HyperFRAME RESEARCH

Sponsored by: dynatrace

RESEARCH BRIEF

Observability in the Age of Generative and Agentic AI

Why AI complexity requires a different observability approach.



Authors:

Stephanie Walter
Practice Leader AI Stack

Steven Dickens
CEO & Principal Analyst

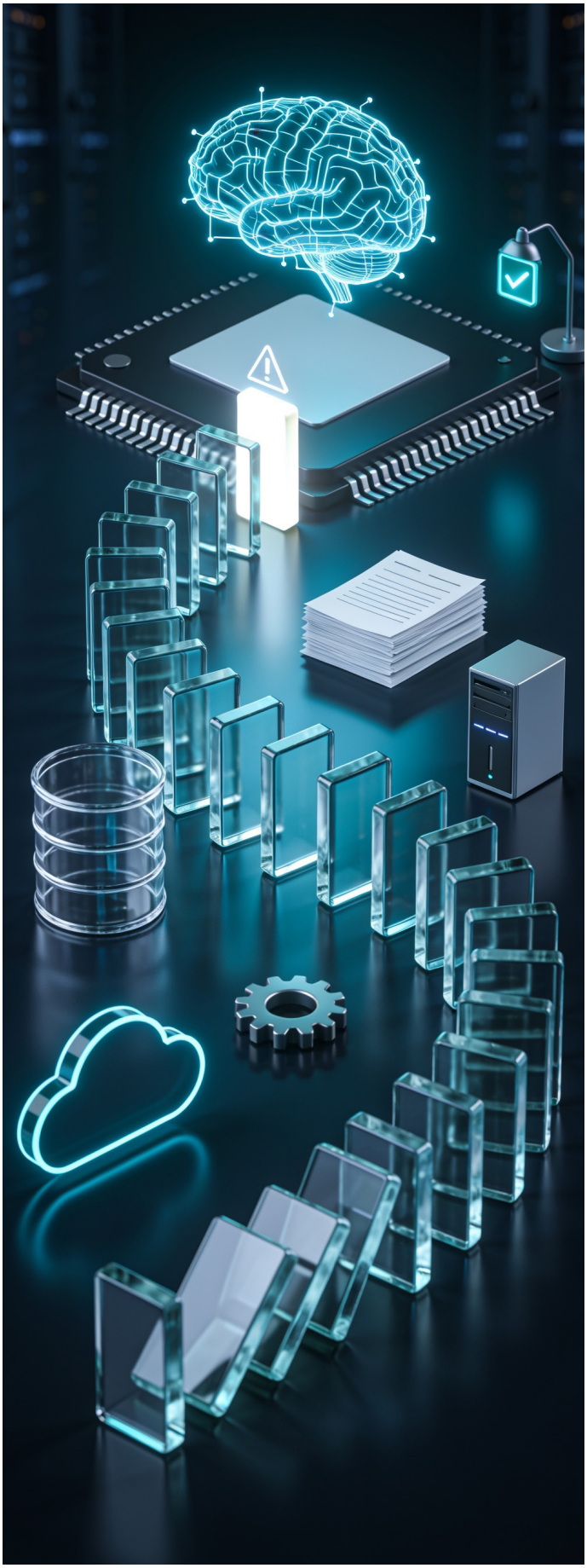
OCTOBER 2026

Executive Summary

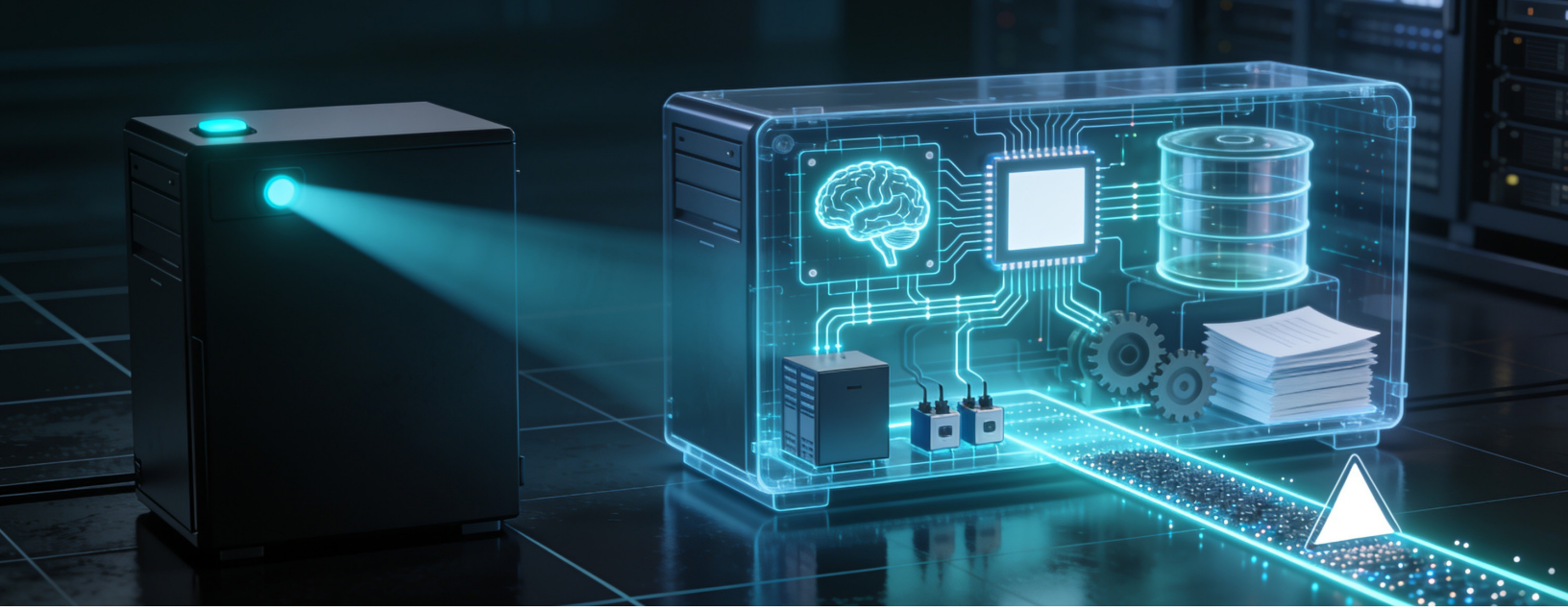
Generative AI (GenAI) has changed what enterprises can build. Agentic AI is changing how those systems behave once deployed. Unlike conventional applications, AI systems don't always follow a fixed execution path or fail in a predictable way. A response can arrive on time and return a technically successful status code while still being inaccurate, unsafe, irrelevant, or unexpectedly expensive. As agents gain access to tools, data, and downstream systems, one poor decision can propagate across an entire workflow.

These risks make observability a production requirement, not an operational enhancement. Practitioners need to see more than whether a model endpoint is available. They need to understand which prompts, models, agents, tools, and services participated in an outcome; why the system selected a particular path; what that path cost; and where quality or policy controls broke down. This requires early instrumentation, correlated telemetry, continuous evaluation, and root cause analysis that spans the AI and traditional application stack.

Enterprises should instrument AI workloads before production, establish baselines for quality, performance, cost, and risk, and select a modern observability platform that can preserve context across rapidly changing environments. Dynatrace addresses this challenge in two related ways. Its AI Observability capabilities are built for AI workloads, providing visibility into models, prompts, agents, tools, quality, performance, and cost. Dynatrace Intelligence is powered by AI, applying causal, predictive, generative, and agentic capabilities to help teams understand problems and support governed action. Both operate on the shared foundation of Grail and Smartscape, which connects AI activity to application, infrastructure, and dependency context. Dynatrace's completed acquisition of Arize adds AI-focused tracing, evaluation, and experimentation capabilities, creating an opportunity to connect AI engineering workflows more closely with production operations.



HyperFRAME Research. All rights reserved. 2026



Why AI Systems Require a Different Observability Approach

Enterprise AI already runs across customer-facing applications, employee workflows, software delivery, security operations, and core business processes. Many of these applications use multiple models, retrieve enterprise data, call tools and APIs, and interact with downstream systems. The challenge is no longer observing one user interacting with one model. It is understanding the complete AI workflow and how its components affect the outcome.

The scale of this transition is becoming measurable. According to the HyperFRAME Research Lens: State of the Enterprise AI Stack 1H 2026, a survey of 544 enterprise technology leaders, only 30% of organizations operate AI at scale, but 66% expect mass AI deployments within the next 12 to 24 months. This gap suggests that observability demands will rise as more interconnected AI workflows reach production. Each additional model, tool, data source, and handoff creates more telemetry and more possible points of failure.

Practitioners now need to observe the complete workflow. A model, tool, or service may appear healthy on its own while the combined system still produces a poor result. One flawed response can also become context for another agent, allowing an error to spread through the workflow.

Cost and performance problems can be equally difficult to see. A looping planner, inefficient retrieval step, or unnecessary model handoff may consume additional tokens and compute while the application continues to function. Conventional monitoring can show that the system is available. It may not reveal that answer quality has declined or that the cost of completing a task has doubled.

Instrumentation is the foundation of operational control. If teams wait until an AI application is in production to decide what they need to measure, they will discover that many of the signals required to reconstruct an outcome were never captured. Prompts, model versions, retrieval context, agent handoffs, tool calls, token consumption, evaluation scores, and downstream service dependencies should be designed into the telemetry model from the beginning.

This is where the distinction between monitoring and observability matters. Monitoring asks whether a known component is available or whether a predefined threshold has been crossed. Observability asks why the system is behaving as it is and whether practitioners have enough context to act quickly. Monitoring can tell a team that latency increased. Observability should help reveal that a prompt revision caused an agent to make three additional tool calls, one of which invoked a slower model and created a downstream database bottleneck.

AI-specific failure modes add another layer. Teams must watch for model and data drift, hallucinations, biased or toxic outputs, prompt injection, PII exposure, and deteriorating relevance or faithfulness. They must connect these quality signals to conventional telemetry such as traces, logs, infrastructure utilization, deployment changes, and service dependencies.

HyperFRAME Research. All rights reserved. 2026

Why Agentic Systems Are Harder to Observe

Variable execution is part of the operating model for agentic AI. Traditional software generally executes logic that engineers can trace through a known code path. Agents instead interpret a goal, assemble context, choose among tools, and revise their approach as conditions change. Two successful runs may follow different paths, while an unsuccessful run may be difficult to reproduce. Practitioners therefore need traces that capture not only what was called, but also the sequence of decisions, handoffs, prompts, retrieved context, evaluations, and downstream effects surrounding the outcome.

Model choice adds another variable. Enterprises increasingly use multiple commercial, open, specialized, and fine-tuned models. Behavior can shift when a provider updates a model, a team changes a system prompt, retrieval returns different context, or routing logic selects a new endpoint. Accuracy emerges from the combination of the model, prompt, context, application logic, tools, and data involved in each run. Tracking drift and reliability therefore requires continuous evaluation tied to production telemetry.

The workflow itself creates another source of complexity. Agentic applications chain together orchestration frameworks, vector or search systems, databases, APIs, cloud infrastructure, identity controls, and business applications. Every handoff introduces latency, cost, permissions, and another possible failure point. A minor anomaly can ripple through the chain: slow retrieval can change an agent's plan, a malformed tool response can contaminate context, or a retry policy can multiply token usage. Looking only at the LLM layer misses the infrastructure and application dependencies that shape the final result.

Enterprises must also connect technical behavior to business requirements. They need evidence that AI behavior complies with policy, protects sensitive data, stays within economic thresholds, and meets service-level expectations. Yet costs and outcomes become harder to predict as agents branch, loop, or escalate to more capable models. Teams need to connect each technical trace to a business transaction and determine whether the expense produced an acceptable outcome.

Point tools and team-built dashboards may be sufficient for monitoring a limited set of model metrics. The requirement changes when AI workflows cross applications, infrastructure, data services, and business systems. Teams then need to correlate model behavior with the rest of the production environment without manually building and maintaining those relationships.

The challenge spans four dimensions: agent behavior, model diversity, system interdependence, and business accountability. Practitioners need automated discovery and correlation across all four so they can investigate problems before autonomous systems continue acting on them.



HyperFRAME Research. All rights reserved. 2026

Recommendations for Practitioners

Practitioners should treat observability as part of AI system design, not a tool added after deployment. Five priorities stand out:

1. **Instrument early**. Capture standardized traces and attributes for model requests, prompts, agent and tool interactions, retrieval steps, token use, cost, errors, and evaluations. OpenTelemetry and emerging GenAI semantic conventions can help reduce dependence on a single framework while creating a consistent foundation across teams. Instrumentation should also connect AI activity to the application, infrastructure, user experience, and business process it affects.

2. **Establish multidimensional service levels**. Availability and latency remain important, but they are insufficient. Teams should define acceptable ranges for relevance, faithfulness, hallucination, toxicity, PII leakage, cost per successful outcome, and other measures appropriate to the use case. A fast answer that is wrong is not a healthy transaction.

3. **Use a modern observability platform**. Rather than stretching legacy monitoring across a new operating model, practitioners need real-time metrics, topology and dependency context, distributed traces, prompt-level visibility, continuous evaluation, proactive anomaly detection, and workflows that can initiate governed remediation. The goal is correlation across the complete production environment.

4. **Prioritize root cause over anomaly volume**. AI systems will generate enormous numbers of signals. Teams need causal explanations that identify the likely source and downstream impact of an issue. Every automated action should carry enough supporting context for humans to verify why it was recommended or executed.

5. **Evaluate agentic workflows continuously in production**. Availability and latency will not reveal declining answer quality. Use evaluations and LLM-as-judge scoring to measure accuracy, relevance, and grounding across retrieval, model routing, tool use, retries, and handoffs. Surface hallucinations and ungrounded responses, review guardrails and permissions, and fix problems at their source.



HyperFRAME Research. All rights reserved. 2026



Dynatrace AI Observability and Dynatrace Intelligence

Dynatrace approaches this problem from two directions. AI Observability is built to monitor customers' AI applications and agents, while Dynatrace Intelligence uses AI within the Dynatrace platform to support analysis and action. Grail brings telemetry and other operational data together, while Smartscape maps relationships across applications, services, infrastructure, and cloud environments. Both capabilities use this shared context.

That context matters because AI issues rarely stay in one layer. A poor answer may stem from a prompt change, retrieval issue, model update, tool failure, or downstream service problem. Dynatrace helps teams connect those signals so they can move faster from symptom to cause.

Teams can instrument AI workloads through OneAgent or standards including OpenTelemetry and OpenInference, bringing traces, metrics, and logs into Grail. Dynatrace provides different ways to examine that data through DQL and dashboards, from individual prompts and traces to service health, quality, and cost.

This supports several practical tasks: monitoring service health and cost, tracing a request from its initial prompt to the final response, identifying where latency or quality changed, and preserving the evidence behind an AI outcome. Topology data helps teams see how agents, models, providers, tools, and services interacted during the workflow.

Arize provides evaluations that help teams assess AI quality during development and in production. These evaluations can surface changes in relevance, faithfulness, safety, and other measures. As integration progresses, the opportunity is to connect these findings with Dynatrace telemetry to investigate issues across the model, prompt, retrieval layer, orchestration logic, and underlying infrastructure.

Arize extends this lifecycle perspective. Its tracing, evaluation, dataset curation, and experimentation workflows help AI engineers investigate agent behavior and test improvements. The acquisition creates an opportunity to connect that work with the application and infrastructure context used by SRE and platform engineering teams. The goal is a loop where a production failure informs the next evaluation or regression test, while a development change can be assessed against its operational effects.

Observability also plays an important governance role. Dynatrace can capture, visualize, and alert on guardrail outcomes, while Arize provides AI evaluations. Enforcement generally remains with the model, provider, or application layer. When the appropriate telemetry and retention policies are configured, prompt, response, trace, and lineage data can also provide an audit record for legal, compliance, and risk teams. Organizations still need to decide which data they are permitted to capture and retain.

Dynatrace Intelligence builds on the context provided by Grail and Smartscape. It combines causal, predictive, generative, and agentic AI capabilities to help teams understand what changed, why it matters, and where to investigate next. This grounds its recommendations in telemetry and dependency data rather than relying on an AI-generated answer alone.

Practitioners can interact with these capabilities through Dynatrace Assist, including natural-language exploration and context-aware guidance within the platform. Dynatrace Intelligence can also use agentic workflows and specialized tools to combine their outputs as part of an agentic investigation. This reduces the manual work required to connect signals across applications, services, infrastructure, logs, and traces.

HyperFRAME Research. All rights reserved. 2026



Looking Ahead

The next phase of AI observability will be defined by a shift from watching individual models to governing complete systems of action. As enterprises move toward mass deployment, practitioners will need to reconstruct how agents form plans, select models, access context, invoke tools, cross application boundaries, and change business state. Telemetry standards will have to evolve beyond technical performance signals to capture intent, decision paths, permissions, policy checks, evaluation results, and the business outcome of each workflow.

Observability is also likely to become more closely tied to runtime governance. Detecting a harmful or excessively costly action after it occurs provides limited protection once agents can change business systems. Over time, observability platforms may help identify behavioral drift and supply the context needed for policy-aware responses. Identity systems, security controls, and model guardrails will still enforce many of those policies.

Economics will emerge as another core observability discipline. Enterprises will move beyond tracking tokens and infrastructure spend toward measuring cost per trusted outcome. That will require connecting model selection, retrieval quality, agent path length, latency, evaluation scores, and remediation costs to the value produced by the workflow. As autonomy grows, buyers are likely to evaluate observability platforms by how well they connect quality, risk, performance, cost, and business outcomes.

Dynatrace provides a credible foundation for this evolution through its full-stack observability platform, unified data and topology context, and Dynatrace Intelligence, complemented by Arize's evaluation capabilities. Its opportunity is to extend that deterministic context into an evidence layer for increasingly autonomous operations. Supporting increasingly autonomous operations requires visibility across heterogeneous agent frameworks and model providers, context for policy-aware action, and human accountability as remediation becomes more automated.

With Arize joining Dynatrace, the opportunity is to connect AI engineering workflows with a broader observability strategy. Connecting evaluation and experimentation with production telemetry can help teams use operational findings to refine test datasets, compare potential fixes, and confirm in production whether those changes improve AI quality and reliability. As autonomous systems take on more work, enterprises will need evidence that they are reliable, appropriately governed, and producing enough value to justify their risk and cost. Observability will be central to that assessment.

HyperFRAME Research. All rights reserved. 2026

ABOUT HYPERFRAME RESEARCH:

HyperFRAME Research delivers in-depth research and insights across the global technology landscape, spanning everything from hyperscale public cloud to the mainframe and everything in between. We offer strategic advisory services, custom research reports, tailored consulting engagements, digital events, go to market planning, message testing, and lead generation programs.

Our industry analysts specialize in rigorous qualitative and quantitative assessments of technology solutions, business challenges, market forces, and end user demands across industry sectors. HyperFRAME Research collaborates closely with your Analyst Relations, Product, and Marketing teams to build and amplify your thought leadership, positioning your expertise to enhance brand and product recognition. Through content that engages readers, viewers, and listeners alike, we ensure your voice resonates across channels.

CONTACT HYPERFRAME RESEARCH:

Steven Dickens
CEO & Principal Analyst | HyperFRAME Research

Email Address:

steven.dickens@hyperframeresearch.com

Telephone Number:

+1 845 505 1678

X: @StevenDickens3

LinkedIn: Steven Dickens

BlueSky: Steven Dickens

CONTRIBUTORS

Stephanie Walter
Practice Leader, AI Stack

Steven Dickens
CEO & Principal Analyst

INQUIRIES

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

LICENSING

This document, including any supporting materials, is owned by HyperFRAME Research. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of HyperFRAME Research.

DISCLOSURES

HyperFRAME Research 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.


HyperFRAME
RESEARCH