

AI Factory Observability

Understand the Complete AI Execution System, from GPU to Model Outcomes

Modern AI factories are complex execution systems spanning applications, orchestration platforms, GPU clusters, CPUs, memory, storage, high-speed networking, data fabrics, model pipelines, and large language models. As organizations scale AI initiatives, they must continuously understand where performance constraints exist, how infrastructure impacts training and inference, and what actions will optimize cost, utilization, and AI service reliability.

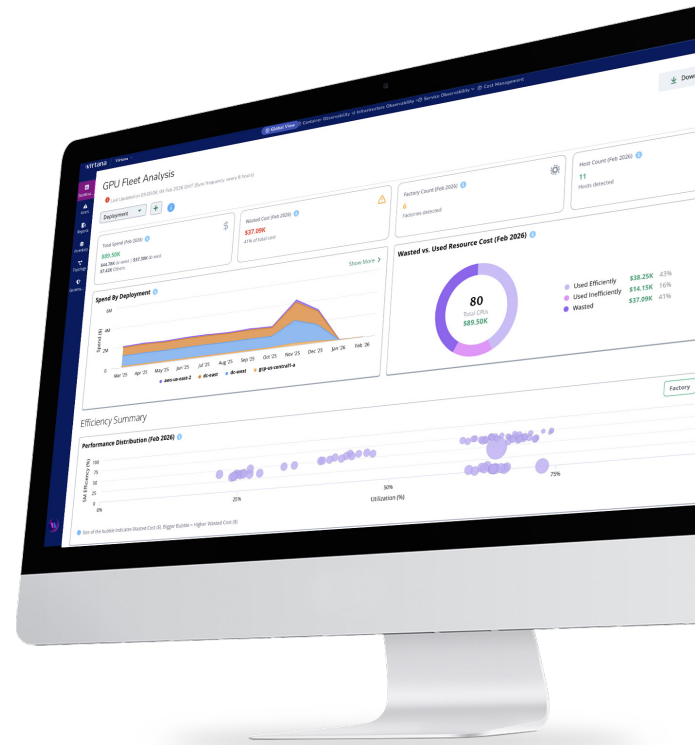
Legacy monitoring tools collect infrastructure metrics and GPU telemetry, but they present isolated views of individual components. Teams must manually correlate GPU utilization, network throughput, storage performance, model behavior, and application performance to determine why training jobs stall, inference latency increases, or infrastructure costs rise. As AI environments become larger and more distributed, troubleshooting becomes slower, GPU resources are underutilized, and operational costs increase.

Virtana AI Factory Observability, delivered through the Virtana Agentic Observability Platform, transforms AI infrastructure telemetry into system understanding. The platform continuously builds a Single Source of Truth by discovering AI applications, orchestration layers, GPU clusters, storage, networking, data fabrics, and model pipelines to create a real-time operational model of the complete AI execution system.

Diagnostic Agents leverage this unified context to perform Agentic Root Cause Analysis, reasoning across GPU infrastructure, model execution, network fabrics, data movement, workload behavior, power consumption, and historical performance to identify the operational constraint responsible for AI performance degradation. Remediation Agents then recommend or orchestrate policy-driven workflows that improve utilization, reduce operational risk, and optimize AI infrastructure efficiency.

Unlike legacy monitoring tools, Virtana continuously understands AI factories as living execution systems, connecting infrastructure behavior directly to model performance, operational efficiency, and AI business outcomes.

Legacy monitoring tools show what happened. Virtana explains why.



Key Capabilities

Full-Stack AI Visibility

Connect technical signals across AI applications, orchestration platforms, GPU clusters, CPUs, memory, storage, networking, and model pipelines into a Single Source of Truth.

Continuous discovery and dynamic topology create a real-time operational model that provides complete visibility into every layer of the AI factory, enabling teams to understand how infrastructure and model execution interact.

GPU Performance & Cost Optimization

Continuously monitor GPU utilization, memory consumption, thermal conditions, power usage, and workload efficiency across heterogeneous GPU environments.

Identify idle capacity, resource contention, throttling, fragmented memory, and scheduling inefficiencies to maximize GPU utilization while reducing infrastructure costs and improving training and inference throughput.

AI Data Fabric Observability

Automatically observe the data movement that powers AI workloads.

Analyze NVLink, InfiniBand, PCIe, Ethernet, and storage throughput to identify congestion, latency, and data transfer bottlenecks that slow distributed training and inference pipelines. Dynamic dependency mapping reveals how infrastructure bottlenecks propagate across the AI execution system.

Model Performance Monitoring

Measure the operational performance of training and inference workloads across the complete execution pipeline.

Track GPU cycles, memory efficiency, workload execution, latency, throughput, and energy consumption while comparing workload efficiency across environments. Detect regressions before they affect production AI services.

Token Economics & Forecasting

Understand the operational cost of AI workloads beyond infrastructure utilization.

Track token consumption, model usage, and AI service costs while forecasting demand and optimizing resource allocation across cloud, hybrid, and on-premises AI environments. Connect operational efficiency directly to AI business value.

Power & Sustainability Intelligence

Monitor power consumption, thermal efficiency, and energy utilization across GPU clusters.

Identify energy waste caused by idle resources, thermal throttling, or inefficient workload placement. Support sustainability initiatives while maximizing AI infrastructure performance and efficiency.

AI Security & Guardrails

Monitor AI execution for operational anomalies, latency spikes, policy violations, and unexpected model behavior.

Track AI agent activity, invocation behavior, intervention events, and operational guardrails to support secure, compliant, and governed AI operations across enterprise and sovereign environments.

Agentic Root Cause Analysis

Using the platform's unified operational model, Diagnostic Agents reason across GPU infrastructure, AI orchestration, network fabrics, storage systems, workload behavior, and model execution to identify the true operational constraint behind AI performance degradation.

Rather than simply correlating infrastructure alerts, they verify root cause, explain failure propagation across the AI execution system, determine operational and business impact, and provide evidence-based recommendations that accelerate resolution.

Autonomous AI Operations

Accelerate AI operations with AI-generated incident summaries, dependency intelligence, operational evidence, and recommended actions.

Remediation Agents recommend or orchestrate policy-driven workflows that optimize GPU allocation, improve workload scheduling, reduce operational costs, and accelerate issue resolution while maintaining governance and human oversight.

Business Outcomes

Organizations using Virtana AI Factory Observability can:

- Reduce Mean Time to Resolution (MTTR) for AI infrastructure incidents
- Maximize GPU utilization while reducing idle capacity
- Improve training and inference performance
- Detect infrastructure bottlenecks before they affect AI workloads
- Optimize token consumption and AI operating costs
- Improve energy efficiency across GPU environments
- Increase AI infrastructure utilization while reducing operational expense
- Accelerate AI troubleshooting through Agentic Root Cause Analysis
- Enable governed automation across AI operations
- Establish the operational foundation for autonomous AI infrastructure management

Why Virtana AI Factory Observability?

Full-Stack AI Visibility

Understand AI applications, orchestration platforms, GPU infrastructure, storage, networking, and model pipelines as one connected execution system.

Single Source of Truth

Maintain a continuously updated operational model connecting AI workloads, infrastructure resources, dependencies, and operational context.

Agentic Root Cause Analysis

Move beyond GPU dashboards and infrastructure alerts. Diagnostic Agents identify the true operational constraint, explain why AI performance changed, determine downstream impact, and recommend the next best action.

GPU Performance & Cost Optimization

Improve infrastructure efficiency by identifying idle GPUs, resource contention, fragmented memory, thermal throttling, and scheduling inefficiencies before they affect model performance.

AI Infrastructure Intelligence

Correlate GPU, CPU, storage, networking, orchestration, and model behavior to optimize the complete AI execution system rather than isolated infrastructure components.

Governed AI Operations

Use Remediation Agents to guide or automate operational workflows while maintaining enterprise governance, policy controls, and operational oversight.

About Virtana

Virtana, provider of the deepest and broadest Agentic Observability Platform for hybrid and multi-cloud environments, delivers system-aware observability spanning applications, services, data pipelines, GPUs, CPUs, networks, and storage. Powered by high-fidelity operational data and agentic AI, Virtana provides unmatched visibility across end-to-end IT services and AI execution systems, correlating health, performance, cost, and user impact in real time. With advanced event intelligence and AI-driven reasoning, Virtana delivers clarity no other provider can match. Trusted by Global 2000 enterprises and public sector organizations, Virtana helps IT operations, Platform Engineering, SRE, and DevOps teams reduce risk, strengthen resilience, improve efficiency, and modernize with confidence across multi-cloud, on-premises, edge, and sovereign environments.

Learn more at virtana.com