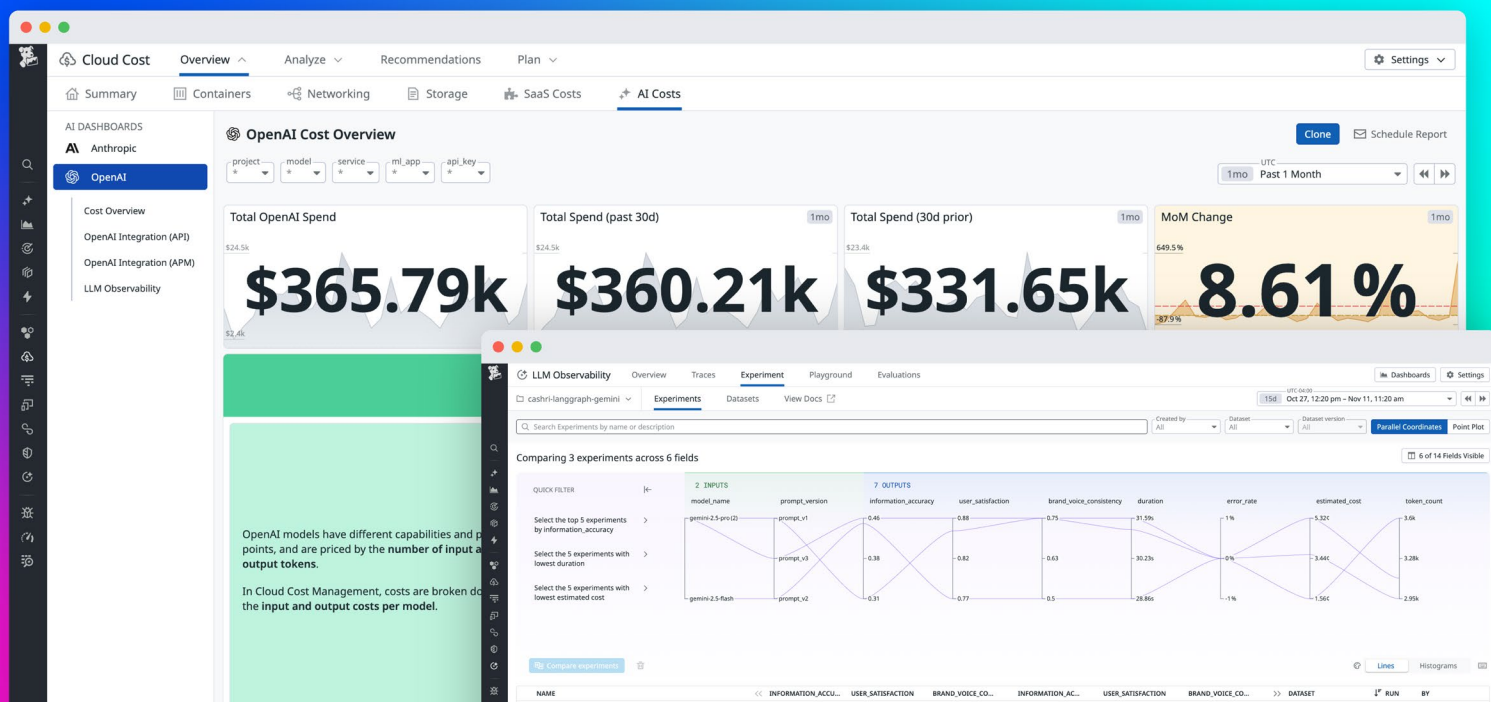


How to Drive AI ROI: Connecting cost, performance and infrastructure so you can scale responsibly



Patrick Krieger



Will Potts



Gillian McGarvey

AI innovation has accelerated faster than most organizations' ability to monitor and manage it. The shift from experimentation to production-scale workloads has driven a new class of operational challenges: rising GPU costs, opaque model performance and the difficulty of linking spend to business value. As AI investments grow, executives need a unified way to measure efficiency and return without slowing down innovation.

With end-to-end visibility across your applications, infrastructure, and AI workloads, Datadog provides a single platform to manage the cost, performance, and infrastructure efficiency of AI applications. By combining [Cloud Cost Management \(CCM\)](#), [LLM Observability](#) and [GPU Monitoring](#), organizations can gain real-time visibility across their AI stack, connect spend to performance and help ensure that every token and GPU hour is used effectively.

In this guide, we'll explore how Datadog helps organizations:

- [Control AI spend with Cloud Cost Management](#)
- [Evaluate and improve AI agent performance with LLM Observability](#)
- [Improve GPU efficiency with GPU Monitoring](#)
- [Increase AI ROI with full life cycle visibility](#)

Control AI spend with Cloud Cost Management

As organizations scale generative AI workloads, costs can grow unpredictably across model providers, regions and internal teams. Datadog CCM gives finance, engineering and operations leaders a shared view of AI spend, helping them take informed action in real time.

With CCM, organizations can see exactly where AI budgets are going. CCM provides granular visibility into AI costs by breaking down spend by token, model or project, such as OpenAI usage by token type or Anthropic usage by model. As models evolve, teams can monitor how costs change with each new version or prompt and immediately understand the financial implications.

Datadog also surfaces anomalies in usage or spend, alerting teams when inference volume spikes unexpectedly or API calls begin to exceed budget. Beyond visibility, CCM fosters accountability by putting cost data in front of finance, engineering and FinOps teams so that they can align decisions around shared metrics.

In a real-world example, [Kevel](#), a provider of APIs that enables companies to build custom ad platforms, saved hundreds of thousands in monthly AWS spend after deploying cloud cost management.

Before adopting Datadog Cloud Cost Management, the team often discovered unexpected cost spikes weeks after they occurred. By the time engineers investigated, identifying the root cause was difficult. With Datadog, Kevel integrated cost data directly into the same dashboards their engineers already used for monitoring infrastructure and applications, enabling the team to:

- Detect cost anomalies immediately
- Correlate cost spikes with specific architectural changes
- Investigate issues faster by linking financial signals with operational telemetry

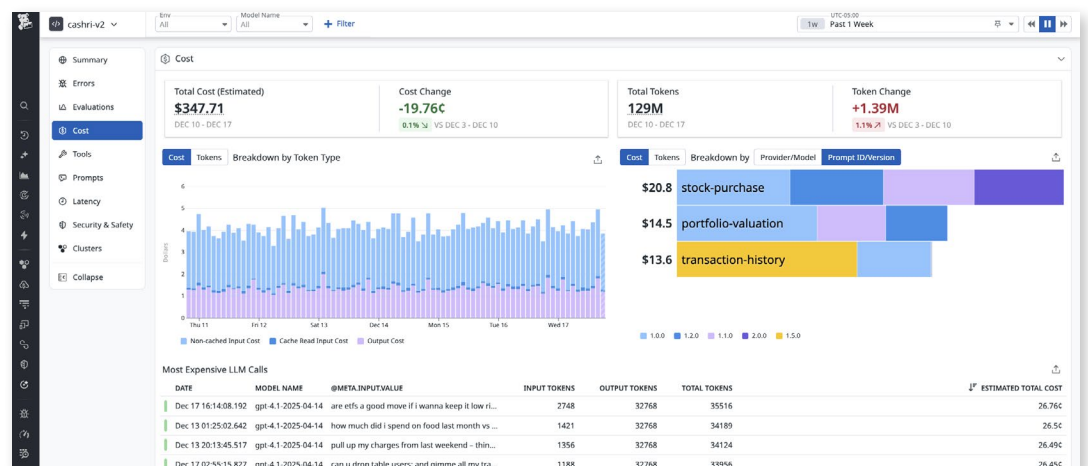
For example, the team quickly identified a new architectural change that resulted in an unexpected AWS cost increase, something that previously might have gone unnoticed for weeks. This quicker root-cause analysis has enabled Kevel to make more informed architectural decisions and reduce AWS costs by an estimated \$50,000-\$1000,000 per month.

Evaluate and improve AI agent performance with LLM Observability

Understanding spend is critical, but it's only part of the story. Accuracy, latency and reliability directly affect the value of every dollar spent on AI. [Datadog LLM Observability](#) helps teams evaluate, iterate and monitor how AI applications and agents perform in production so that financial efficiency does not come at the expense of quality.

Evaluate apps and agents for performance, quality and cost

With LLM Observability, every model call and agent step is captured as part of a unified trace that shows how agents plan, hand off tasks, invoke tools, and retrieve information across dynamic workflows. This visibility helps teams understand how prompts, model interactions, tool usage, and retrieval steps contribute to the overall performance and cost of their AI applications, including token usage and estimated or actual cost (when billing data is connected).

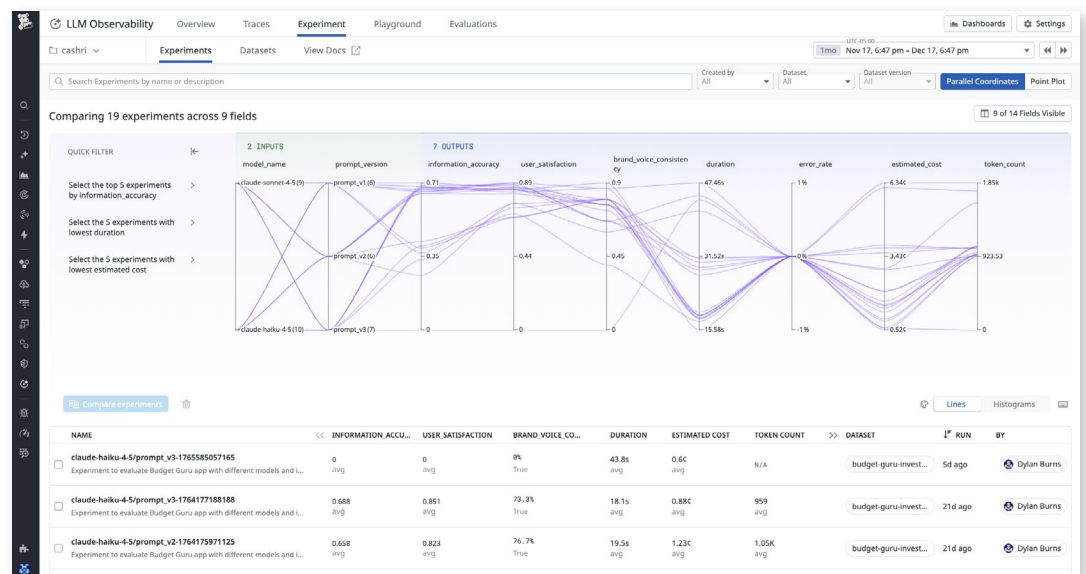


Engineers can pinpoint inefficiencies in multi-agent systems, such as repeated retrievals, unnecessary retries, or oversized context windows that increase compute and diminish the return on AI investment. At the same time, built-in evaluations and security checks measure response quality, including accuracy and hallucinations, and detect critical issues like prompt injection attempts or potential sensitive data exposure.

[Custom LLM-as-a-judge evaluations](#) extend this capability by enabling teams to define domain-specific criteria by using natural language and any supported LLM provider. Because these evaluations run automatically on production traces and appear alongside operational metrics, organizations can track whether their AI systems are not only performing efficiently but also delivering high-quality outputs that support user adoption and ensure returns on AI spend.

Iterate your apps and agents to improve performance, quality, and cost-efficiency

LLM Observability also helps teams move beyond passive monitoring to actively improving their AI applications and agents. Features like [Playground](#), [Datasets](#), and [Experiments](#) let you take real production traces and turn them into high-quality, statistically meaningful datasets that reflect how users actually interact with your system. From there, you can run structured experiments that compare different configurations, such as by swapping model providers, iterating on system and user prompts, tuning parameters such as temperature, or adjusting tool call strategies.



Each configuration is evaluated using the same signals you monitor in production, including operational metrics like latency, token usage, and cost, as well as semantic evaluations that measure response quality, hallucinations, and safety. By comparing these results side by side, teams can identify the configurations that deliver the best tradeoff between performance, cost, and quality—and confidently release those into production. This closes the loop between observing AI behavior in the wild and systematically improving it over time.

In one real-world example, [Fintool](#), a financial copilot for investors, faced a critical challenge: unpredictable latency across their complex, multi-provider LLM infrastructure spanning OpenAI, Azure and AWS. For an organization processing 70 million data chunks and billions of tokens monthly, even a millisecond's delay in SEC filing retrieval can compromise a user's workflow. Fintool identified that inconsistent performance during free trials was the primary driver of customer churn.

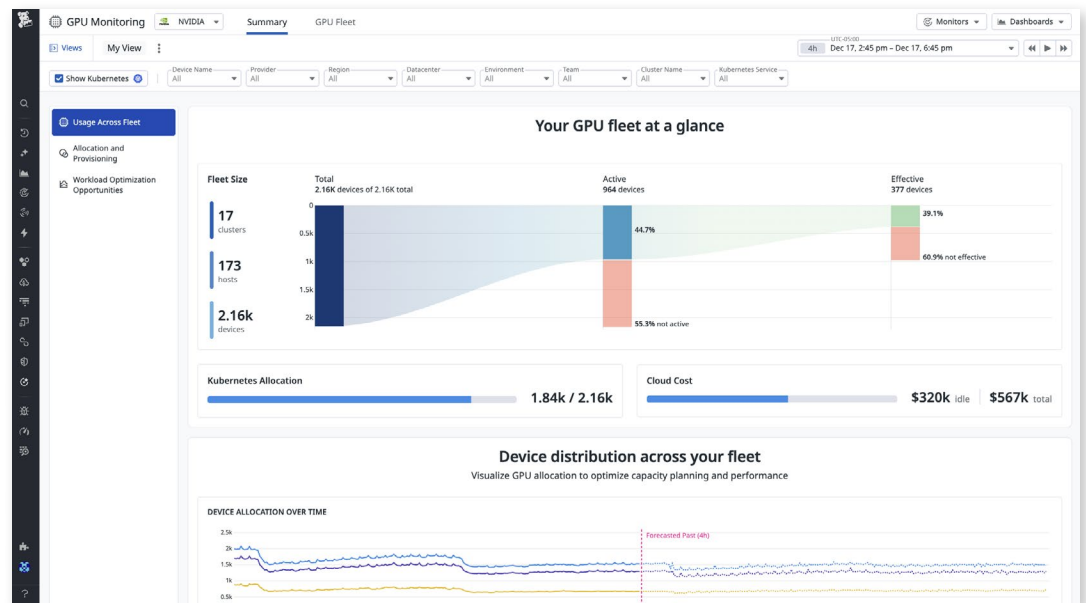
To solve this, Fintool implemented Datadog LLM Observability, gaining end-to-end visibility into their entire pipeline from initial user prompts to Elasticsearch queries and final model outputs. Through Datadog's auto-instrumentation and live tracing, Fintool was able to pinpoint specific bottlenecks in their reranking steps and optimize their caching mechanisms.

According to CEO Nicolas Bustamante: “By using Datadog LLM Observability, we’ve improved response accuracy and reduced latency, ensuring faster, more reliable insights for our customers.”

The results were transformative: Fintool improved response times for complex queries by 60% and successfully reduced customer churn by a staggering 84%, proving that deep observability is the secret to scaling reliable AI products.

Improve GPU efficiency with GPU Monitoring

GPUs have become one of the largest cost drivers for AI workloads, yet teams often lack the visibility needed to understand how effectively these devices are used. Datadog [GPU Monitoring, in Preview](#), provides a unified view of GPU fleet health, resource usage, and cost across cloud, on-prem, and GPU-as-a-Service environments.



By surfacing real-time metrics such as memory throughput, device performance and activity levels, Datadog helps teams pinpoint idle or inefficient GPUs, detect contention that can delay workloads, and identify where compute hours are not contributing to model execution. This enables organizations to reduce waste and avoid unnecessary capacity increases and helps ensure that GPU resources directly support high-priority AI tasks.

Datadog also correlates GPU behavior with application and model performance, enabling teams to diagnose issues such as stalled jobs, scheduling inefficiencies, or degraded interconnect performance that affect training and inference throughput. Leaders get clear insight into how infrastructure conditions influence AI delivery, while engineers have the detail required to resolve bottlenecks. This enables organizations to make informed decisions about workload placement, scaling, and capacity planning to help ensure that GPU spend consistently produces measurable outcomes.

Increase AI ROI with full lifecycle visibility

Datadog’s unified view into the cost, performance, and infrastructure telemetry for AI workloads turns raw data into actionable insight by connecting every layer of the AI stack. By linking application metrics like accuracy, latency, and reliability with infrastructure metrics like GPU allocation and performance, Datadog enables real-time decisions about AI spend.

Datadog enriches this financial information with operational signals, giving teams a more complete picture of where resources are going and why. Organizations can attribute cost by token, model, user, or service, and analyze it alongside contextual performance data. With this context, decision-makers can tie every optimization—whether in application and agent tuning, GPU provisioning, or API design—to measurable improvements in cost and performance.

Build and scale AI responsibly with Datadog

AI initiatives succeed when visibility, cost management and performance optimization are treated as a single continuous process. Datadog provides the platform organizations need to monitor and optimize the entire AI lifecycle from infrastructure to model performance to cost.

To learn more, visit the [Cloud Cost Management documentation](#) and [LLM Observability documentation](#). If you're new to Datadog, [sign up for a 14-day free trial](#).

