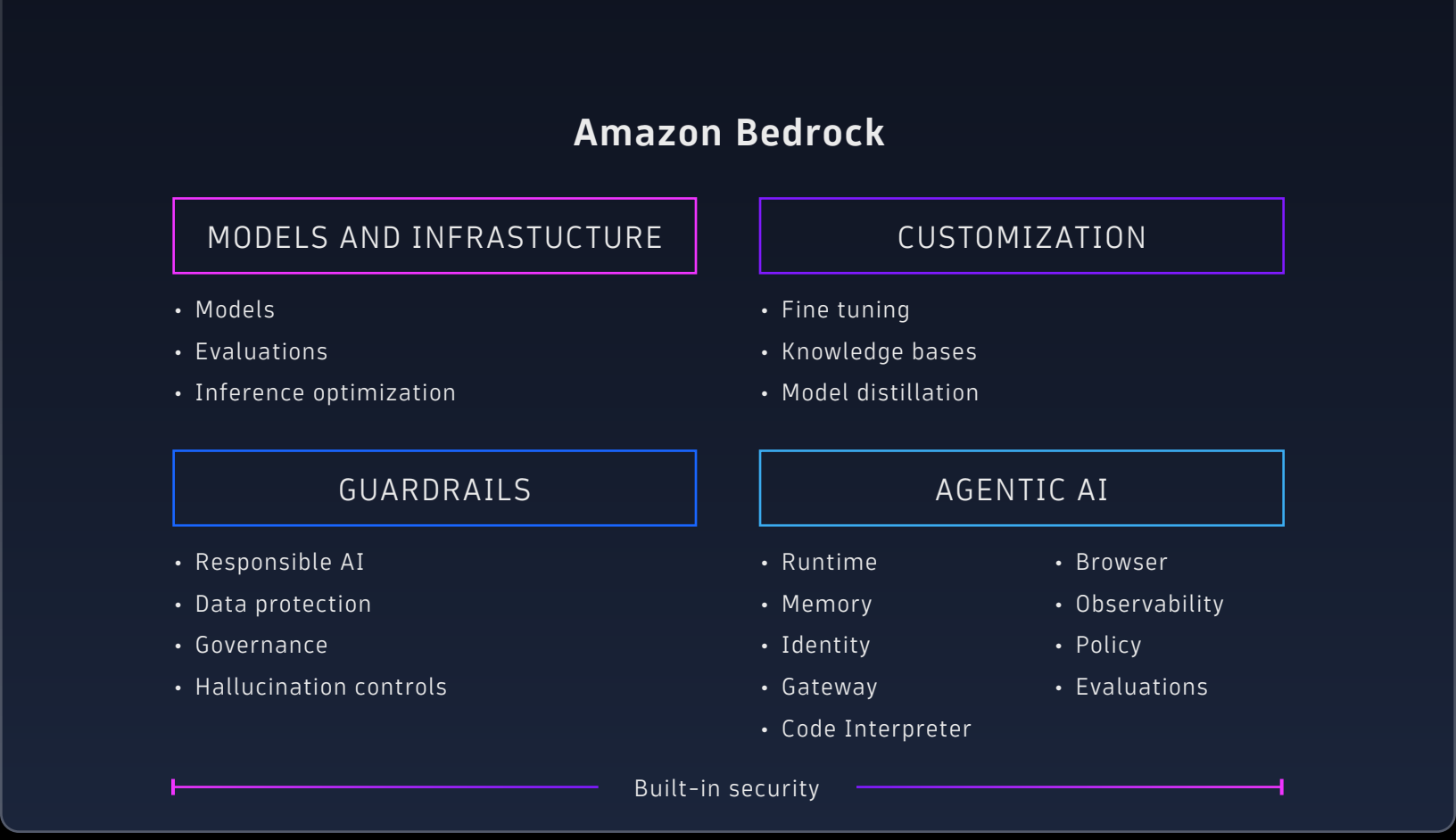
dynatrace

Maximize the power of Amazon Bedrock deployments

with Dynatrace AI-powered observability

©2026 Dynatrace



INTRODUCTION

What is Amazon Bedrock?

Amazon Bedrock is a fully managed cloud service from Amazon Web Services designed to help engineering teams easily build and scale generative AI applications.

It provides access to a range of foundational large language models (LLMs) and other generative AI models from leading providers, including Amazon, Anthropic, Cohere, Stability AI, and AI21 Labs.

With Amazon Bedrock, developers can experiment with, customize, and deploy generative AI models for tasks such as text generation, image creation, code synthesis, and more — all without managing underlying infrastructure.

SECTION ONE

Why engineering teams value Amazon Bedrock and its key features

Key features of Amazon Bedrock include the following:

- **Model variety.** Access multiple high-quality, pre-trained models from top AI companies, allowing users to choose the best fit for their use case.
- **Customization.** Amazon Bedrock enables organizations to fine-tune models with their own data, ensuring outputs are tailored to specific business needs.
- **Serverless deployment.** As a managed service, Amazon Bedrock abstracts infrastructure management, allowing developers to focus on building applications rather than provisioning hardware.
- **Secure and scalable.** Built on Amazon's secure cloud platform, Amazon Bedrock offers enterprise-grade security and the ability to scale workloads as needed.
- **Amazon Ecosystem integration**. Amazon Bedrock integrates seamlessly with other AWS services, such as Amazon S3, AWS Lambda, and Amazon SageMaker, enabling robust, end-to-end AI solutions.

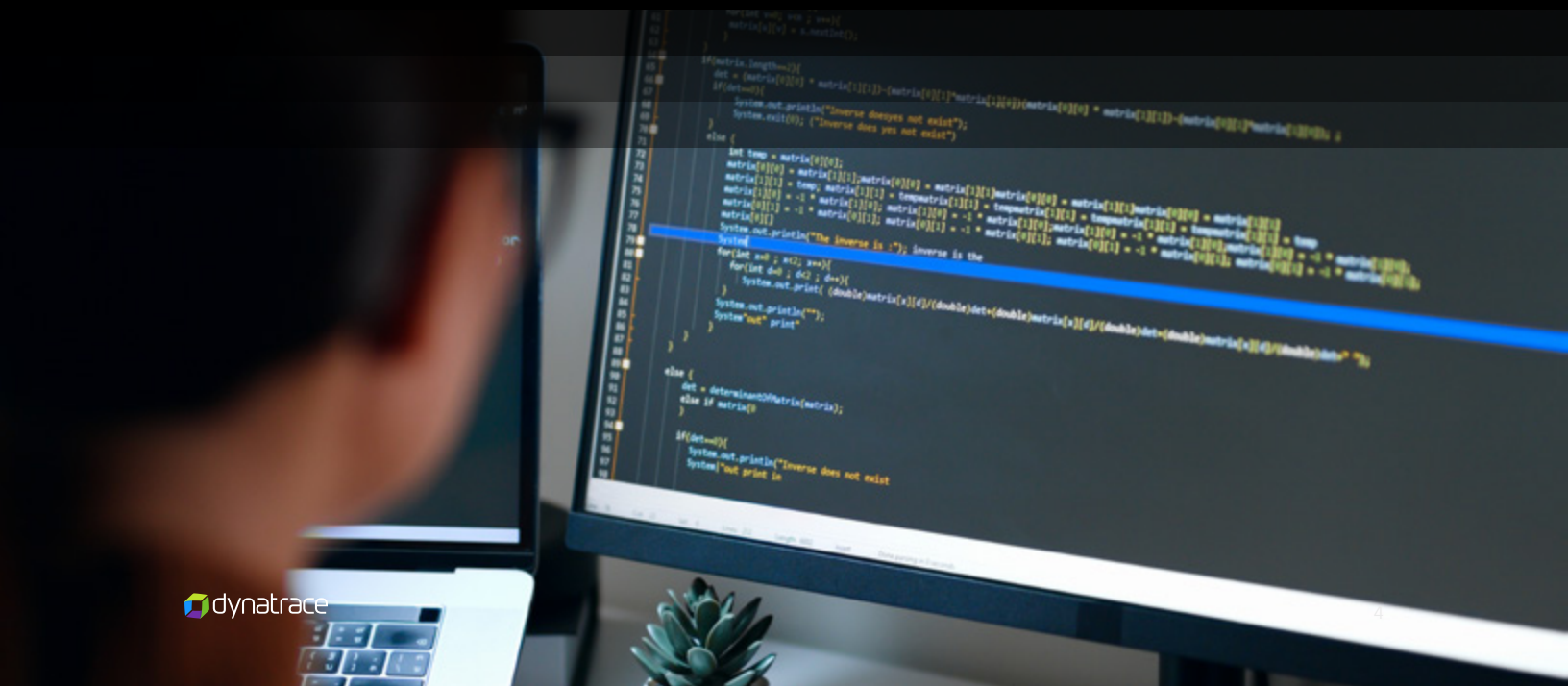
Many engineering teams turn to Amazon Bedrock as they look to integrate generative AI into their workflows and applications. By providing seamless access to a diverse set of advanced AI models, Amazon Bedrock accelerates innovation and shortens development cycles, making it easier to launch new features and products quickly.

Amazon Bedrock also simplifies the deployment of AI technologies. By leveraging a serverless architecture, the technology lowers barriers to entry, enabling companies without deep artificial intelligence expertise to build and deploy robust AI solutions while only paying for what they use.

Additionally, its integration with Amazon's secure cloud infrastructure ensures enterprise-grade security, compliance, and scalable customization, giving organizations confidence in both operational efficiency and data protection.

This flexible architecture enables enterprises to develop AI-powered applications at a significantly faster pace. However, the proper guardrails need to be put in place to ensure AI-enabled applications operate securely and effectively.

When deploying an application or agent to Amazon Bedrock, monitoring key performance indicators (KPIs) is essential to ensure optimal operation, efficiency, and security.



Critical key performance indicators

INFRASTRUCTURE AND APM

- Model latency
- Throughput
- Resource utilization
- Technical error rate

MODEL AND LLM OBSERVABILITY

- Token consumption and cost
- Security and compliance guardrails
- Model accuracy and quality
- Model version comparison

SECTION TWO

Critical KPIs to maximize the value of Amazon Bedrock deployments

To maximize the value delivered by Amazon Bedrock deployments, organizations must move beyond a one-size-fits-all monitoring approach. True visibility requires distinguishing between the foundational performance of your application infrastructure and the unique, dynamic behaviors of generative AI models.

By tracking these distinct categories of KPIs, you can maintain high standards of service, optimize resource allocation, and innovate with confidence.

Infrastructure and application performance (APM)

Ensure your underlying systems are resilient and scalable by monitoring these core operational metrics:

1. MODEL LATENCY

Track the response time of generative AI models to user queries. Low latency is crucial for delivering a smooth, responsive user experience, particularly in real-time or customer-facing applications like service agents.

2. THROUGHPUT

Measure the number of requests processed per second. High throughput indicates your system's ability to handle increased usage or traffic spikes without performance degradation, ensuring reliability at scale.

3. RESOURCE UTILIZATION

Monitor critical infrastructure metrics, including CPU, memory, and network usage. Visibility into these resources helps optimize costs, prevent bottlenecks, and ensure your applications scale efficiently across your cloud environment.

4. TECHNICAL ERROR RATE

Observe the percentage of failed requests or system operations. A rising error rate often signals integration issues or system instability that requires immediate attention to prevent downtime.

Model and LLM observability

Gain insight into the cognitive behavior, cost, and safety of AI deployments with these specialized metrics:

1. TOKEN CONSUMPTION AND COST

Track spending on compute, storage, and API usage, specifically focusing on token consumption. Monitoring these cost-related KPIs allows for precise budget control and informs strategic decisions regarding scaling and optimization.

2. SECURITY AND COMPLIANCE GUARDRAILS

Monitor guardrails to ensure safety, privacy, and compliance. Acting as a critical layer of control, these metrics track how effectively a system intercepts and filters user inputs and model outputs to block sensitive or inappropriate content.

3. MODEL ACCURACY AND QUALITY

Assess the relevance of AI outputs against business objectives and trustworthy data sources. Measuring the "grounding" of a model ensures outputs are explainable and derived from corporate data sources rather than hallucinations, building trust in AI agents.

4. MODEL VERSION COMPARISON

With new models and versions releasing frequently, it is vital to ensure the application utilizes the most effective option. Leverage practices like A/B testing to compare models in real time, ensuring continuous improvement and adaptation.

SECTION THREE

How to operationalize AI observability

Unified observability

Many organizations rely on different solutions when developing AI applications or agents, rather than using the same observability solution that will be used in production. This approach can lead to inconsistent monitoring, gaps in visibility, and challenges in identifying and resolving issues that may only surface under real-world workloads.

Without unified observability from development through production, teams may miss critical metrics, encounter difficulty in correlating data across environments, and experience delays in troubleshooting or optimizing system performance.

Ultimately, this disconnect can undermine the reliability, security, and efficiency of AI deployments, making it harder to ensure the application meets business objectives and compliance standards.

Standardized frameworks

OpenTelemetry and OpenLLMetry are invaluable tools for instrumenting AI applications and agents, providing standardized frameworks for collecting, analyzing, and correlating critical observability data.

By leveraging these solutions, teams can ensure consistent monitoring from development through production, enabling comprehensive visibility into model performance, user interactions, and system health. This unified observability helps quickly identify issues, optimize resource usage, and maintain compliance, supporting reliable and effective AI deployments.

End-to-end traceability

End-to-end traceability is crucial for AI and agentic applications because it enables teams to track the entire lifecycle of data, decisions, and interactions within the API, frontend stack, and backend applications. This comprehensive visibility ensures that performance bottlenecks, errors, and anomalies can be quickly identified and addressed, minimizing downtime and maintaining high service quality.

By correlating **traces, logs, events, and metrics** across all components, organizations can ensure their AI applications function as intended, meet business objectives, and remain compliant with regulatory standards. Effective traceability also supports faster troubleshooting, enhances accountability, and provides the insights needed to continuously optimize both the model and underlying infrastructure.

Feedback loops and proactive alerting

Feedback loops, alerting, and human-in-the-loop automation are essential components when operating agentic-based applications. Robust feedback loops enable continuous learning and adaptation by providing actionable insights into model performance and user interactions.

Integrated alerting systems facilitate rapid detection of anomalies or degradations in service, allowing teams to proactively address issues before they affect end users. Incorporating human-in-the-loop automation further enhances reliability and accountability, as it ensures that critical decisions or edge cases receive expert review, mitigating risks and supporting trustworthy AI outcomes.

Unified view and visibility

Visualizing **model behavior, infrastructure health, user experience, performance, and cost trends** in a single unified view is critical when observing agentic applications. Centralized visualization streamlines correlation across signals, making it easier to detect anomalies, trace root causes, and understand the interplay between different system components.

When these key observability facets are distributed across disparate tools, teams face fragmented visibility, increased context-switching, and run the risk of missing important cross-domain insights. This disconnect can delay troubleshooting and lead to inconsistent data interpretation, jeopardizing reliability, cost-efficiency, and user satisfaction.

By consolidating all relevant observability data, organizations ensure a holistic, actionable view that supports faster decision-making and more resilient AI operations.

END-TO-END OBSERVABILITY
across the entire
AI infrastructure

REAL-TIME INSIGHT
into model behavior
and performance

CAUSAL AI
to surface root cause

UNIFIED DASHBOARDS
for observability
and accountability

SECTION FOUR

The power of AI observability with Dynatrace

Dynatrace provides robust observability for agentic-based applications by monitoring every layer of the technology stack — from infrastructure and cloud resources to AI models and user interactions. By correlating data from metrics, traces, logs, and events in real time, Dynatrace empowers teams to visualize system health, performance, and behavior in a unified dashboard.

This end-to-end visibility enables faster root-cause analysis, proactive anomaly detection, and seamless troubleshooting, while AI-powered analytics drive smarter automation and alerting. As a result, organizations can optimize resource utilization, enhance reliability, ensure compliance, and deliver consistent, high-quality user experiences across their entire AI ecosystem.

End-to-end observability of AI-powered stacks

Dynatrace captures telemetry across the entire AI infrastructure, extending from foundational cloud resources to granular, model-level signals such as latency, versioning, cost metrics, and detailed input/output traces.

By instrumenting every layer of the technology stack, Dynatrace seamlessly ingests and correlates data from infrastructure components, application services, and AI models, providing real-time insights into operational health and performance.

This unified approach enables teams to monitor individual model performance — including tracking version changes, latency spikes, resource consumption, and the flow of inputs and outputs — ensuring accurate root-cause analysis, cost management, and optimized user experiences throughout the AI lifecycle.

Dynatrace supports both self-hosted deployments and API-based services such as Amazon Bedrock. For self-hosted environments, Dynatrace can be instrumented directly within on-premises data centers or private clouds, capturing telemetry from custom AI models, application servers, and supporting infrastructure.

For API-based services, Dynatrace integrates seamlessly with cloud-native platforms, ingesting and correlating key performance indicators. This flexibility ensures that organizations benefit from unified observability regardless of deployment mode.

Real-time insight into model behavior and performance

Dynatrace excels at detecting anomalies in critical AI model metrics, such as token usage, input/output length, latency, and cost per request. By continuously ingesting telemetry from every layer of the technology stack, Dynatrace applies advanced AI-powered analytics to monitor and correlate these signals in real time.

This enables teams to quickly identify deviations from normal patterns, such as unexpected spikes in token consumption, unusually long or short input/output sequences, latency fluctuations, or abnormal cost per request. Automated alerts and visualizations empower operators to respond proactively, minimizing downtime and optimizing resource allocation for generative AI applications.

Additionally, Dynatrace provides robust tracking of model versions, allowing organizations to maintain visibility into version changes and their impact on application performance. Teams can monitor the behavior of individual model variants and easily compare key metrics across different versions.

This capability supports efficient model A/B testing, enabling teams to evaluate the effectiveness of new models side by side. As a result, organizations can make informed decisions about model deployment.

Casual AI to surface root cause

The Dynatrace Davis® AI engine automatically correlates performance and behavior anomalies with their underlying causes, enabling rapid and informed responses.

For instance, if there's a sudden spike in AI model hallucinations, Davis AI can trace the anomaly back to a recent version rollout or changes in input formation. This pinpointed context equips teams with actionable insights, allowing them to address issues quickly and confidently, reducing downtime and ensuring continued reliability of agentic applications.

Unified dashboards for observability and accountability

Dynatrace offers a variety of pre-built views tailored to meet the needs of different users, including AI native developers, platform teams or site reliability engineers, and executives. These dashboards seamlessly combine infrastructure health metrics with AI model outputs and business impact indicators in a single, unified interface.

Developers can drill down into granular telemetry and debug model behavior, while SREs monitor system reliability and resource utilization alongside operational anomalies. Executives, meanwhile, gain high-level visibility into business KPIs and the direct effects of AI performance on customer experience, enabling data-driven decision making across the organization.

Transform complexity into your greatest asset

The rise of generative AI marks a pivotal moment for every enterprise. Amazon Bedrock provides a powerful, flexible foundation for building and scaling AI-powered applications, but unlocking its full potential requires more than just adoption. It demands a new level of understanding — a unified view of your entire technology stack, from the underlying infrastructure to the cognitive behavior of AI models.

Without this clarity, you risk flying blind. Performance bottlenecks, security vulnerabilities, and rising costs can quickly undermine your AI initiatives, turning innovation into a source of frustration. But with the right observability strategy, you can turn this complexity into your greatest asset.

Dynatrace offers the end-to-end observability needed to master your AI-driven ecosystem. By providing unified insights into your application performance, model behavior, and business impact, Dynatrace empowers your teams to innovate with confidence. You can move from being reactive to proactive, automating intelligently and delivering digital experiences that are not only powerful but also reliable and secure.

Ready to harness the full power of your AI investments?

Contact us today for more information about how Dynatrace can help you maximize your Amazon Bedrock deployment.

ABOUT DYNATRACE

Dynatrace is advancing observability for today's digital businesses, helping to transform the complexity of modern digital ecosystems into powerful business assets. By leveraging AI-powered insights, Dynatrace enables organizations to analyze, automate, and innovate faster to drive their business forward. Learn more at www.dynatrace.com.

dynatrace

dynatrace.com/blog @dynatrace linkedin.com/company/dynatrace instagram.com/dynatrace

01.15.26 BAE22470_WP_USLet_cs