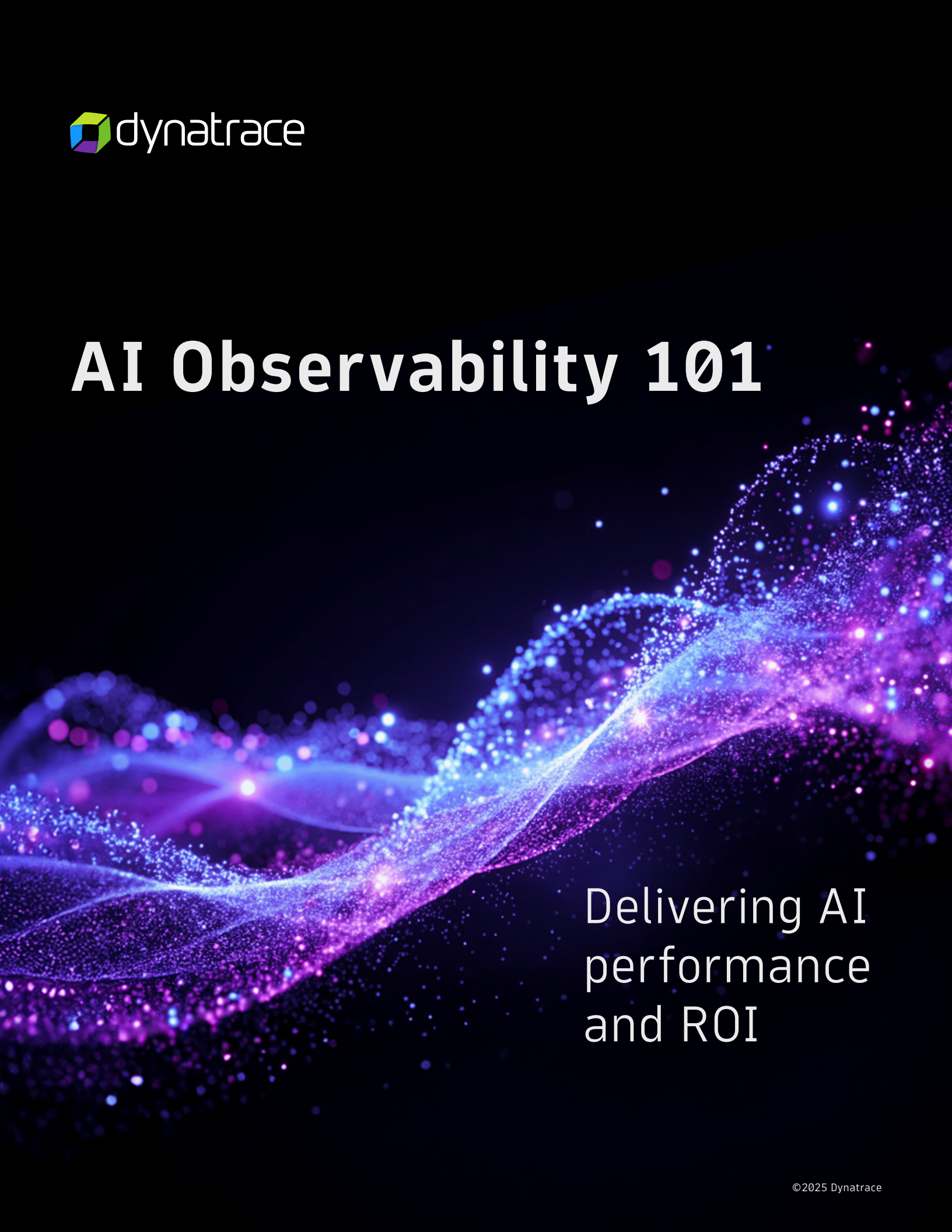
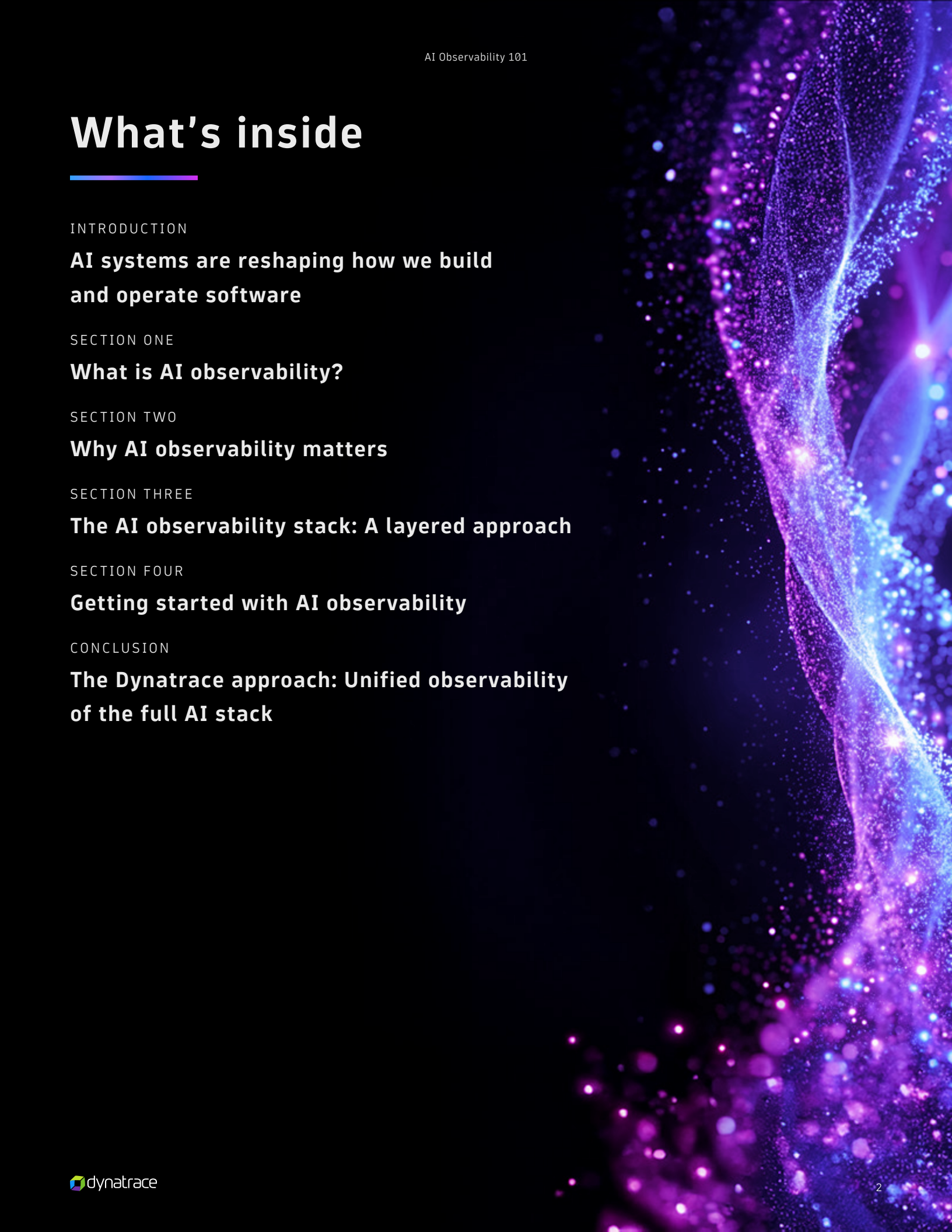
dynatrace

AI Observability 101



Delivering AI performance and ROI

What's inside



INTRODUCTION

AI systems are reshaping how we build and operate software

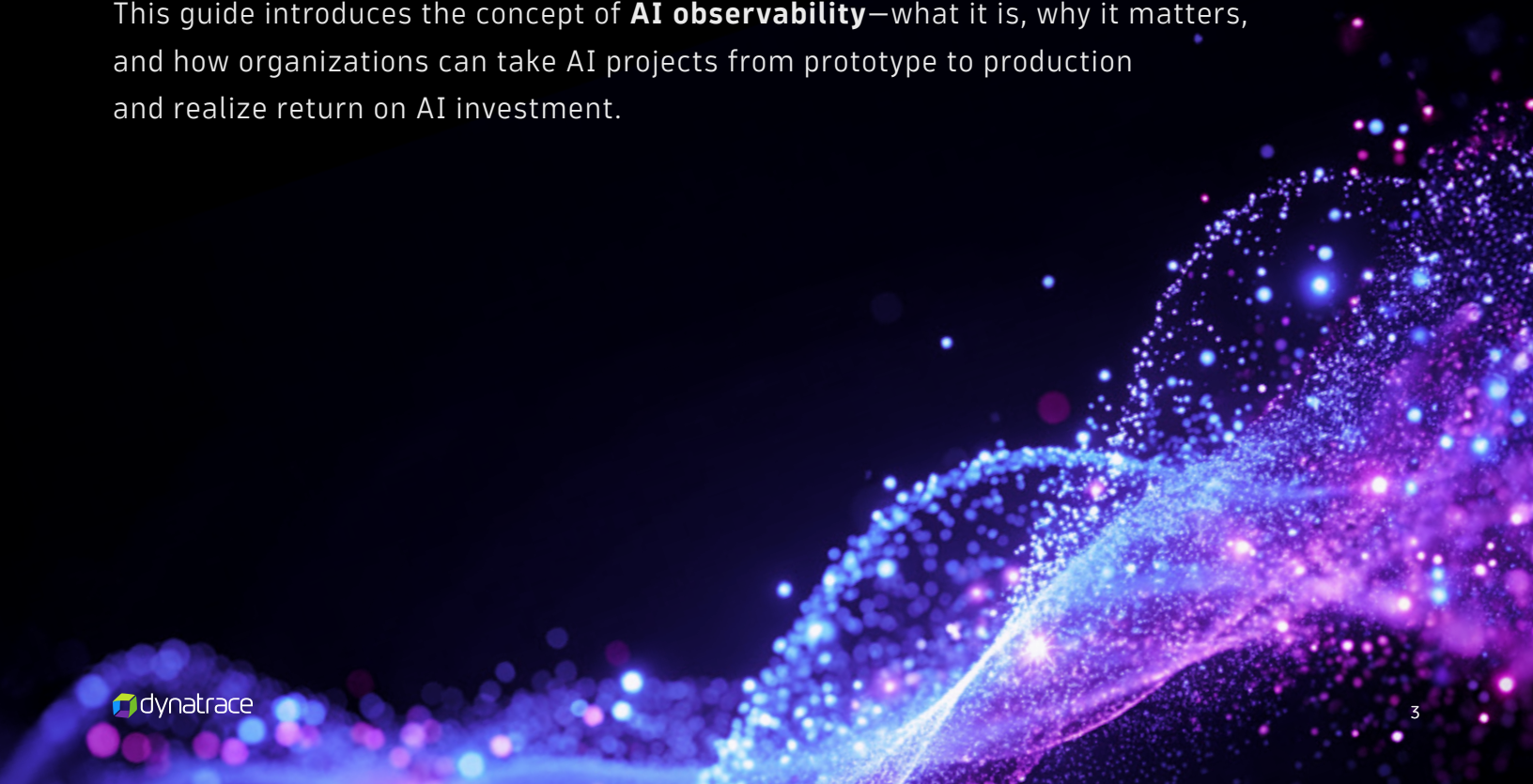
Organizations are turning to AI for greater efficiency, reduced operational toil, improved software performance, and better customer experiences.

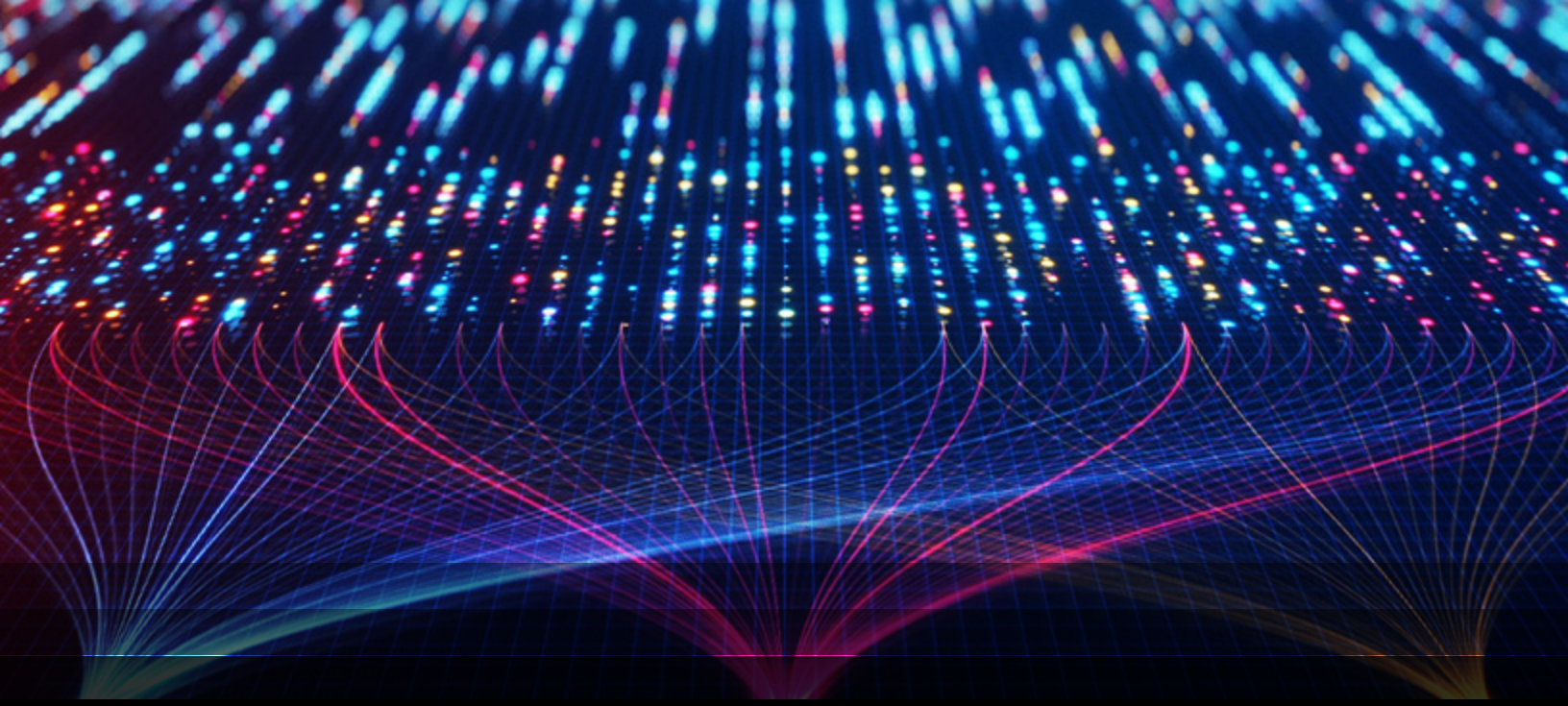
But as AI initiatives scale, engineering teams and leaders are discovering that **AI systems also introduce new complexity, risks, and blind spots.**

The result: **without visibility into models, pipelines, and infrastructure, teams are operating in the dark,** and AI projects often stall in the prototype phase instead of delivering on their promise.

AI observability changes that—giving teams the visibility and control they need to understand precisely what's happening inside these black-box systems, optimize for the outcomes that matter, and operate AI in production with confidence.

This guide introduces the concept of **AI observability**—what it is, why it matters, and how organizations can take AI projects from prototype to production and realize return on AI investment.





SECTION ONE

What is AI observability?

AI observability is the practice of monitoring, measuring, and understanding the behavior of AI systems, from individual model predictions to complex multi-agent workflows.

Observability customarily focuses on infrastructure metrics and application performance. Aided by open-source instrumentation standards such as OpenTelemetry, traditional observability tools excel at tracking telemetry signals, such as metrics, logs, traces, and events. But they can't answer whether an AI model's output is accurate, why a prompt generated an unexpected response, or whether reasoning chains are functioning as intended.

AI observability monitors the unique characteristics of AI systems based on telemetry from large language model (LLM) applications and their agents, infrastructure, and orchestration layers. This data includes model outputs, prompt interactions, token consumption, inference latency, and the reasoning chains that drive AI decision-making. Observability of these AI signals is likewise aided by open-source instrumentation standards, such as OpenLLMetry.

With AI observability, teams can answer critical questions traditional tools can't address, for example:

- Is the model hallucinating?
- Why did latency spike on this request?
- Which prompts are driving up costs?
- How is model performance changing over time?

AI observability provides the real-time insight to operate AI systems safely, efficiently, and at scale. This visibility transforms AI from an operational black box into a manageable, measurable component of your technology stack.

What is OpenLLMetry?

OpenLLMetry stands for "open large-language model telemetry." It's an open-source observability toolkit built on the OpenTelemetry standard that offers specialized instrumentation for LLM-based applications.

It extends OpenTelemetry's standard tracing, metrics, and logging capabilities by capturing LLM-specific data points—such as model details, prompt and completion tokens, latency, and errors. It enables developers to monitor, debug, and optimize LLM workflows across diverse observability backends.

SECTION TWO

Why AI observability matters

As AI systems evolve from basic assistants to autonomous agents executing complex tasks, the operational stakes intensify.

A single hallucinated response can erode trust. Silent model drift can materially reduce accuracy before anyone notices. Prompt-injection can exfiltrate sensitive data. And without visibility into token consumption and GPU utilization, costs can spiral out of control.

AI effectiveness demands model quality and operational reliability

Unlike standard web applications, AI systems generate probabilistic, non-deterministic responses that demand new approaches to quality assurance and performance monitoring. For instance, a customer service chatbot that begins hallucinating product features could damage customer trust and create support burdens before traditional monitoring detects any anomaly.

With AI observability, teams can detect hallucinations and prompt failures in real time, monitor accuracy and drift across model versions, and track latency and throughput across inference pipelines. This awareness enables organizations to enhance the reliability of AI-generated content and reduce the mean time to detect and resolve AI-related issues.

Accountability for security, compliance, and governance

AI systems also introduce novel security and compliance risks that traditional security tools may miss. AI observability helps teams to detect and mitigate PII leaks (where models inadvertently expose sensitive information), prompt injections (where malicious inputs manipulate model behavior), and policy violations while maintaining full audit trails across prompts, tools, and agents.

For regulated industries—such as healthcare, financial services, and government—this traceability ensures compliance with internal and regulatory requirements for responsible AI deployment.

Business value meets strategic impact

Beyond technical operations, AI observability drives measurable business outcomes. For example, reducing debugging cycles from days to hours can accelerate time-to-market for new AI features. Likewise, with visibility into cost-per-query, teams can optimize spend across thousands or millions of daily interactions.

AI observability enables organizations to transform AI from a risk source into a reliable business capability that drives innovation and sustainable growth.

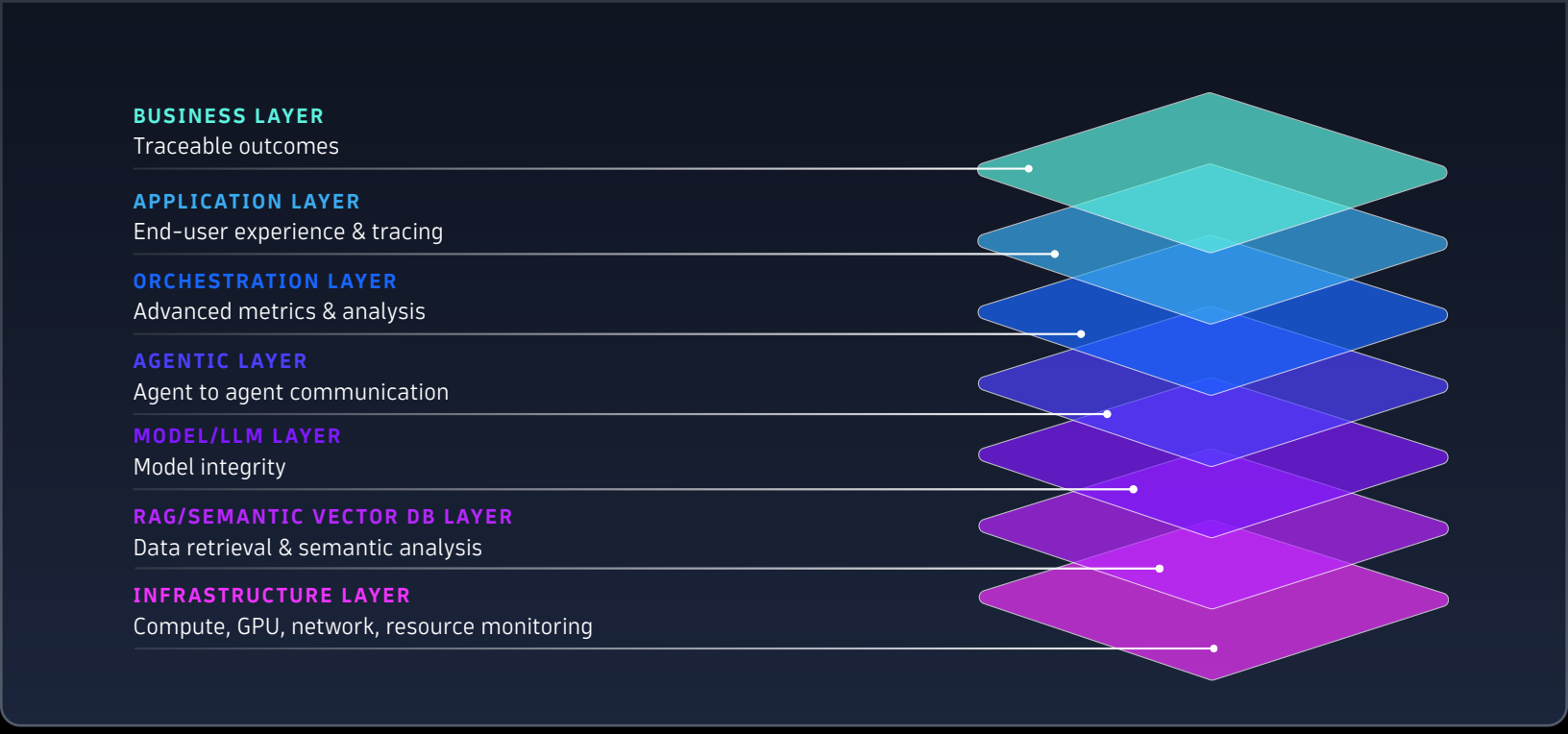
SECTION THREE

The AI observability stack: A layered approach

AI is not a single component—it's a complex system spanning multiple layers and technologies.

Effective AI observability requires visibility across the complete stack, from business outcomes to infrastructure, all within a unified context.

Taken in layers, teams can pinpoint the origin of issues—whether they stem from model quality, orchestration logic, infrastructure constraints, or business outcomes. Each layer generates unique telemetry that, taken together, reveal the complete picture of AI system health.



AI observability layers

- **Business impact.** Track how AI creates productivity gains, deflects support tickets, acts autonomously, and delivers return on investment.
- **Application performance.** Trace end-user experience, availability, and reliability of AI-powered applications.
- **Orchestration layer.** Track chain performance, guardrails, and prompt caching across orchestration frameworks.
- **Agent-to-agent communication.** Observe agent protocols, command execution, tool usage, and multi-agent communications.
- **Model integrity.** Assess token usage, cost, stability, latency, invocation errors, and resource utilization of model outputs.
- **Semantic caches and vector databases.** Monitor RAG pipelines, data volume, distribution, and retrieval patterns.
- **Infrastructure monitoring.** Track utilization, saturation, and errors across GPUs, TPUs, and compute resources.

Observability of these layers together helps teams understand what's happening with individual AI components and how they interact with traditional application and infrastructure layers to pinpoint issues and their downstream effects.

Five ways AI observability drives successful AI projects

When unified across all seven layers, AI observability transforms projects from prototype to delivering business value in production in several ways.

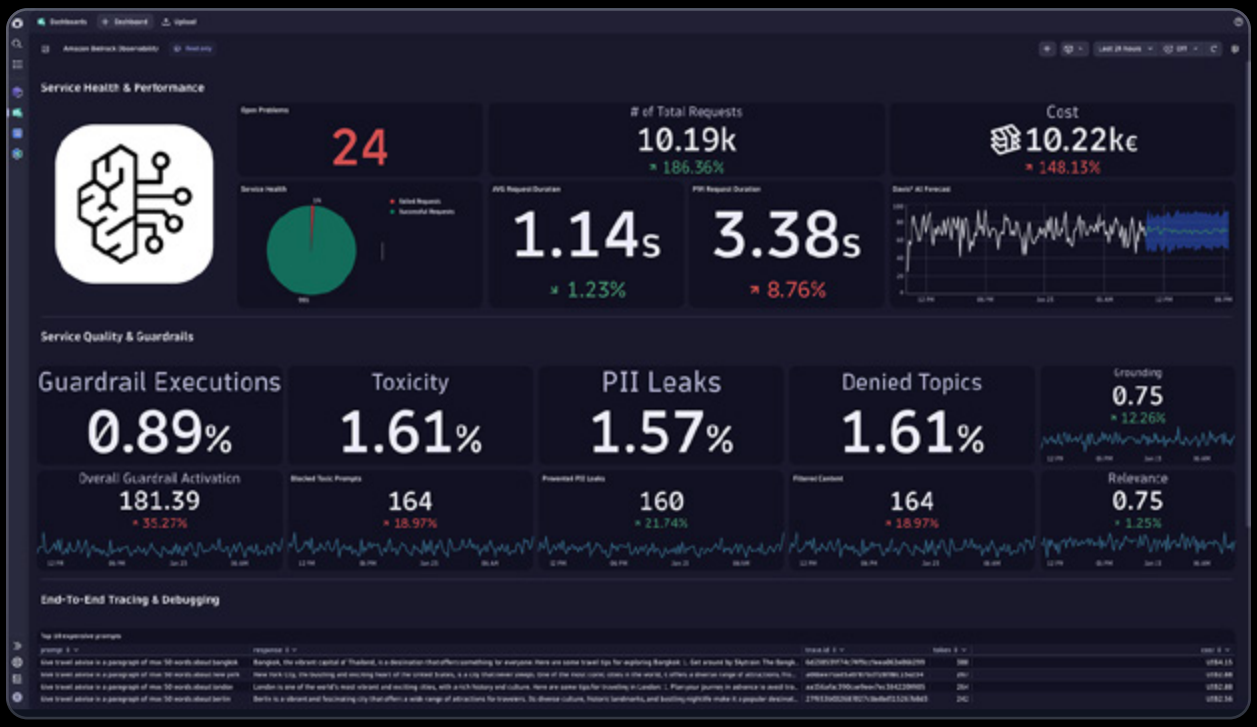


Figure 1: Amazon Bedrock observability dashboard

1. Building trust with intelligent guardrails

- Safeguard AI application quality by monitoring guardrail metrics—automated checks that detect hallucinations, malicious prompt injections, PII leakage, and toxic language.
- Analyze guardrail effectiveness and make adjustments to ensure optimal user experience and safety while mitigating potential biases and misuse.

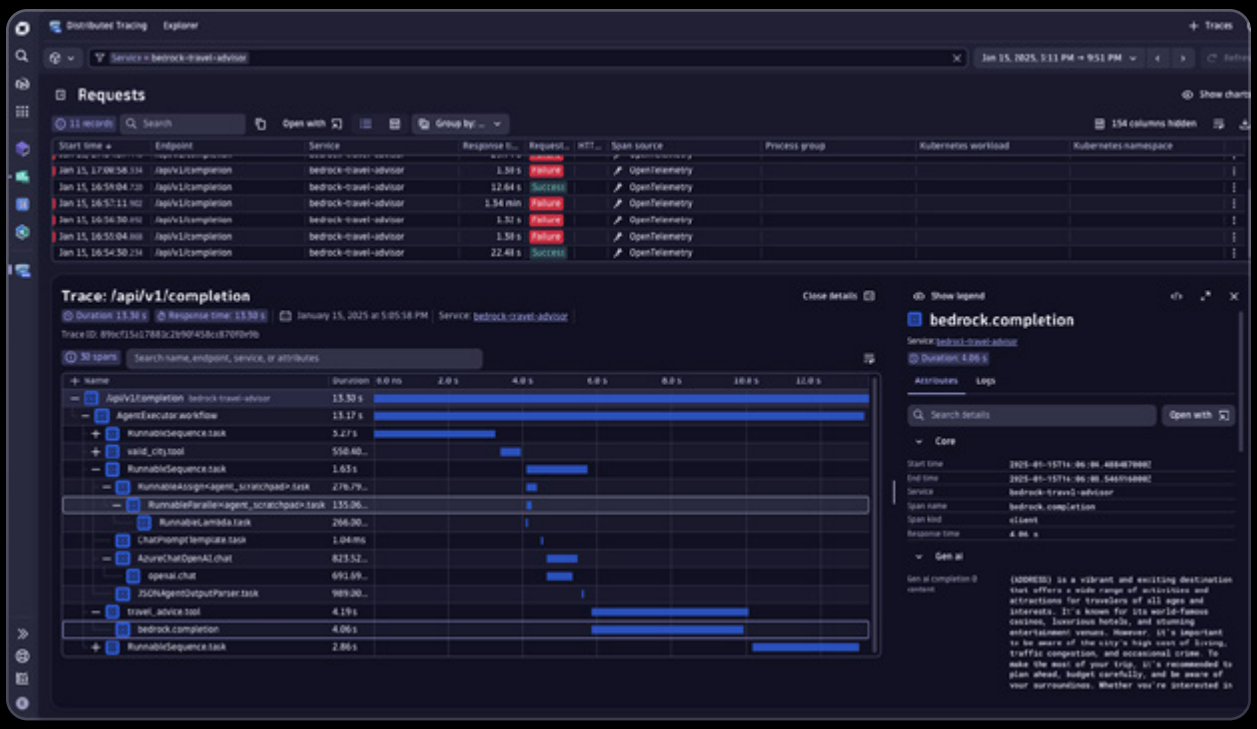


Figure 2. Distributed tracing explorer

2. Tracing and explaining AI outputs

- Gain end-to-end visibility into each user request with full traces covering the application stack, orchestration, semantic cache, and model layers.

  For example, when a customer receives an incorrect recommendation, teams can trace the request backward through the RAG pipeline, identify which document chunks were retrieved, and determine whether the issue originated from retrieval quality, prompt design, or model behavior.

- Map dependencies across multiple LLMs, RAG pipelines, and agentic frameworks.
- Automatically pinpoint the root cause of errors and failures to accelerate resolution before impacting customers.

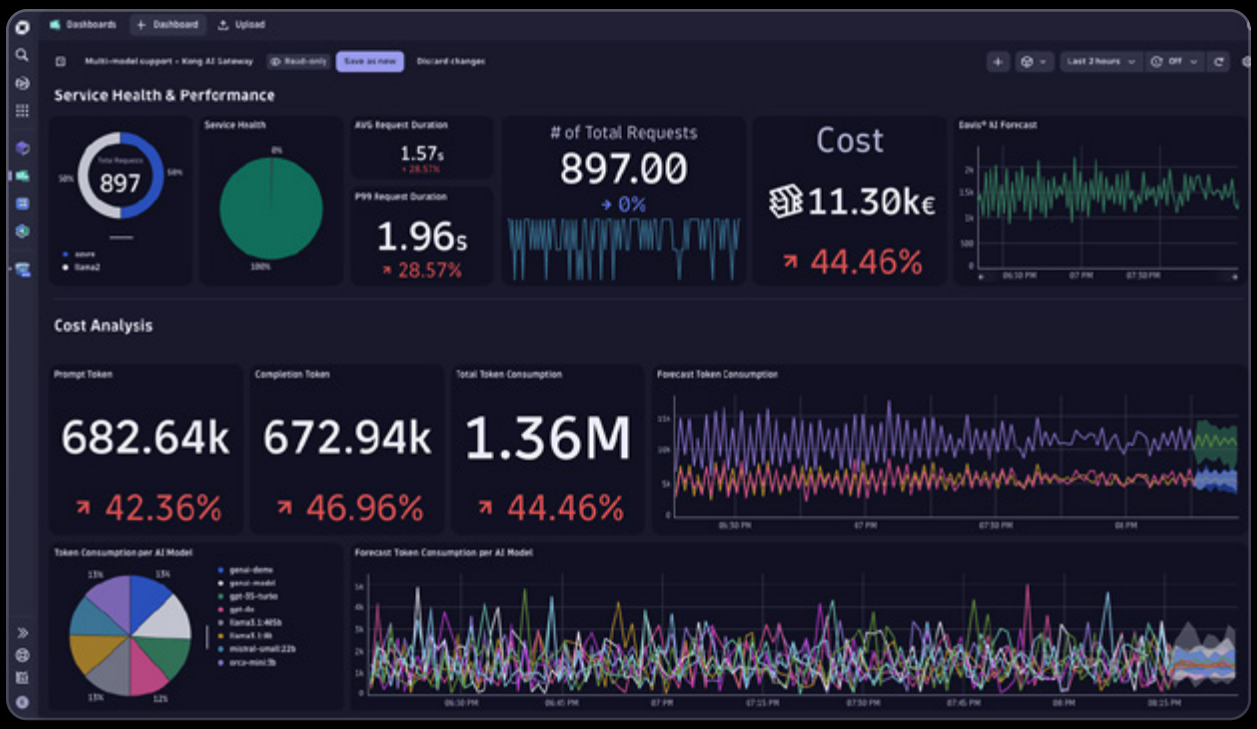


Figure 3. Multi-model support for Kong AI gateway

3. Reducing cost and improving performance

- Monitor token consumption, request duration, and GPU utilization to control expenses and optimize performance.
- Analyze traces for the slowest requests and errors to reduce model response times and improve reliability.
- Predict cost increases and trigger automated workflows based on predefined thresholds or unexpected behaviors.

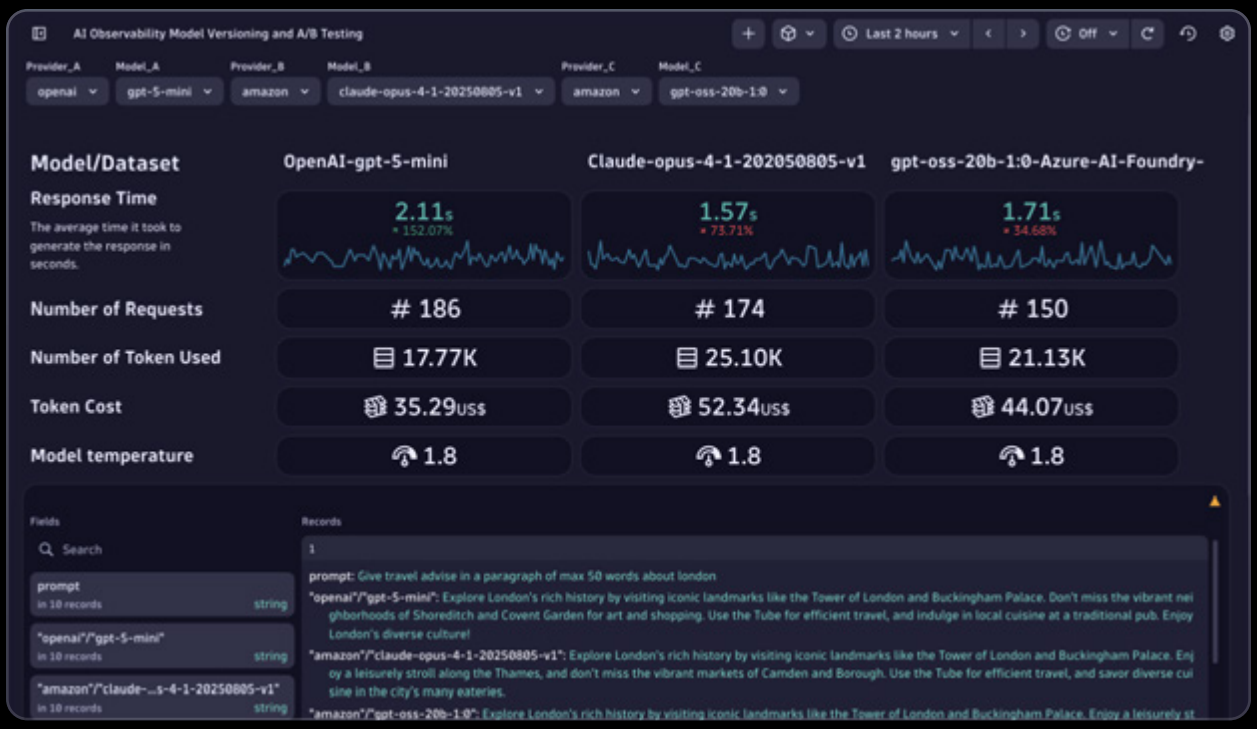


Figure 4. Model versioning and A/B testing

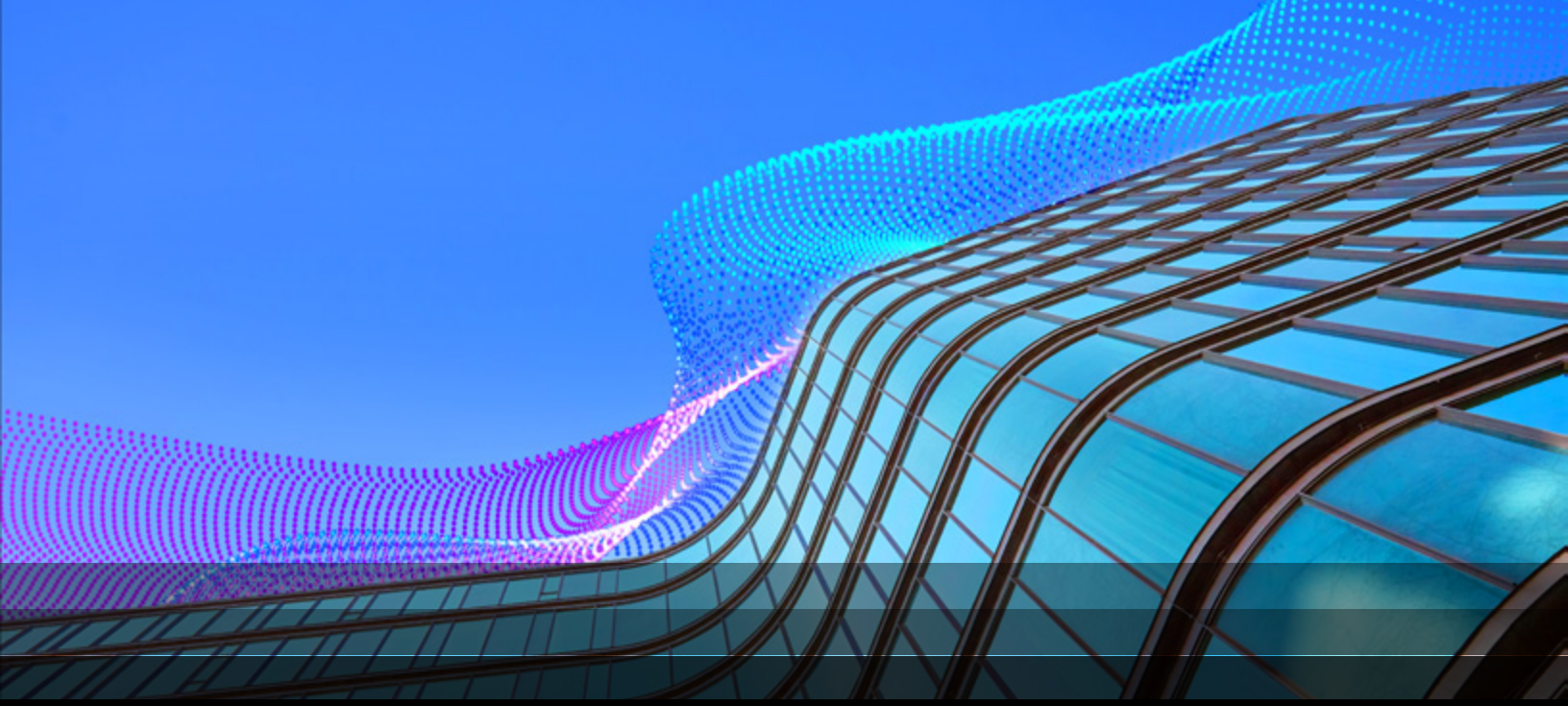
4. AI Model Versioning and A/B testing for more reliable models

- Compare models and versions to validate improvements and spot bottlenecks across latency, reliability, token usage, cost, and output quality signals.
- Drill into prompt-level differences to confirm why a variant wins.
- If something breaks, follow the request end-to-end with distributed tracing from input through orchestration steps and model calls to completion, to pinpoint exactly where an error or slowdown originated.

Figure 5. GenAI compliance audit report

5. Ensuring compliance and governance

- Comprehensively document AI inputs and outputs, maintaining full data lineage from prompt to response to build clear audit trails for regulatory compliance.
- Visualize AI system behavior and performance to increase transparency and demonstrate responsible AI practices.
- Track infrastructure metrics—including temperature, memory utilization, and process usage—to support both operational efficiency and sustainability initiatives.



SECTION FOUR

Getting started with AI observability

Successfully implementing end-to-end AI observability requires balancing comprehensiveness with practicality.

Organizations that operationalize AI effectively follow key patterns:

Instrument early and comprehensively

Build observability into AI systems during development, capturing prompts, outputs, model confidence scores, response times, and version metadata. The OpenLLMetry extensions built within the OpenTelemetry framework provide vendor-neutral instrumentation that simplifies implementation across languages and frameworks.

Enable end-to-end traceability

A single user request may trigger multiple model calls, vector database searches, and API requests. Trace requests through the entire chain, linking logs and traces with model versions, tracking data lineage in RAG systems, and correlating business outcomes with technical metrics.

Monitor drift and degradation

Unlike traditional software failures that trigger immediate errors, AI drift degrades gradually—a model trained on summer data may silently underperform when fall patterns emerge, or a prompt that worked well with GPT-4 may produce inconsistent results after a version update.

Establish baselines early and continuously monitor input drift, prediction drift, data quality issues, latency trends, and fairness metrics to catch degradation before it impacts users.

Define custom KPIs that reflect business value

Beyond infrastructure metrics, track indicators specific to your use case: hallucination rates, task completion rates, human intervention frequency, and domain-specific quality metrics that connect to business outcomes.

Create intelligent feedback loops

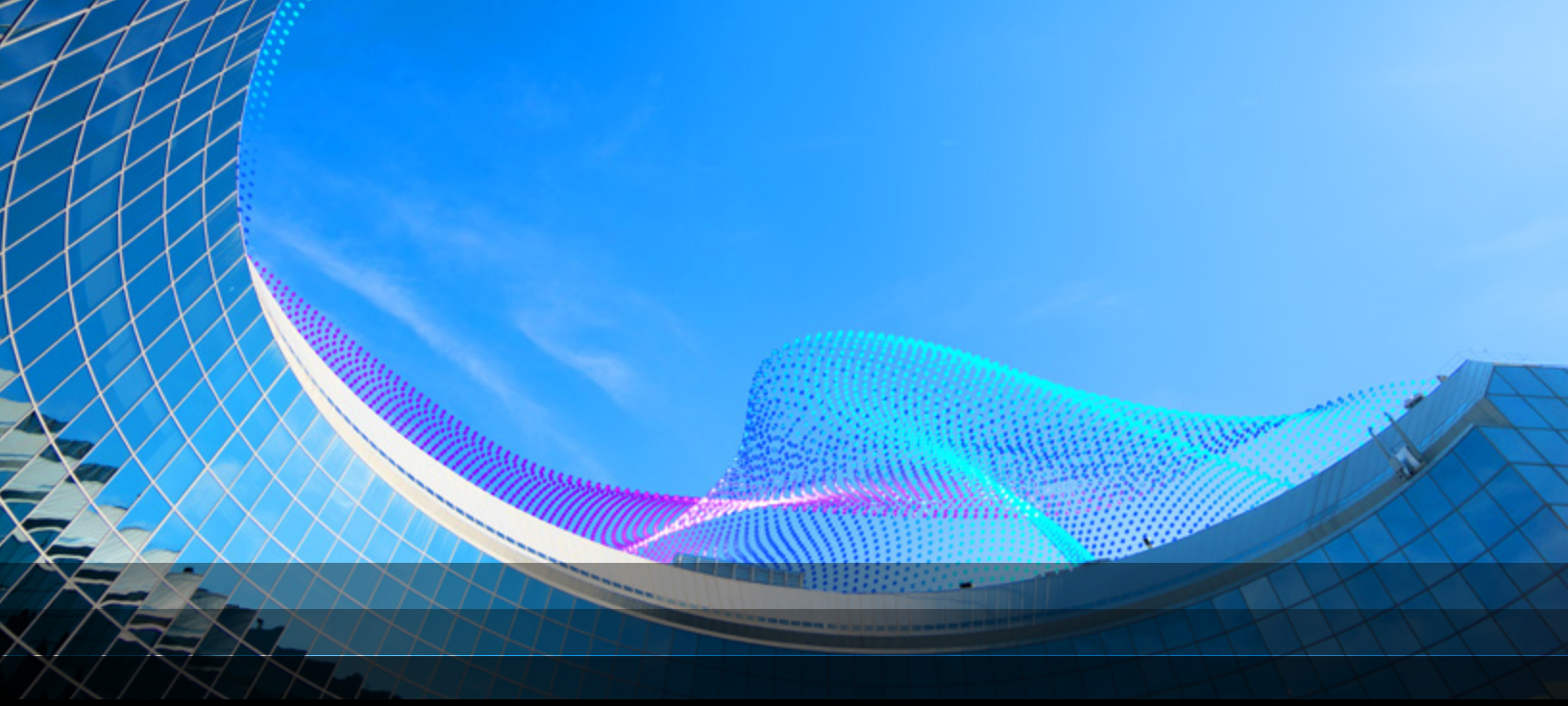
Implement context-rich alerting, automatic rollback capabilities, and adaptive thresholds. Enable users to flag problematic outputs and feed those signals back into monitoring systems for continuous improvement.

These loops enable continuous learning: when users flag a problematic response, the system can automatically correlate that feedback with the specific model version, prompt template, and context—allowing teams to identify patterns and make targeted improvements rather than relying on anecdotal reports.

Visualize everything in unified dashboards

Fragmented observability creates cognitive overhead. Correlate model behavior, infrastructure health, user experience, and cost trends in a single interface that supports different team perspectives—from SREs to product managers to data scientists.

The goal isn't to cram everything onto a single screen, but to make it easy to move from high-level indicators to detailed diagnostics without switching tools. When latency spikes, you should be able to drill down from "user experience degraded" to "vector search is slow" to "database connection pool is saturated"—all within the same interface.



CONCLUSION

The Dynatrace approach: Unified observability of the full AI stack

Dynatrace provides comprehensive AI observability across the complete stack—unifying model, agent, and infrastructure telemetry in context with conventional telemetry signals in a single platform to address the critical challenges outlined above.

Full-stack observability from a single source of truth

Dynatrace automatically traces every request from user interaction through model inference with PurePath® distributed tracing. Smartscape® topology mapping visualizes relationships across agents, models, and tool chains, while support for Model Context Protocol (MCP) and Agent2Agent (A2A) reveals how autonomous agents reason and collaborate.

Intelligent cost and performance optimization

With Dynatrace, you can monitor token usage, cost per prompt, and GPU/TPU utilization in real time. Davis® AI detects anomalies, forecasts costs, and identifies inefficiencies like batch size problems or underperforming agent loops. Grail® stores and analyzes high-cardinality telemetry at scale, enabling teams to reduce waste and right-size infrastructure.

Accelerated troubleshooting with causal AI

Davis automatically links model errors to root causes—tracing hallucinations to version rollouts or prompt changes. Davis Copilot interprets observability data through natural language queries, while Dynatrace Notebooks enable repeatable, shareable investigations.

Built-in governance and compliance

Grail maintains durable audit trails of all inputs, outputs, versions, and decisions. Integrated guardrail monitoring detects unsafe responses, PII exposure, and policy violations in real time. OneAgent® tags all telemetry with semantic context for traceability across the AI lifecycle.

Business-aligned insights

Unified dashboards correlate model performance, infrastructure health, and business KPIs. Teams visualize usage patterns, quality trends, and cost impacts without switching tools, while Davis Copilot auto-generates summaries connecting technical metrics to business outcomes.

Dynatrace delivers end-to-end oversight of AI systems—from development through production—ensuring reliability, trust, and measurable business value at scale.

NEXT STEPS

Ready to take control of your AI systems?

Choose your path forward:

GET STARTED NOW

Experience Dynatrace AI observability firsthand with a 15-day free trial. No credit card required.

Start a free trial →

EXPLORE THE AI OBSERVABILITY PLAYGROUND

See AI observability in action with interactive demos and pre-configured environments.

Visit playground →

ALREADY A DYNATRACE CUSTOMER?

Install the AI Observability app from Dynatrace Hub to start monitoring your AI systems immediately.

Install AI Observability app →

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

01.06.26 BAE21153_WP_USLet_cs