November 2025 | eBook

AI with Confidence: Why a Platform Approach Is Essential to Enterprise AI

NUTANIX

The New AI Reality

AI has already permeated the enterprise. Employees use generative AI built into productivity apps, developers experiment with generative models, and many organizations deploy large language models (LLMs) in the public cloud. These moves have broadened access to AI but haven't necessarily fulfilled the key goal of GenAI: high productivity with a measurable return on investment.

Despite widespread adoption of AI, the business transformation leaders envisioned remains elusive, and three-quarters of organizations haven't yet unlocked any real value from AI, says a report from Boston Consulting Group. Meanwhile, infrastructure costs are rising unpredictably, governance and sovereignty challenges are mounting, and boards are pressing CIOs and IT leaders with a pointed question: What's the return on investment?

Table of Contents

The ROI Stays Elusive

For IT, the stakes are high. AI promises to automate core tasks, unlock new business models, and transform customer engagement. But it also introduces the very real risks of spiraling costs, compliance exposure, and operational complexity that erode trust and predictability.

The problem isn't adoption, it's value.

- SaaS AI apps deliver localized gains but may not transform enterprise workflows and raise privacy concerns.
- Cloud-first LLM deployments expand capabilities but come with steep trade-offs of cost unpredictability, sovereignty concerns, and management complexity.

IT leaders are asking:

- How do we move from experimentation to enterprise-wide transformation?
- Can we control costs as usage scales?
- How do we deliver sovereignty and compliance with sensitive data?
- What happens when AI evolves from smart assistants to autonomous agents?
- Can we utilize current talent and staff without hiring new, expensive employees?

The reality: Adoption alone isn't enough.

To deliver sustainable business value, enterprises need enterprise AI—and that means AI that is observable, governed, efficient, and resilient like any other mission-critical workload. AI that can be depended upon and woven into intricate enterprise workflows and deliver real results.

This eBook explores the challenges IT leaders face today—and how a platform-first strategy enables them to not just get localized gains but leverage the transformative power of AI at the enterprise level securely, sustainably, and at scale.

Challenge 1: Enterprise AI Readiness

LLMs without enterprise discipline
LLMs are often deployed on infrastructure that wasn't built to support its scale or complexity. Without built-in automation, observability, and resilience, IT teams face mounting challenges as models move from pilot to production.

Sourcing LLMs from different geopolitical sources may be disallowed based on governance policies or the sovereign political climate of a respective country. The result can be inconsistent performance, limited governance, and difficulty integrating AI into the systems that drive real business value.

The next wave: agentic AI
AI is evolving rapidly from basic GenAI interaction that assists with tasks into agentic systems that interact autonomously with enterprise data, workflows, and even each other. Imagine dozens of specialized AI agents—handling IT service tickets, managing logistics, personalizing customer interactions—operating in parallel and in real time.

Without enterprise-grade orchestration and governance, these systems can overwhelm IT operations and amplify risk.

Infrastructure uncertainty
Even when enterprises train or fine-tune their own models, infrastructure planning is fraught with risk:

- Should inference run in the cloud for elasticity, on-prem for sovereignty, or at the edge for latency-sensitive use cases, or across any of those environments?
- How should enterprises design for agility in a fast-changing world with model proliferation, a developing regulatory landscape, and high rate of innovation across the infrastructure landscape?
- Enterprises struggle to balance GPUs, CPUs, and new accelerators amid power, cooling, and supply chain constraints.

Solution: A Platform Approach

- **Run AI as a first-class workload.** To deliver real ROI, AI must move beyond SaaS apps and cloud-only models to become a first-class enterprise workload—run across datacenter, edge, and cloud with the right balance of performance, cost, and compliance. For instance, a global bank can manage fraud detection models across continents with unified observability, minimizing downtime and proving compliance to regulators.
- **Enable agentic AI readiness.** Enterprises must be ready for agentic AI, autonomous systems that demand secure, modular orchestration and seamless scaling across datacenter, cloud, and edge to ensure responsiveness, efficiency, and control. A logistics company could coordinate warehouse robots, delivery vehicles, and predictive demand models in real time, using orchestration to minimize latency and improve efficiency.
- **Provide infrastructure agility and hardware readiness.** Scalable AI requires an adaptable infrastructure strategy that provides model and hardware agility, and rightsizes GPUs, CPUs, and emerging accelerators while managing power, cooling, and supply chain risks to stay flexible as technology evolves. A healthcare network could align GPU usage with AI imaging workloads, avoiding costly overprovisioning while ensuring diagnostic models always perform reliably.

Challenge 2: Governance, Security, and Trust

Sovereignty and compliance pressures
AI workloads often touch the most sensitive enterprise data—customer records, financial transactions, patient histories. SaaS and public-cloud-first approaches give enterprises limited control over where data resides or how it's accessed. The result is sovereignty challenges, growing compliance risk, and sleepless nights for CISOs.

Trust in the models themselves
Accuracy is only the starting point. LLMs can drift, hallucinate, or perpetuate bias if not continuously validated. Without governance pipelines, enterprises cannot ensure that outputs remain explainable, reliable, or compliant with regulatory frameworks.

Fragmented governance
Most AI lifecycles span disconnected tools for data ingestion, model training, deployment, and inference. This fragmentation makes it difficult for IT to enforce consistent policies or produce audit trails sufficient for regulatory compliance.

Solution: Sovereign and Secure By Design with Day 2 Operations

- **Enforce sovereignty by design.** AI success depends on strict data sovereignty— enforcing where data lives, how it's accessed, and managing compliance across hybrid environments amid rising regulatory and geopolitical demands. Unified platforms give IT end-to-end control over data residency, lineage, and access. For example, a European energy provider could design its systems so that consumption data stays within its region to satisfy GDPR, while still enabling advanced AI analytics.
- **Build in risk management.** AI infrastructure must embed risk management at every layer to provide secure access, real-time monitoring, and lifecycle controls to reduce exposure and build stakeholder confidence. A financial services firm could enforce RBAC policies and monitor real-time usage to prevent unauthorized algorithm deployments, reducing risk and reinforcing regulatory confidence.
- **Validate and monitor models continuously.** Enterprises must continuously validate LLMs against drift, bias, and hallucinations using monitoring, audit trails, and techniques like retrieval-augmented generation (RAG) to deliver accuracy, compliance, and trust. Platforms help ground outputs in trusted enterprise data, monitor drift, and detect bias. A pharmaceutical company could run continuous validation pipelines so that AI-driven trial analysis remains accurate, explainable, and compliant with safety regulations.

Challenge 3: Performance and Cost Predictability

The AI data gap

AI projects often run into a data gap, where traditional storage can't keep pace with GPU-hungry workloads. Training stalls and inference slows as GPUs wait idly for data pipelines to catch up. Beyond raw throughput, utilization suffers when inference is fragmented into isolated silos across the enterprise. Without a shared inference service and AI-ready storage, GPU resources can be underutilized, delaying outcomes and limiting scale.

Limited visibility

AI performance depends on compute, storage, and networking working seamlessly together. But in fragmented environments, IT teams often lack end-to-end visibility, which means bottlenecks remain hidden, inefficiencies multiply, and costs rise unpredictably. Without a clear view across infrastructure and data pipelines, IT leaders struggle to tune performance, control costs, or scale workloads with confidence.

Operational complexity

AI pushes Kubernetes® to its limits. On-premises, setup assumes deep expertise and command-line mastery, while managed services only mask part of the challenge. Once deployed, GPU scheduling, upgrades, and policy enforcement can quickly become a Day 2 burden—manual, inconsistent, and error-prone across datacenter, edge, and cloud. The result is delayed time to first inference, wasted GPU resources, and mounting operational debt.

Solution: Optimize and Tune the AI Lifecycle

- **Optimize AI storage.** AI success depends on fast, scalable storage that balances performance and cost while efficiently feeding GPUs and supporting RAG workloads across datacenter, edge, and cloud. AI-ready storage keeps GPUs fully utilized while balancing cost through tiered strategies. A retailer could run real-time personalization models on high-performance storage while archiving historical training data to colder tiers—driving cost efficiencies without slowing innovation.
- **Tune performance end-to-end.** Enterprises need end-to-end observability across models, infrastructure, and data pipelines to optimize resources, control costs, and deliver predictable AI performance. Unified observability connects infrastructure telemetry with model behavior, helping IT rightsize resources. A telecom provider could detect bottlenecks in network traffic models and shift workloads to edge nodes, delivering fast customer response while optimizing GPU usage.
- **Simplify Kubernetes lifecycle management.** Enterprises need streamlined Kubernetes lifecycle management to simplify operations, minimize lock-in, and scale AI consistently across hybrid and multicloud environments. Standardized operations and Day-2 automation can reduce sprawl and risk. A government agency could manage clusters consistently across datacenter and cloud to accelerate citizen-facing AI services without piling on operational debt.

The Platform Path to Enterprise AI

AI adoption is no longer the hurdle. The challenge is translating adoption into sustainable business value without runaway costs, compliance risks, or operational silos. A successful enterprise AI platform requires sovereignty, security, and observability that simply performs and adapts.

For IT leaders, the mandate is clear: Treat AI as a mission-critical workload. Govern it, observe it, secure it, and make it resilient, just like all other mission-critical applications.

A platform approach makes this possible by unifying compute, storage, orchestration, and governance across datacenter, cloud, and edge. With the right foundation, enterprises can:

- **Prove ROI** with efficient, observable deployments.
- **Build trust** with integrated governance and sovereignty.
- **Scale sustainably** across datacenter, cloud, and edge.
- **Future-proof** against rapid change with modular flexibility.

To learn how you can turn innovation into long-term enterprise value, go to nutanix.com/enterprise-ai

NUTANIX

info@nutanix.com | www.nutanix.com | @nutanix