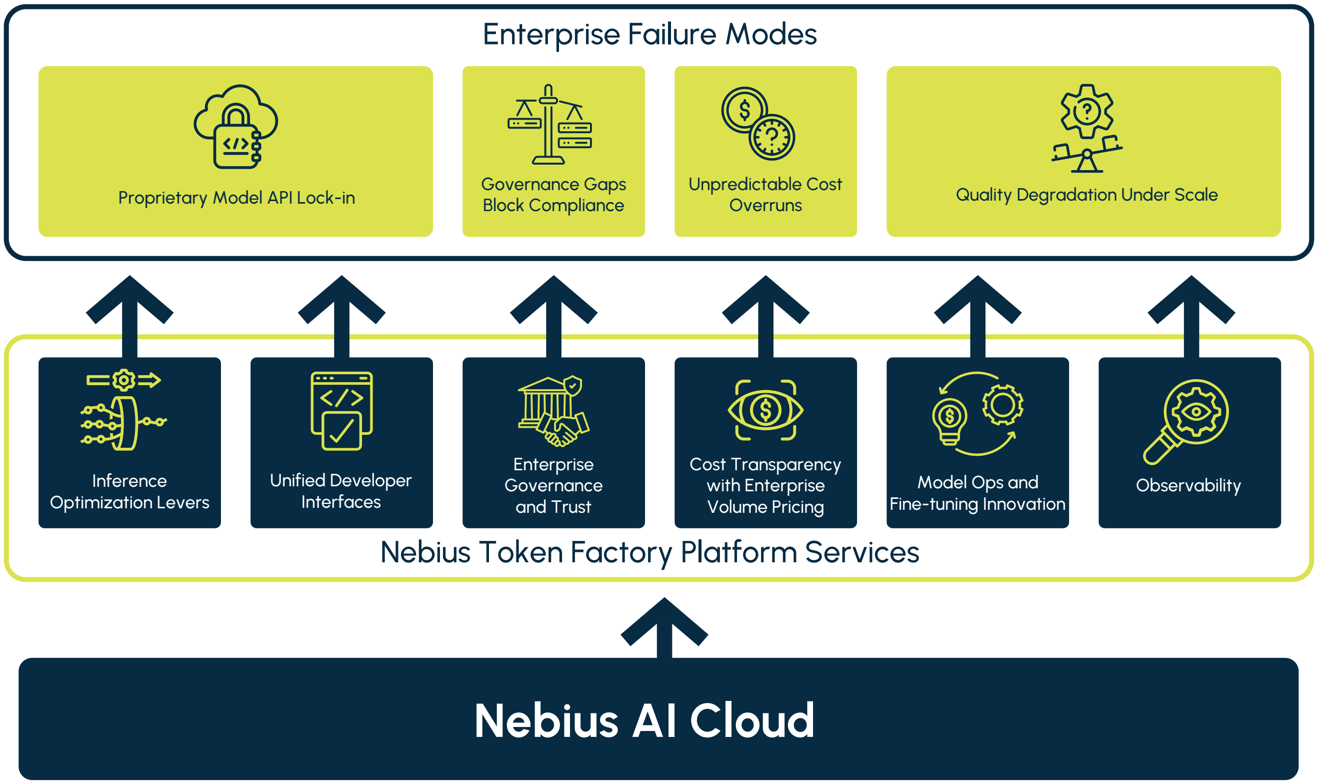
Nebius Token Factory Completes the Full Stack for Inference Deployment

Nebius Token Factory is a purpose-built inference platform that addresses all five failure modes with model behavior tuning, cost control, and a comprehensive set of platform services (see Figure 2). Token Factory delivers cost transparency through token-level metering, observability, and enterprise governance and trust with integrated compliance controls that satisfy regulatory requirements. A unified developer interface eliminates API lock-in with optimization levers for open-source models, including KV cache offloading, speculative decoding, and quantization.

All these services benefit from a vertically integrated platform built on Nebius AI Cloud that combines infrastructure designed for AI workloads across data centers, hardware, networking, storage, and cloud platform software while optimizing endpoints for cost and performance. In production, Token Factory runs on Nebius AI Cloud infrastructure that includes InfiniBand-connected clusters powered by NVIDIA GPUs, high-speed networking, and storage designed to support large-scale training and inference workloads.

Nebius combines gigawatt-scale, frontier infrastructure built on NVIDIA-accelerated computing with custom inference endpoints that adapt to engineering requirements. The platform transforms inference from an unpredictable cost center into a measurable production system that shapes model behavior and controls solution economics.

Figure 2: How Nebius Token Factory Addresses Common Failure Modes


Enterprise Failure Modes
Proprietary Model API Lock-in
Governance Gaps Block Compliance
Unpredictable Cost Overruns
Quality Degradation Under Scale
Inference Optimization Levers
Unified Developer Interfaces
Enterprise Governance and Trust
Cost Transparency with Enterprise Volume Pricing
Model Ops and Fine-tuning Innovation
Observability
Nebius Token Factory Platform Services
Nebius AI Cloud

Source: Futurum Research

Aligning Token Economics with Organizational Discipline

Scaling inference demands a new organizational discipline: treating model behavior as a measurable, manageable production system rather than an experimental black box.

The era of vibes-based AI is giving way to a phase of ROI measurement where tangible financial value now outweighs perceived productivity gains, as found by a recent Futurum survey.[5] Token Factory enables tracking of six operational KPIs that translate model behavior into business outcomes via state-of-the-art performance optimization techniques (see Table 1).

Table 1: Token Factory Operational Levers

KPI	Definition	Token Factory Lever(s)
Cache hit rate	The percentage of requests served from semantic cache, now a critical FinOps KPI that directly links infrastructure efficiency to the balance sheet.	▪ Cache-aware routing ▪ Embedding reuse
Cost per workflow	Total token spend per end-to-end task.	▪ Fine-tuning ▪ Token-based billing ▪ Workload-aware autoscaling
Latency	Average, variance, and tail latency measures user experience holds under load.	▪ Model distillation ▪ Latency SLAs ▪ TensorRT-LLM/vLLM/SGLang
Input/output token mix	The ratio that reveals whether models are under-utilizing enterprise context and/or generating too many reasoning tokens.	▪ Inference stack tuning ▪ Prefill-decode disaggregation
Tokens per task	Total number of input and output tokens consumed to complete a single discrete unit of work.	▪ Batch inference ▪ Quantization ▪ Speculative decoding
Quality and uptime validation	Continuous monitoring for degradation and SLA compliance.	▪ A/B tests ▪ Model selection and routing ▪ Observability ▪ Sampler correctness and validation

Together, these metrics create a shared language between engineering, finance, and product teams, enabling the organizational behavior change required to cross from pilot to production. Combining open-weight models with Token Factory services gives builders direct control over the cost-performance curve, turning each of these levers from engineering abstractions into actionable optimization surfaces that can serve millions of users while maintaining healthy margins.

[5] - The "Hard ROI" Pivot: 1H 2026 Enterprise Software Decision Makers Survey, Futurum Research, February 2026

Strategic Implications

As in the cloud-native era, inference ROI will ultimately be determined by platform scale. Just as cloud computing shifted from experimental virtual machines to production-grade platforms that defined winners and losers, inference infrastructure will bundle services that align with the business yet adapt to AI's architectural evolution. Three strategic imperatives now define the platform selection decision:

1

Achieving compelling unit economics and profit margins for company growth.

Without token-level cost transparency and optimization controls, margins erode faster than AI-native products can generate revenue. Tokenomics will come under review from internal finance leaders; builders will need to respond with accurate forecasting and cost-saving levers to serve millions of users while maintaining healthy margins.

2

Future-proofing infrastructure for next-generation models, hardware, and software optimization.

The market is entering an era of structural model fragmentation with competitive open-weights models from multiple providers permanently altering the landscape. Enterprises will manage mixed model portfolios, making infrastructure that supports heterogeneous model deployment and continuous optimization a core requirement.

3

Aligning with emerging standards and regulations while preparing for stricter board oversight.

As agentic AI introduces autonomous decision-making, the governance gap widens. Data security, information leaks, and privacy risks remained the leading AI concern for CIOs in every quarter last year, according to Futurum's survey. As with other IT shifts, deploying applications with data governance and privacy SLAs will become a board-level concern.

Nebius Token Factory, a vertically integrated inference-first cloud that addresses all three imperatives by providing immediate access to production-grade inference infrastructure with token-level observability, built-in optimization controls, and transparent pricing. The platform anticipates how inference will scale in the era of gigawatt-scale AI factories.

Click here to learn more about Token Factory.

Important Information About This Report

AUTHORS

Daniel Newman
CEO | The Futurum Group

Brendan Burke
Research Director, Semiconductors, Supply Chain & Emerging Tech | The Futurum Group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group 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 "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

DISCLOSURES

The Futurum Group 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 NEBIUS

Nebius is the Ultimate AI Cloud that combines the performance of a supercomputer with the flexibility of hyperscalers to power AI production at scale. Founded around deep in-house technological expertise, Nebius brings a strong engineering culture rooted in designing and operating large-scale platforms that run reliably at global scale. The company serves AI builders and enterprises worldwide across industries including healthcare and life sciences, robotics and physical AI, financial services, media & entertainment, retail and many others.

ABOUT TOKEN FACTORY:

Nebius Token Factory unifies inference, data capture, post-training, and deployment so AI systems can continuously improve in production.

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION: **The Futurum Group LLC I futurumgroup.com I (833) 722-5337**