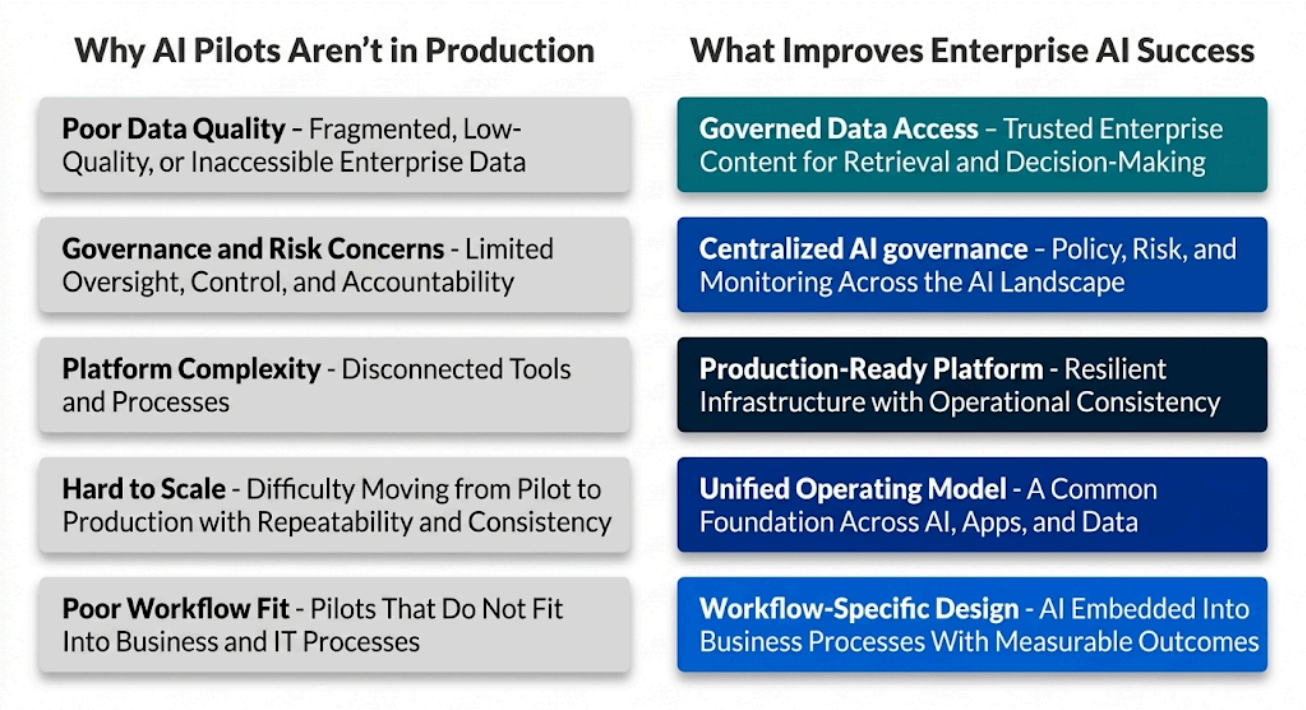
Enterprise AI Success at Scale

How an open AI operating model leveraging IBM Fusion HCI, Content-Aware Storage, and IBM watsonx can improve enterprise AI execution, governance, and time to value

This report integrates current public research with IBM's recent AI and infrastructure developments to evaluate how a unified, open operating model can improve enterprise AI execution, governance, and time to value at scale.

Executive Summary

The decisive factor in enterprise AI is execution, not experimentation. Large enterprises have access to numerous language models and no shortage of AI pilot activity. The harder challenge is establishing the data foundation, governance, operating discipline, and deployment architecture required to turn isolated initiatives into repeatable business value at scale. **Gartner's 2026 analysis showed that by the end of 2025, at least 50% of generative AI projects were abandoned after proof of concept** because of poor data quality, weak risk controls, escalating costs, or unclear business value.



The challenge is consistent with broader market evidence. McKinsey's November 2025 research found that AI use is now widespread, but nearly two-thirds of organizations still had not begun scaling AI across the enterprise. **BCG's September 2025 study found that only 5% of firms were "future-built," while 60% were seeing little material value despite significant investment.**

Deloitte's October 2025 analysis found that many organizations were generating ROI from AI, but that value was often fragmented because legacy systems, skills shortages, security concerns, and weak strategy continued to slow enterprise adoption. MIT NANDA's July 2025 research reached a similar conclusion from a different angle: success came from deploying AI into specific workflows, measuring business outcomes, and using systems that improved through feedback over time, not from running more pilots.

For large enterprises, the practical requirement is an open AI operating model that supports multiple models, multiple data environments, multiple deployment choices, and centralized governance without forcing the organization into a single proprietary path. In that model, the production platform, the enterprise data layer, and the AI control plane all matter.

IBM Fusion HCI, Content-Aware Storage (CAS), and watsonx align to those needs: Fusion HCI provides the OpenShift-based production foundation, CAS improves access to governed enterprise context for retrieval and agentic workflows, and watsonx provides open model access, hybrid data, governance, and orchestration. Fusion HCI also provides a unified OpenShift-based platform for AI, data services, modern applications, and virtualized workloads, with consistent operations, enterprise resilience, and stronger data control.

1 - Why Use IBM's Approach to Enterprise AI

IBM does not eliminate every enterprise AI challenge, but its emphasis on openness, governance, and production infrastructure makes it a credible foundation for on-premises AI deployment. That matters because large enterprises rarely need a single AI product; they need an operating model that can support multiple models, enterprise data, hybrid environments, and controlled execution at scale.

Enterprise AI outcomes depend on governed data access, workflow alignment, operational resilience, and a platform model that can scale. IBM brings together capabilities across Fusion HCI, Content-Aware Storage, watsonx, NVIDIA, and Red Hat AI that align with these requirements, including production-grade infrastructure, enterprise context management, open model access, governance, and orchestration.

Another advantage of IBM's approach is its modularity. These capabilities can be adopted together as an integrated architecture or used selectively alongside existing platforms, depending on enterprise requirements, modernization priorities, and governance needs. That flexibility supports a more practical path to AI adoption by reducing lock-in and allowing organizations to align deployment choices with business and operational realities.

2 - The Challenge Has Shifted from Experimentation to Operationalization

Most organizations no longer need to be convinced that AI can deliver productivity gains. The issue is scaling AI in a way that produces durable business value. McKinsey's November 2025 research showed that many organizations were already using AI regularly, yet most remained in experimentation or pilot phases rather than enterprise-wide scale. BCG's September 2025 study

showed the same gap between investment and realized value, with only a small leading group materially outperforming peers.

Deloitte's October 2025 analysis reached a related conclusion from the technology-estate side. AI initiatives were generating returns in many organizations, but those returns were often constrained by the surrounding environment. Legacy systems, limited internal expertise, security concerns, weak strategy, and budget tradeoffs continued to delay production adoption. In practical terms, many AI initiatives underperformed not because the models were inadequate, but because the enterprise environment around them was not yet built for scale.

MIT NANDA's July 2025 research sharpened that point by describing a "GenAI Divide": adoption was high, but transformation was low. The study found that generic AI tools were easy to trial, but task-specific enterprise systems often failed to reach production when they did not fit real workflows, did not retain feedback, and did not improve in context over time.

The study also found that successful customers typically demanded process-specific customization and evaluated AI by business outcomes rather than software benchmarks. That is an important distinction for enterprise architecture decisions: the winning pattern is not more experimentation, but a stronger operating model.

The implication for enterprise leaders is straightforward. AI value comes from combining workflow fit, governed data access, operational resilience, and clear accountability for outcomes. Gartner, McKinsey, BCG, Deloitte, and MIT NANDA all support that conclusion from different perspectives.

3 - Why Open AI Matters

Large enterprises rarely want an AI architecture that locks future decisions into one cloud, one model family, one inference runtime, or one governance boundary. An open AI operating model reduces that dependency risk.

It allows organizations to adopt AI now while preserving flexibility in model choice, deployment location, data access, and control. That flexibility becomes more valuable as model economics, regulations, and business priorities and technologies continue to evolve.

watsonx.ai Model Gateway enables access to IBM and third-party models through a unified interface. watsonx.data provides a hybrid, open data platform. watsonx.governance extends oversight across IBM and third-party AI environments. watsonx Orchestrate connects AI into enterprise workflows rather than isolating it in a chat interface.

Red Hat AI adds an open, hybrid inference layer for distributed production AI workloads. Fusion HCI strengthens this approach by supporting watsonx, Red Hat AI, and NVIDIA AI on a common platform, while allowing enterprises to use the models and accelerators that best fit each workload and keep enterprise data under organizational control.

For enterprise customers, the value of openness is practical rather than philosophical. It allows the organization to use IBM models where they fit, third-party models where they fit, and hybrid deployment models where governance, economics, or latency require them. That lowers dependency risk and improves the ability to adapt over time without re-architecting the entire AI estate.

IBM's Open AI Model Creates Partnerships

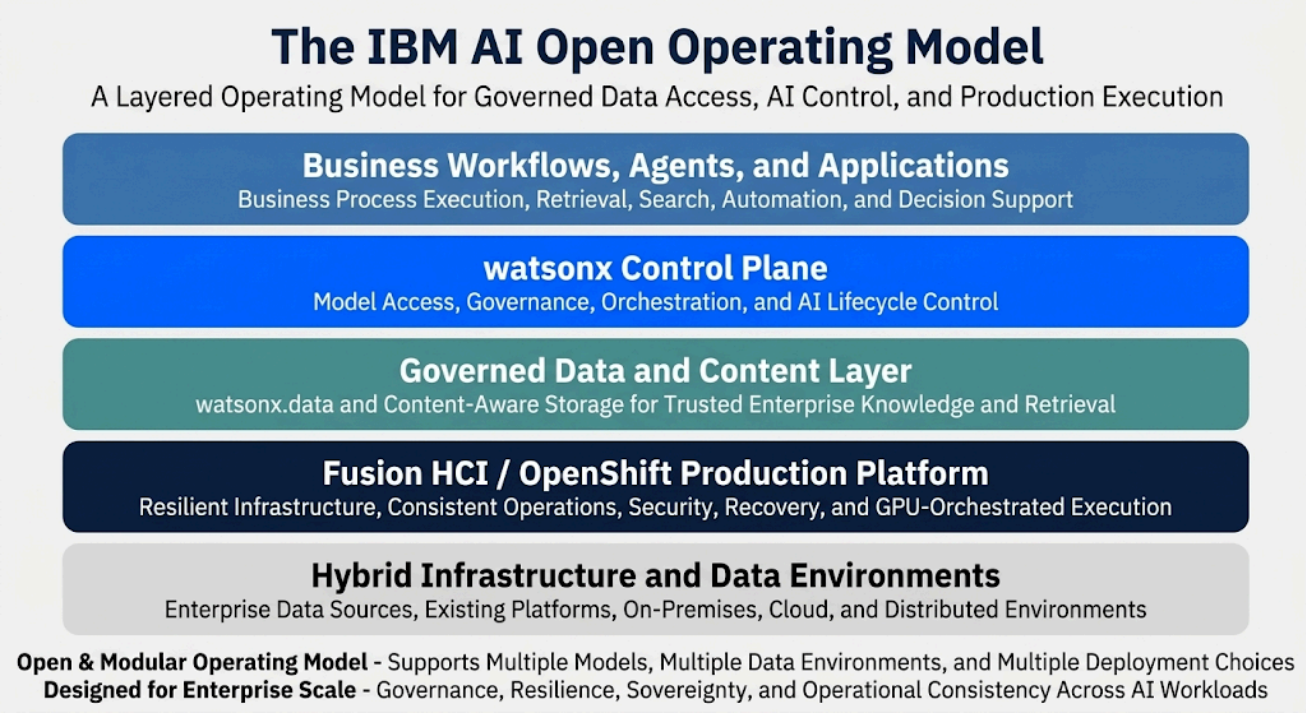
IBM's open AI position is not only about options, it' also about alignment across the layers that matter most in production. Red Hat reinforces the hybrid platform and operating-model foundation. NVIDIA reinforces the AI infrastructure layer. Anthropic adds value in a different way: as a strategic model and open-integration partner for enterprise AI.

In October 2025, IBM and Anthropic announced a strategic partnership to integrate Anthropic's models into select IBM software products with security, governance, and cost controls built in, beginning with Project Bob.

IBM also said it would contribute enterprise-grade assets to the MCP community, reinforcing openness not just at the model layer, but at the integration layer as well. For customers, the significance is practical: the operating model is designed to work across infrastructure partners, third-party models, and emerging open standards rather than forcing everything into a single vendor path.

Enterprises can pursue several paths to AI production, including cloud-native managed stacks, point-product AI platforms, and custom Kubernetes-based architectures. The reason to consider this model is not that those alternatives are invalid, but that a unified OpenShift-based operating model can reduce integration burden, improve governance consistency, and support AI, data services, and broader modernization on a common platform foundation. That is particularly relevant where the objective is not just to run models, but to create a governed, repeatable operating model for production AI.

4 - Why Production Infrastructure Matters



As AI initiatives move from experimentation to execution, infrastructure design begins to shape business outcomes directly. Systems that are acceptable for demonstration environments often fail to meet enterprise requirements for availability, governance, recovery, lifecycle control, and repeatable operations.

That gap is one reason many organizations remain stuck between pilot and scale. Gartner identifies the production transition as a core failure point, and McKinsey confirms that most organizations are still in early-stage scaling.

IBM Fusion HCI addresses that production gap by providing an OpenShift-based platform that combines compute, storage, networking, GPUs, backup, disaster recovery, cataloging, security, and multi-cluster operations into one operating model.

Fusion HCI supports AI, modern applications, data services, and virtualized workloads together, rather than forcing enterprises to build and operate separate stacks for each.

That matters because enterprise AI rarely exists in isolation. Retrieval services, model-serving tiers, APIs, workflow engines, data services, and adjacent applications all need to be operated, patched, secured, monitored, and recovered together. A fragmented infrastructure increases integration effort, extends timelines, and raises operational risk.

A unified OpenShift-based platform reduces those friction points and allows AI to run as part of the broader IT estate rather than as a separate infrastructure island. Fusion HCI also helps accelerate deployment, reduce operational overhead, improve resilience, and simplify support through a more integrated platform model.

MIT NANDA reinforces the same point from the business-outcome side. Its study found that successful customers typically chose systems that are integrated with existing processes and improved over time. It also found that external partnerships with customized, learning-capable tools reached deployment materially more often than internally built tools in its sample.

For enterprise architecture, that means pre-integrated production platforms are strategically important because they reduce the burden of assembly, integration, and workflow adaptation.

5 - IBM CAS for Retrieval, Enterprise Search, and Agentic AI

Many AI programs underperform not because the model is weak, but because enterprise content is fragmented, poorly governed, or difficult to retrieve efficiently. This is especially true for retrieval-augmented generation, enterprise search, and agentic workflows that depend on accurate access to policies, contracts, product content, research, case histories, and operational records. Gartner identifies data quality as one of the main reasons GenAI programs fail, and Deloitte highlights the drag created by disconnected information environments.

Content-Aware Storage directly addresses that challenge by strengthening the enterprise context layer. CAS makes unstructured enterprise content searchable and usable for AI by supporting knowledge-based retrieval, content tagging, cataloging, and RAG-oriented access patterns. It also

helps reduce the cost and complexity of moving large content sets away from their current locations.

For enterprise environments, that matters because the highest-value AI use cases usually depend on proprietary knowledge rather than public information. Accuracy, explainability, and trust improve when the retrieval layer is governed, connected, and aligned to how the enterprise already manages information.

CAS can therefore improve AI outcomes by reducing friction between enterprise content and enterprise AI, improving retrieval quality, and keeping knowledge under the same governance and security model as the broader platform. CAS also strengthens RAG, supports faster analytics across enterprise content, and improves data control in regulated environments.

6 - watsonx as the Enterprise AI Control Plane

A production AI environment needs more than inference. It needs coordinated control across models, data, governance, and execution. That is where watsonx provides the most value.

watsonx.ai: Model Choice Without Hard Dependency

watsonx.ai supports a multi-model strategy. IBM's public model-gateway approach gives enterprises flexible access to IBM and third-party models, which reduces switching friction and allows teams to choose the right model for each use case.

That flexibility is useful when cost, latency, governance, or domain performance differ by workload. On Fusion HCI, enterprises can support IBM and open-source models, prompt workflows, tuning, and model-serving patterns within a more consistent OpenShift-based operating environment.

watsonx.governance: Centralized Oversight Across Hybrid AI Estates

Openness without governance can quickly become sprawl. watsonx.governance helps enterprises govern models, applications, and agents across IBM and third-party environments, including cloud platforms and external model providers. Centralized control is increasingly important as organizations expand AI usage across departments and workflows. EY's October 2025 research supports this direction by showing that stronger responsible-AI practices are associated with better business outcomes.

Fusion HCI strengthens watsonx governance by providing a resilient platform foundation with encrypted data access and support for transparency, explainability, and operational control across AI workloads.

watsonx.data: Open Hybrid Data Access Instead of Forced Data Relocation

watsonx.data provides a hybrid, open data layer for agentic AI, retrieval, analytics, and governed data access. That architecture is valuable because it lets enterprises improve access across distributed data without forcing every workload into one proprietary store or requiring broad data relocation. The strongest use case is not replacing every data platform. It is improving governed access across the existing data estate.

watsonx.data helps organize company data, query across lake and database environments, control storage costs, and work with existing tools. When deployed on Fusion HCI, it also helps unify enterprise data access while supporting sovereignty, resilience, and cost control.

watsonx Orchestrate: Moving from AI Responses to AI Execution

Enterprise AI value increasingly comes from completing work across systems, not from generating a standalone answer. watsonx Orchestrate addresses that requirement by coordinating agents, workflows, and business systems.

watsonx Orchestrate is especially valuable for use cases that require real-time responsiveness, integration with mission-critical systems, governance, and support for multiple AI workloads. Fusion HCI improves reliability, security, and scale for those environments by providing fast governed data access, GPU orchestration, resilience, and support for sensitive enterprise processes.

Taken together, watsonx.ai, watsonx.governance, watsonx.data, and watsonx Orchestrate form a coherent control plane for enterprise AI. That control plane becomes more valuable when it runs on infrastructure built for production operations and connected to governed enterprise context.

7 – AI Challenges for Large Enterprises

Large enterprises face AI challenges that are structurally different. Their data is more fragmented. Their governance requirements are more demanding, their application environments are more heterogeneous, their risk tolerance is lower, and the cost of platform sprawl is materially higher.

Deloitte, EY, PwC, and MIT NANDA each highlight different aspects of the same enterprise challenge, including governance, skills, workflow fit, risk control, and operating complexity, all of which become harder to manage as organizations scale AI.

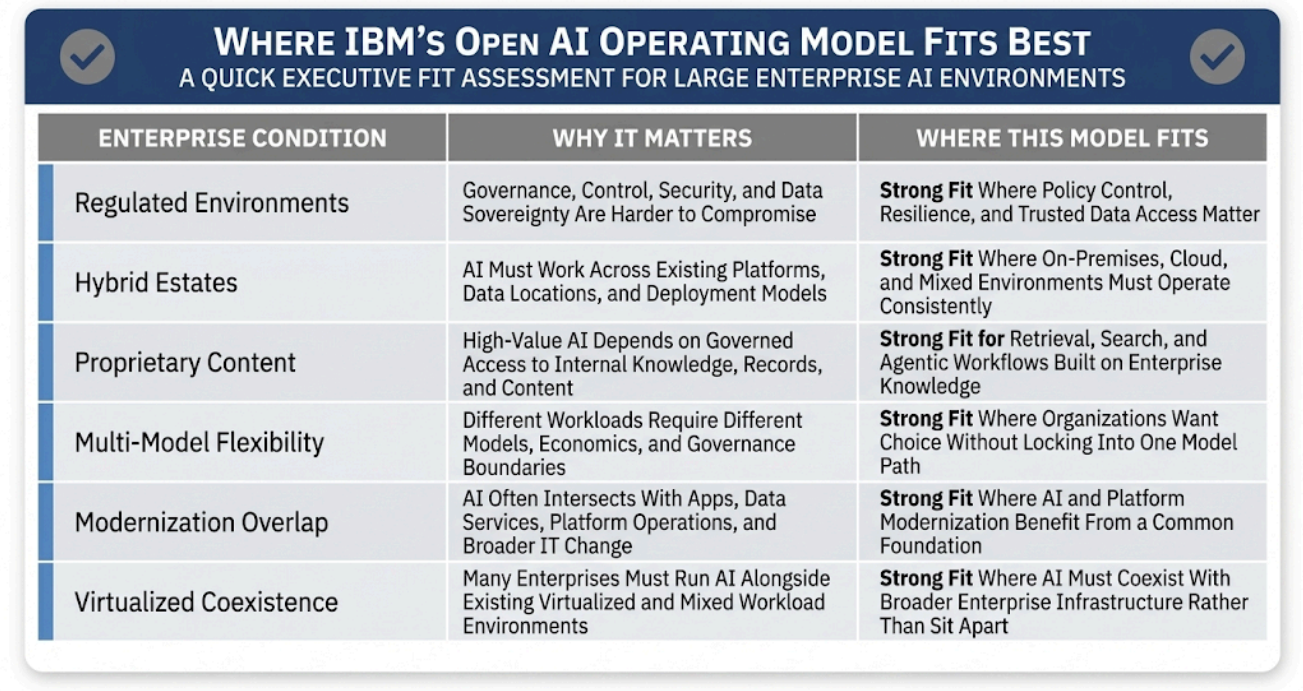
These enterprises also tend to need multiple AI patterns at once: retrieval and search over enterprise content, agentic workflow execution, model serving, hybrid data access, modernization of adjacent application estates, and in some cases coexistence with virtualized workloads.

Fusion HCI supports these requirements through a broader OpenShift-centered operating model, while the surrounding IBM and Red Hat portfolio can be adopted as an integrated stack or in selected combinations based on enterprise requirements, existing platform investments, and governance needs.

It can support workloads across watsonx, Red Hat AI, NVIDIA-related environments, IBM software, ISV software, and custom industry applications on a common platform foundation.

That broader platform perspective matters because large enterprises are not usually selecting a single AI tool. They are shaping an AI operating model that must live inside a much larger IT and risk architecture. An open approach built around OpenShift, hybrid data access, governance, and multi-model support is more aligned to that reality than a narrow point solution. Fusion HCI also supports sovereignty, faster time to insight, enterprise resilience, and AI stacks beyond IBM-only tooling.

8 - IBM's Open AI Use Cases



WHERE IBM'S OPEN AI OPERATING MODEL FITS BEST
A QUICK EXECUTIVE FIT ASSESSMENT FOR LARGE ENTERPRISE AI ENVIRONMENTS

ENTERPRISE CONDITION	WHY IT MATTERS	WHERE THIS MODEL FITS
Regulated Environments	Governance, Control, Security, and Data Sovereignty Are Harder to Compromise	**Strong Fit** Where Policy Control, Resilience, and Trusted Data Access Matter
Hybrid Estates	AI Must Work Across Existing Platforms, Data Locations, and Deployment Models	**Strong Fit** Where On-Premises, Cloud, and Mixed Environments Must Operate Consistently
Proprietary Content	High-Value AI Depends on Governed Access to Internal Knowledge, Records, and Content	**Strong Fit for** Retrieval, Search, and Agentic Workflows Built on Enterprise Knowledge
Multi-Model Flexibility	Different Workloads Require Different Models, Economics, and Governance Boundaries	**Strong Fit** Where Organizations Want Choice Without Locking Into One Model Path
Modernization Overlap	AI Often Intersects With Apps, Data Services, Platform Operations, and Broader IT Change	**Strong Fit** Where AI and Platform Modernization Benefit From a Common Foundation
Virtualized Coexistence	Many Enterprises Must Run AI Alongside Existing Virtualized and Mixed Workload Environments	**Strong Fit** Where AI Must Coexist With Broader Enterprise Infrastructure Rather Than Sit Apart

This operating model is most relevant for large enterprises that need governed access to proprietary enterprise content, flexibility across models and deployment locations, and stronger alignment between AI initiatives and broader platform modernization. It is especially applicable in regulated, sovereignty-sensitive, or hybrid environments where governance, resilience, and operational consistency matter as much as model choice.

It is likely to be a strong fit where organizations want AI, data services, modern applications, and virtualized workloads to operate on a more common platform foundation rather than as separate infrastructure silos. It is also well suited to enterprises that expect AI to support retrieval, agentic workflows, and business process execution rather than remain limited to isolated experimentation.

This model is less compelling where AI requirements are specific, SaaS-native, and self-contained; where cloud-native managed services are already the preferred operating model; or where the organization has both the capability and the appetite to integrate and operate a custom AI stack on its own. In those situations, the value of a unified OpenShift-based platform may be lower than the value of speed through a simpler managed-service path.

9 - What an AI Enterprise Value Case Looks Like

A credible enterprise value case is not built primarily on aggressive benchmark claims or isolated cost comparisons. It is built on measurable improvement in time to production, platform simplification, governance maturity, data accessibility, and reuse of a common operating model

across AI and modernization. Gartner, McKinsey, Deloitte, EY, and MIT NANDA all point to those factors as leading indicators of durable AI success.

For large enterprises, the value of Fusion HCI, CAS, and watsonx is most evident in five areas: faster movement from pilot to production through a pre-integrated OpenShift-based platform; better access to governed enterprise context for retrieval and agentic workflows; centralized control across model choice, data access, governance, and orchestration; lower operational complexity through a more unified platform model; and reduced dependency risk through support for open models, third-party models, hybrid data, and mixed deployment patterns.

This creates a stronger business case than claims based mainly on isolated benchmarks. It aligns more closely to how large enterprises evaluate AI investments: by their ability to deliver outcomes safely, repeatedly, and without creating a new layer of architectural debt. MIT NANDA reinforces this point by showing that successful customers emphasize workflow-specific customization, continuous learning, and measurable business outcomes over software performance claims alone.

Conclusion

Enterprise AI success now depends less on access to models and more on the ability to operationalize AI on trusted data with appropriate controls, resilient infrastructure, and a scalable operating model. The research is consistent on this point. Gartner identifies poor data, inadequate governance, cost escalation, and weak business-value discipline as primary reasons for failure.

McKinsey and BCG show that many organizations are still stuck between experimentation and scale. Deloitte, EY, and PwC show that governance, skills, and complexity remain central barriers. MIT NANDA shows that successful organizations focus on workflow-specific customization, continuous learning, and measurable business outcomes rather than isolated software performance claims.

An open AI operating model is therefore more than a technical preference. It is a practical strategy for reducing dependency risk while improving execution. For large enterprises, that means combining a production-ready OpenShift platform, governed enterprise data access, multi-model flexibility, centralized AI governance, and workflow orchestration into one coherent architecture.

IBM Fusion HCI, Content-Aware Storage, watsonx, and Red Hat AI support that direction through a modular architecture that can be deployed as a unified operating model or adopted selectively alongside existing enterprise platforms, depending on the organization’s requirements for control, interoperability, and modernization pace.

For large enterprises, the advantage of this approach is not only technical. It is operational and strategic: faster execution, better governance, stronger alignment with hybrid enterprise realities, and lower dependency risk across models, data platforms, and deployment choices. Those are the conditions that improve the odds that AI becomes a durable business capability rather than another pilot program.

About the Author

Armando Arias is an IT leader, advisor, and strategist focused on enterprise infrastructure, hybrid cloud, and AI infrastructure adoption at scale. His background includes large OpenShift deployments and IBM Fusion HCI solutions, supported by deep familiarity with IBM and Red Hat platform strategy, OpenShift operating models, and modern enterprise architecture.

He helps organizations evaluate and shape practical paths to AI production by addressing the core infrastructure requirements for success, including data readiness, governance, cost control, architectural flexibility, and operating discipline.

Armando Arias - Li9 Technology Solutions
President & CEO
armando.arias@Li9.com

Note on Research and Content

This report was developed by Armando Arias using current public research, direct analysis, and AI-assisted support. ChatGPT, GROK and Google Gemini were used to create images, refine structure, language, and presentation elements. All findings, interpretations, recommendations, and final decisions were reviewed and approved by Armando Arias.

Appendix: Sources Referenced

Sources, publication dates, and links are listed below for reference.

Source	Date	Title	URL / Notes
Gartner	January 26, 2026	Why 50% of GenAI Projects Fail — And How to Beat the Odds	https://www.gartner.com/en/articles/genai-project-failure
McKinsey & Company	November 5, 2025	The state of AI in 2025: Agents, innovation, and transformation	https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
Boston Consulting Group (BCG)	September 30, 2025	Are You Generating Value from AI? The Widening Gap	https://www.bcg.com/publications/2025/are-you-generating-value-from-ai-the-widening-gap
Deloitte	October 16, 2025	AI is capturing the digital dollar. What's left for the rest of the tech estate?	https://www.deloitte.com/us/en/insights/topics/digital-transformation/ai-tech-investment-roi.html
Ernst & Young	October 8, 2025	EY survey: companies advancing responsible AI governance linked to better business outcomes	https://www.ey.com/en_gl/newsroom/2025/10/ey-survey-companies-advancing-responsible-ai-governance-linked-to-better-business-outcomes

Source	Date	Title	URL / Notes
PwC	October 1, 2025	2026 Global Digital Trust Insights	https://www.pwc.com/us/en/services/consulting/cybersecurity-risk-regulatory/library/global-digital-trust-insights.html
KPMG	October 7, 2025	Global CEOs double down on AI and talent drive despite economic challenges / 2025 Global CEO Outlook	https://kpmg.com/xx/en/media/press-releases/2025/10/global-ceos-double-down-on-ai-and-talent-drive-despite-economic-challenges.html
MIT NANDA	July 2025	The GenAI Divide: State of AI in Business 2025	https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf
IBM	2025 public announcement	Access any model, anywhere on watsonx.ai (watsonx.ai Model Gateway)	https://www.ibm.com/new/announcements/watsonx-ai-model-gateway-in-public-preview
IBM	October 7, 2025	From orchestration to outcomes: New agentic workflows and domain agents in IBM watsonx Orchestrate	https://www.ibm.com/new/announcements/new-agentic-workflows-and-domain-agents-in-ibm-watsonx-orchestrate
IBM	December 10, 2025	watsonx.data v2.3: New innovations powering enterprise AI at scale	https://www.ibm.com/new/announcements/watsonx-data-v2-3-new-innovations-powering-enterprise-ai-at-scale
IBM	Current product information	watsonx.governance product page	https://www.ibm.com/products/watsonx-governance
Red Hat	October 14, 2025	Red Hat Brings Distributed AI Inference to Production AI Workloads with Red Hat AI 3	https://www.redhat.com/en/about/press-releases/red-hat-brings-distributed-ai-inference-production-ai-workloads-red-hat-ai-3
IBM	October 2, 2025	IBM Granite 4.0: hyper-efficient, high performance hybrid models for enterprise	https://www.ibm.com/new/announcements/ibm-granite-4-0-hyper-efficient-high-performance-hybrid-models
IBM internal material	March 2026	Fusion for AI Inference	Internal document used for this assessment