NUTANIX

Executive Brief

The Paradox of AI for Banking CIOs: Why 95% of Pilots Never Reach Production… and What to Do About It

By Sean O'Dowd, Head of Global Financial Services Strategy & Solutions, Nutanix

Most major banks have an AI strategy. Most have a growing portfolio of pilots. Yet industry data tells a sobering story: research suggests up to 95% of those pilots never make it to production. The financial services industry calls it "Pilot Purgatory," and it has become a defining infrastructure challenge of 2026 and beyond.

The budget picture makes it worse. According to Capgemini's 2026 research, 43% of corporate and investment bank IT budgets are consumed maintaining legacy systems, while only 29% is allocated to transformative technologies. The result: more than 80% of executives report that new technology implementations are not consistently delivering anticipated revenue gains. Nutanix's 2026 Financial Services Enterprise Cloud Index finds 68% of FSI IT leaders view their current infrastructure as not fully ready to support AI workloads on-premises, even as 71% expect to be running more than five AI-enabled applications within three years.

For the CIO, every stalled pilots represents sunk capital, eroded board confidence, and competitive ground ceded to institutions that have solved the infrastructure problem.

The challenge is often not a shortage of ambition or innovative models. But instead is rooted in an infrastructure problem, and solving it requires a fundamentally different architectural approach.

The Inflection Point: From Experimentation to Industrialization

AI adoption in financial services has reached critical mass. By late 2025, 85% of global banks had integrated some form of AI into their operations, with the financial sector projected to invest $97 billion in AI infrastructure by 2027. But the nature of that adoption is shifting fast.

The conversation has moved beyond generative AI chatbots to Agentic AI or industry autonomous operating systems. These are reasoning engines that execute multi-step workflows to solve systemic inefficiencies:

- **Credit Underwriting:** Helping significantly accelerate decisioning cycle times without adding headcount, addressing the manual back-office bottleneck that caps loan volume growth.
- **Perpetual KYC Monitoring:** Shifting from point-in-time compliance snapshots to continuous, real-time risk assessment, managing regulatory exposure and the cost of periodic remediation cycles.
- **Security Operations:** Compressing the multi-day handshake between data gathering, risk modeling, and approval routing into near-real-time threat response, freeing analyst capacity for higher-order investigation.

These agents operate continuously, make autonomous decisions within strict guardrails, and facilitate comprehensive auditability across disparate data systems. For the CIO, this marks a shift from experimental to operational: AI can serve as a capacity multiplier, not a science project.

This level of operational intensity requires infrastructure designed for modern workloads, not the systems originally built for legacy batch processing or simple human-initiated queries.

What's Shaping the Pace of Adoption

Across the financial services sector, three patterns are increasingly visible from our conversations. First, the pace of AI progress within a financial institution can depend significantly on organizational risk tolerance and cultural appetite for experimentation. Institutions whose leadership takes direct ownership of AI strategy and accepts calibrated early-stage risk tend to move materially faster than those where decision authority sits primarily with legacy IT ownership models. But appetite alone is not enough; the institutions that convert it into momentum are the ones with the freedom to maneuver, to run models on-premises where data residency demands it, to train and fine-tune specialized models on proprietary data, and to choose or swap the underlying model without re-architecting the application. Notably, the institutions moving fastest tend to have infrastructure that gives them the flexibility to absorb that risk without compounding it. Increasingly, this flexibility is delivered by hybrid multicloud infrastructure rather than legacy single-stack. Experimentation can be less costly, and adaptability may be more achievable, when the underlying platform supports it.

Second, ROI measurement remains an open question; a 12-to-18-month window appears to be emerging as a common benchmark for moving from headline to demonstrated value, though timelines vary materially by use case. Token-level cost governance is emerging as a CIO-level concern as approaches vary widely across institutions. Here, infrastructure architecture itself is becoming a potential cost lever: institutions are building dedicated AI infrastructure, what is increasingly referred to as an 'AI factory' pattern, to manage predictability of fixed infrastructure economics against variable token economics.

Third, a question increasingly being asked is not whether to deploy AI agents, but how to safeguard them at scale. Recent industry signals appear to reinforce this concern. Anthropic's release of Claude Mythos, a frontier model so capable that Anthropic declined to release it for general use and instead route access to Project Glasswing, a critical-infrastructure coalition that includes JPMorgan Chase, amongst others, has reframed agentic AI risk from a theoretical governance question to an operational one for boards and regulators alike.

Mythos is one such signal; the March 2026 kinetic strikes on AWS facilities in the UAE and Bahrain are another. Together they have surfaced a question that infrastructure architecture, not model selection, is positioned to answer: where do agentic workloads run, who controls the substrate they run on, and how is concentration risk managed when a single provider, a single region, or a single failure mode can compromise an entire AI-dependent function?

Five Infrastructure Barriers Keeping Financial Institutions Stuck

A staggering 98% of firms struggle to move GenAI from lab to production. The barriers are structural, not strategic, and each carries a compounding cost in delayed time-to-value, duplicated spend, and regulatory exposure:

1. **Architectural Rigidity:** Legacy debt is a primary barrier to AI. Many institutions rely on 1980s-era code (COBOL/Perl) with stagnant update cycles. Research indicates one-third of legacy systems cannot support AI, risking significant operational limitations as computational demands scale.

2. **Data Sovereignty & Mobility:** Data gravity and silos can limit enterprise-wide AI. As workloads grow, the cost of moving data across edge, on-prem, and cloud environments compounds. Data preparation infrastructure alone is projected to increase initial AI project costs by over 30%.
3. **Identity Risk:** Security vulnerabilities are acute; in finance, machine identities already outnumber humans 80:1. Despite this, 68% of organizations lack the necessary identity controls to secure AI and LLM workloads effectively. Compounding this, Nutanix's 2026 Financial Services Enterprise Cloud Index finds 66% of IT executives encounter AI applications or agents deployed by employees in non-IT functions, and 86% believe unsanctioned AI use creates material business risk.
4. **Regulatory Compliance and Sovereignty:** Organizations subject to the Digital Operational Resilience Act (DORA) and the EU AI Act face increasing data residency requirements and decision documentation obligations. Nutanix's 2026 Financial Services Enterprise Cloud Index finds 79% of FSI IT leaders rate data sovereignty as a high priority or must-have, yet 62% currently run containerized AI applications in public cloud, a growing "Sovereignty Debt", residency and resilience requirements that institutions will likely need to address.
5. **The Scaling Gap:** Fragmented, non-interoperable "point solutions" create a technical patchwork that can impede scaling of enterprise-scale deployments.

The Trusted Stack: A New Industry Standard

Industry leaders are coalescing around a new architectural paradigm: the Trusted Stack. Emerging from practitioner and analyst dialogue at events like the 2025 Momentum AI Finance conference in New York, this framework embeds security, governance, and transparency into every layer of AI infrastructure, from data ingestion to model output.

The Trusted Stack operates across four layers:

1. Data (provenance, lineage, access controls)
2. Models (auditability, bias monitoring, version control)
3. Applications (identity management, permissioning, guardrails)
4. Observability (transparency, explainability, regulatory reporting)

Each layer reinforces the others. A model built on poorly governed data inherits its flaws, and observability bolted on after deployment may not meet regulatory expectations that are increasingly demanding auditability by design.

Engineering the Trusted Stack with Nutanix

Translating this framework into production reality requires hybrid AI architectures that bring compute to the data, rather than forcing sensitive financial data to the cloud. The Nutanix Cloud Platform solution provides this across four operational capabilities:

Data Control and Observability through the Nutanix Unified Storage solution and the Nutanix Data Lens solution provides a unified data service where AI workloads access data in native formats without costly transformations, while policy-based controls are designed to help restrict data to specific jurisdictions and integrated encryption helps support security posture across hybrid environments.

Governance and Validation through the Nutanix Enterprise AI (NAI) software delivers a centralized control plane spanning hosted provider models (OpenAI, AWS Bedrock), self-hosted open-source models, manage agent sprawl and organization-specific fine-tuned models through a Nutanix Agent Gateway. Capabilities include an immutable registry of validated models, pre-flight testing for accuracy before branch or trading floor deployment, token-based rate limiting for per-user cost governance, and standardized API endpoints that allow institutions to swap models without rewriting application code.

Security and Resilience through Nutanix Cloud Infrastructure and the Kubernetes® orchestration software provides hardened foundations designed to automatically remediate, including air-gapped "Dark Site" deployments, zero-trust microsegmentation via the Nutanix Flow solution, and identity-centric access controls. For agentic workflows, unified RBAC and audit trails extend to Model Context Protocol (MCP) server connections, governing how AI agents access enterprise tools and private data sources within auditable boundaries, a capability now delivered through the generally available Nutanix Agent Gateway in Enterprise AI 2.7.

Cost and Performance Predictability through the Nutanix Agentic AI solution offers a turnkey, validated full-stack solution with right-sized compute economics. A tiered approach runs lighter inference on standard CPUs while bursting heavier training workloads to cloud, providing opportunities for cost efficiency across the full AI lifecycle. The web-scale architecture allows institutions to start small and scale node-by-node, avoiding unpredictable cloud spend.

Nutanix Agent Gateway allows organizations to shift AI workloads between public hosted models and private self-hosted inference to avoid vendor lock-in, manage GPU scarcity, and predict token costs.

Key Takeaways for Financial Services IT Leaders

- **Infrastructure is the moat, not models.** As LLMs commoditize, competitive advantage shifts to data governance, legacy modernization, and integrated observability, the same shift strategic analysts describe when they note that advantage now comes from turning common models into uncommon moats faster than competitors can.
- **Sovereignty is a critical consideration.** DORA, the EU AI Act, NYDFS guidance, and emerging state legislation require infrastructure that is designed to support data residency requirements, provide auditability by design, and support air-gapped operations. Hybrid architectures are a path forward.
- **Trust is the foundational currency.** Trust embedded at every layer of the stack is what separates production AI from pilot purgatory.

The Narrowing Window

The gap between AI leaders and laggards in financial services is widening, and infrastructure is a dividing line. A small number of global banks hold a disproportionate share of AI patents and report billions in realized value, while many institutions still have challenges realizing tangible ROI from their AI investments.

The question is no longer whether to invest in AI infrastructure, but whether your current architecture can support what comes next.

Download the full report: Engineering the AI Trusted Stack for Financial Services to explore the complete reference architecture and implementation framework.

NUTANIX

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