ANALYST CONNECTION
Sponsored by: Red Hat and Dynatrace

Enterprises that invest in centralized technologies, processes, and tools will be best positioned to speed the cycle of AI experimentation to production.

Achieving Transparency for Efficient, Safe AI at Scale

December 2025

Questions posed by: Red Hat and Dynatrace

Answers by: Nancy Gohring, Senior Research Director, AI

Q. What are the principal challenges enterprises encounter when scaling AI initiatives?

A. After spending the past few years experimenting with emerging AI technologies, many enterprises now face challenges with efficiently and safely deploying their AI applications for production use. Recent IDC research revealed that 45% of global respondents have only slightly improved their ability to scale new AI initiatives, with limited capabilities and no defined process yet. In fact, their challenges mean that 37% are now focusing on making existing AI projects work better, rather than starting new AI projects (29%), with 19% doing fewer AI projects and 6% stopping or reducing investment in AI (source: IDC's *Future Enterprise Resiliency and Spending Survey, Wave 7,* September 2025).

The key to scaling AI is building a foundation that eliminates the need to recreate all the requirements of an AI application. The foundation should remove barriers by allowing for rapid deployment to approved infrastructure within a governance framework that ensures compliance with security, privacy, and regulatory requirements. It should enable visibility into the performance of the AI application and access to tools that allow for continuous ongoing iteration.

Successful organizations are those that create this type of foundation to support rapid experimentation and learning, leading to the ability to quickly decide whether an experiment should be invested in or discontinued. The ability to adapt, iterate, and scale is what distinguishes AI leaders from those who struggle to move beyond isolated experiments.

Q. What should that foundation look like?

A. Think of an AI foundation as a launchpad. An effective AI foundation is dynamic and future-proof and encompasses technologies, processes, and tools. It should be designed to support a broad spectrum of workloads and use cases, from data-heavy tuning and low-latency inferencing to internal and external-facing applications. This means supporting hybrid and multicloud environments; leveraging accelerators, such as GPUs and TPUs; and enabling seamless integration with a variety of data sources and models.

Technology is only part of the equation. A robust governance framework is essential, encompassing ethical considerations, data stewardship, security, and accountability for outcomes. Governance should be embedded into every layer of the AI foundation, not as a barrier but as an enabler of responsible innovation. This includes establishing clear policies for data usage, model transparency, and risk management, as well as mechanisms for monitoring and auditing AI systems.

The foundation should also include cultural readiness. Organizations must feel empowered to experiment, fail, and learn from experience. That sense of empowerment most often comes when the C-suite is vocally supportive of AI adoption. An AI center of excellence can also be considered part of an AI foundation, bringing together expertise from across the organization to develop best practices that support experimentation and AI deployment. By building a foundation that is both flexible and governed, enterprises position themselves to scale AI effectively and sustainably.

Q. What should organizations consider when it comes to infrastructure that supports and scales enterprise AI aspirations?

A. AI is not a one-size-fits-all technology. Predictive models, generative applications, and agentic systems each place unique demands on infrastructure, as do different AI-related workloads, including training, tuning, inferencing, and RAG. The key is to align resources with the specific needs of each workload, balancing performance, scalability, and cost.

Forward-thinking enterprises are investing in modular, adaptable infrastructure that can evolve as AI use cases mature and diversify. This approach often involves leveraging cloud-native architectures, containerization, and orchestration platforms that enable rapid deployment and scaling of AI workloads. The ability to dynamically allocate resources based on workload requirements is critical for optimizing performance and controlling costs.

The goal is to create an infrastructure that is resilient, scalable, and future ready, capable of supporting the evolving needs of AI initiatives as they move from experimentation to production and beyond.

Q. Once AI applications hit production, how should enterprises think about balancing accuracy, latency, and cost?

A. Achieving controlled balance across accuracy, latency, and cost is key to sustainably scaling AI applications. The three metrics are intertwined. For instance, making changes to improve accuracy can at the same time increase cost and latency. Depending on the use case, some applications may be able to tolerate higher latency or even lower accuracy, while others will require the best possible accuracy, even at a high cost. Organizations with robust visibility are equipped to make informed trade-offs, ensuring that AI applications deliver the desired outcomes.

AI observability goes beyond traditional monitoring, providing granular insights into model performance, data quality, and system behavior. By collecting and analyzing telemetry data, engineers can identify sources of inaccuracy — such as biased training data or model drift — and take corrective action. Observability also helps pinpoint latency issues, allowing teams to optimize infrastructure or adjust deployment strategies to meet performance requirements. Cost management is another area where observability delivers value. Detailed telemetry enables organizations to track resource

consumption and experiment with different deployment options, such as shifting workloads or adjusting models to reduce expenses.

Q. How will the deployment of agentic AI in particular impact transparency?

A. The rise of agentic AI introduces new dimensions of opportunity and risk for enterprises. In multiagent environments, transparency becomes mission critical because without it, companies lack control. Organizations must be able to trace every step of an agent's actions, understanding how it made decisions, what information it accessed, what actions it took, and whether it achieved the desired outcomes.

This level of visibility is critical for compliance and auditing, particularly as regulatory scrutiny of AI systems intensifies. It also allows enterprises to manage cost and accuracy in agentic AI systems. Observability data can reveal inefficiencies, such as redundant agent interactions or excessive API calls, allowing organizations to streamline operations and control expenses. At the same time, visibility into agent performance helps ensure that the business is meeting its objectives and that AI systems are delivering tangible ROI.

As enterprises deploy more AI agents and scale their operations, the volume and complexity of telemetry data will increase. It becomes essential to collect and store the right data sets — those that provide actionable insights — while avoiding unnecessary overhead. Strategic observability empowers organizations to continuously refine their AI systems, maintaining the optimal balance between accuracy, speed, and cost.

Ultimately, the successful deployment of agentic AI hinges on the ability to maintain transparency, manage complexity, and align technology with strategic goals. Enterprises that invest in advanced observability and governance will be best positioned to harness the full potential of agentic AI while mitigating risks and maximizing value.

About the Analyst

Nancy Gohring, *Senior Research Director, AI*

Nancy Gohring is a senior research director, leading IDC's GenAI and Agentic AI Strategies program. Gohring covers big picture trends related to enterprise adoption of AI, including GenAI and agentic AI. Key research themes include business, organizational, and technology architecture transformation. As part of the Worldwide AI, Automation, Data, and Analytics Research practice, Gohring supports a range of clients across the technology stack, including hyperscalers, developer tool providers, enterprise application vendors, professional services organizations, automation frameworks providers, and infrastructure suppliers.

MESSAGE FROM THE SPONSOR

Red Hat and Dynatrace offer solutions that help organizations accelerate AI adoption, shorten time-to-value, and simplify operations. Together, we help you build trusted, scalable, and cost-optimized AI systems. Red Hat provides a comprehensive platform for managing the AI/ML lifecycle at scale across on-premise, cloud, hybrid, and edge environments. Dynatrace offers unparalleled observability across those environments, applications, and user sessions for AI workloads. Develop, train, and fine-tune predictive and generative AI models faster, bringing models from experiments to production more quickly—with end-to-end visibility, automated insights, and proactive issue detection, identification, and resolution. Learn more.

IDC Custom Solutions

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494
T 508.872.8200
F 508.935.4015
blogs.idc.com
www.idc.com

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. This IDC material is licensed for external use, and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives.