virtana

2026 Research Report

AI FACTORY REALITYCHECK

How enterprises are scaling AI faster than the systems managing it

Executive Summary

In March 2026, Virtana's "AI Is Breaking Human-Managed Operations" research found three in four enterprises reporting double-digit AI job failure rates, with executive confidence rising while operational fragility rose faster. The study found that human-scale operations cannot sustain machine-scale AI systems.

Two months later, a more consequential pattern has emerged. AI factories are scaling faster than the systems managing them.

Enterprises are racing to deploy AI factories, acquiring GPUs as quickly as they become available while rapidly building the infrastructure to support them. Yet Virtana's "AI Factory Reality Check," based on a survey of 788 U.S.-based decision-makers and practitioners, finds that the governance, operational controls, and system-level observability needed to run AI factories at scale are failing to keep pace. As AI demands accelerate, organizations are struggling to establish the visibility, accountability, and control necessary to understand, optimize, and govern AI systems, exposing a widening gap between AI factory expansion and the operational foundation needed to sustain it.

This study examines the condition of the systems now carrying this load. The findings show that more than half of enterprises are already scaling AI across teams, with many managing the most operationally demanding state: running live production workloads while expanding infrastructure at the same time.

At the same time those systems are scaling, they are being actively reconfigured. Rising hardware costs are pushing enterprises to rebalance workload placement across the hybrid environments they already operate and to consolidate infrastructure in pursuit of efficiency. These changes are happening while systems are live, under load, and expanding, introducing additional complexity into environments that are already difficult to observe and control.

The operational systems required to manage this complexity have not kept pace. Teams responsible for AI factory environments are operating with limited visibility, inconsistent performance baselines, and a continued reliance on manual investigation when systems fail.

Where the March research found a divide in readiness confidence, this study finds one in operational reality. Practitioners report materially lower levels of visibility and control than the leaders authorizing their expansion, and that divergence shapes investment decisions made against an incomplete picture of how these systems behave.

Enterprises are under increasing pressure to demonstrate returns on significant AI infrastructure investments. The instrumentation required to produce that proof, including cost governance, performance visibility, and cross-domain accountability, is precisely what is being deferred.

Organizations are deploying AI systems with the expectation they will deliver business value, yet they lack the operational visibility to prove those outcomes are occurring. You cannot govern what you cannot see.

Capital continues to flow toward models, applications, and outputs, the most visible layers of AI. When AI factories are treated as outputs rather than systems, risk accumulates in the least visible layers, including infrastructure orchestration, workload coordination, resource utilization, and operational governance. You cannot govern what you cannot see, and the data shows enterprises are scaling systems they cannot yet fully see.

What enterprises report needing is consistent across roles and revenue bands: unified visibility across all layers of the AI factory and automated root cause analysis that reduces dependence on manual investigation. Until that foundation is in place, enterprises will continue to scale AI systems faster than they can operate them.

THE HEADLINES

Key Findings

Scale & Operational State

- 54% of enterprises are already scaling AI across teams, with a further 23% managing live production workloads alongside active infrastructure expansion, the most operationally demanding state of all.

The Visibility Crisis

- 1 in 4 enterprises still runs a manual investigation across tools as its first response when an AI workload underperforms or fails. Teams are manually assembling context across disconnected consoles while expensive compute sits underutilized.
- Nearly 6 in 10 enterprises (59%) cannot automatically identify root cause across all infrastructure domains when an AI workload alert fires. Only 41% have automated, cross-domain root cause identification in place.
- 69% of Infra/SRE practitioners cannot automatically identify root cause across all infrastructure domains, compared to 52% of executives reporting the same limitation, a 17-point divergence between the teams managing the factory floor and the leaders authorizing its expansion.
- Two-thirds of enterprises are operating AI factory infrastructure without reliable performance baselines. Only 34% describe AI workload performance as highly predictable, a figure that falls to 25% at organizations above 50,000 employees. The larger the factory, the harder it is to anticipate what it will do next.
- Cost and efficiency metrics (57%) and GPU utilization tracking (56%) are the top two monitoring challenges. Organizations are acquiring GPU infrastructure at significant cost without the visibility to confirm it is running efficiently.
- Monitoring visibility gaps reach every layer of the AI factory stack: data pipeline visibility (52%), storage and throughput (47%), network bottleneck detection (44%), and lack of centralized dashboards (34%) follow closely behind cost and GPU tracking.

Where Governance Breaks Down

- Managing GPU cost and utilization is the hardest operational challenge for more than 1 in 3 enterprises (35%). Executives feel it as financial accountability pressure (39%). Architects feel it as the complexity of integrating high-performance hardware into distributed environments (36%). Infra/SRE teams feel it as the daily challenge of scaling reliably (22%).

Hardware Costs Reshaping Investment

- 80% of enterprises say the cost of premium AI hardware has changed how they approach infrastructure investment decisions, the highest rate of any segment in the study. Among those affected, 60% are optimizing workload placement across hybrid environments and 58% are accelerating consolidation to improve per-unit efficiency.
- 31% of enterprises cite clearer ROI metrics from existing AI investments as the single most important prerequisite for confidently scaling AI in the next 12 months, ahead of IT-business alignment (22%), infrastructure visibility (21%), and mature tooling (18%).

The Path Forward

- 70% of enterprises converge on the same two immediate operational priorities: a unified platform with visibility across all AI and infrastructure layers (38%) and AI-powered root cause analysis that works without manual correlation (32%). This is consistent across all role groups and revenue bands.

SECTION 1

AI Factories Are Scaling Without the Systems to Control Them

Enterprise AI has entered its operational phase. More than half of enterprises are already scaling AI across teams, and a further 23% are running live AI workloads in production while simultaneously expanding their infrastructure footprint. These organizations face the most concentrated operational demands to maintain production reliability while continuing to grow the surface they are responsible for.

54%

Already scaling AI across teams

23%

Running production workloads while expanding infrastructure

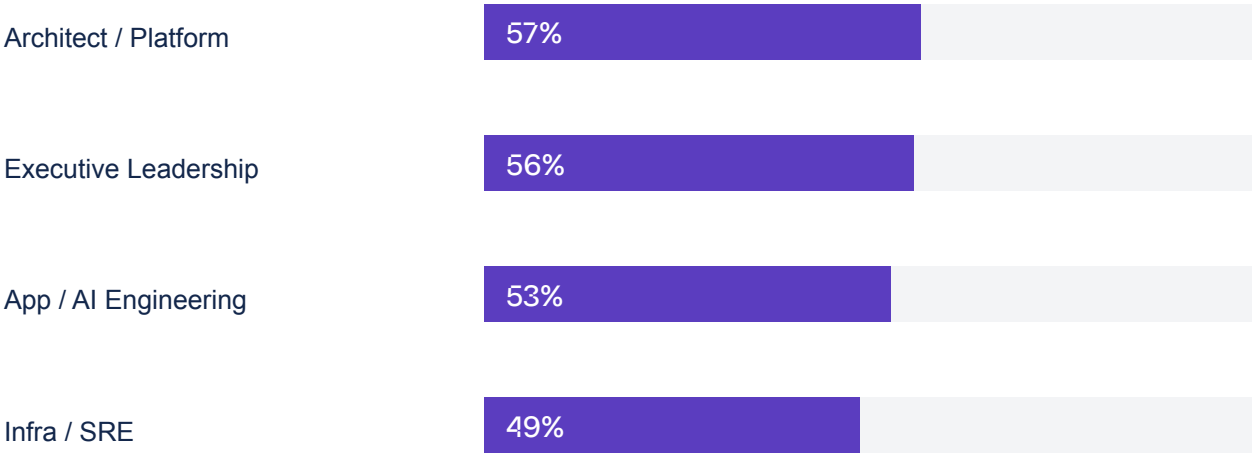
14%

Operating early production workloads

At the largest enterprises, the complexity intensifies. Organizations above $10 billion in revenue are the most likely to be managing simultaneous production and expansion, reflecting AI factories of exceptional scope.

As factories scale, adding GPU clusters, expanding data pipelines, and onboarding concurrent workloads, the operational surface grows faster than the observability infrastructure managing it. The practitioner's view makes this visible. Infra/SRE teams report scaling at 49%, the lowest of any role group, with 21% still managing early production workloads. Executives report scaling at 56%. The people running the factory floor and the leaders overseeing it are describing two different operational environments.

AI journey stage by role

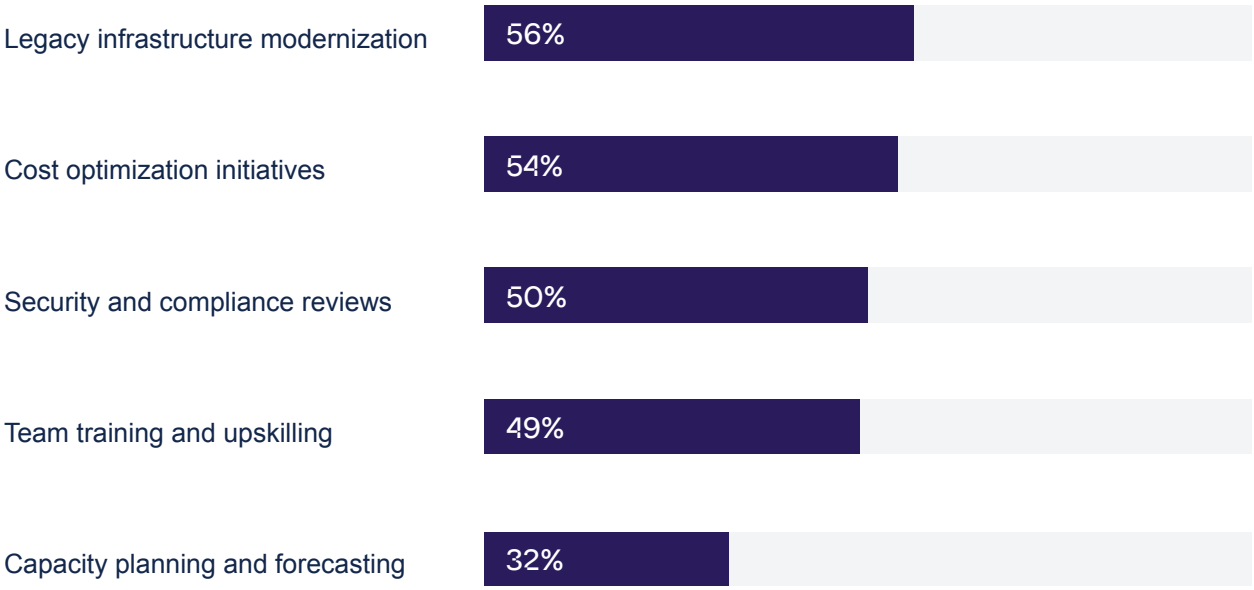


Percentages indicate respondents reporting their organization is scaling AI across teams.

Governance and controls are being deferred at the same time

As AI factory demands grow, the work required to establish visibility, accountability, and control is being sidelined. Cost optimization is deprioritized by 54%, legacy infrastructure modernization is deferred by 56%, and security and compliance reviews are scaled back by 50%. Enterprises are racing to build out AI factories while simultaneously cutting the governance and instrumentation work that would let them operate those factories with confidence.

What teams are deprioritizing as AI factory demands grow



Security and compliance reviews are being deprioritized by 50% of US enterprises.

SECTION 2

The Investment Picture Behind the Build-Out

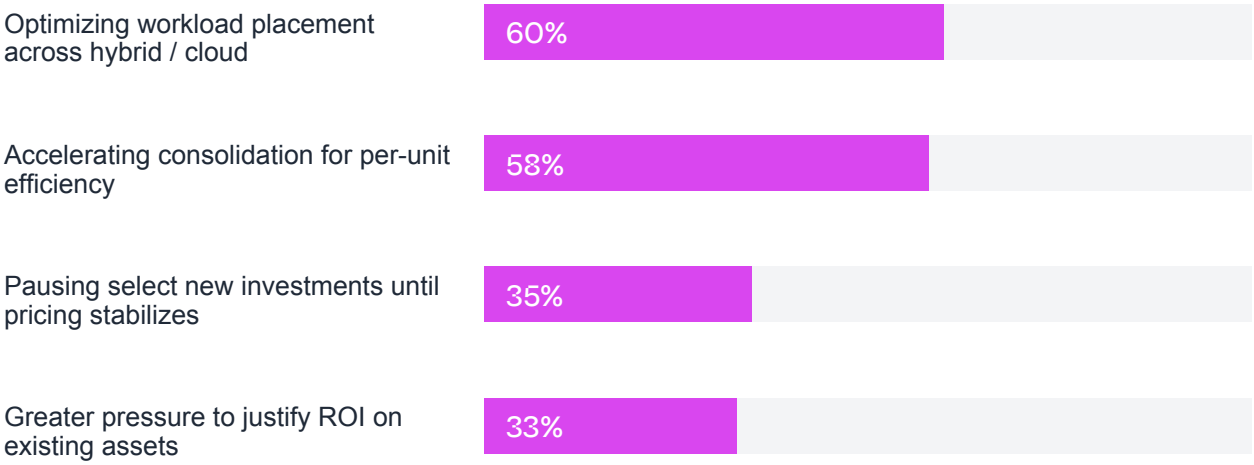
The build-out documented in the previous section is happening against a financial backdrop that shapes how enterprises approach every infrastructure decision. Premium AI hardware is in short supply and high demand, and enterprises are paying close attention to how every dollar of capacity is allocated. They are rebalancing workload placement across the hybrid environments they already operate and consolidating systems to improve per-unit efficiency, all while AI factories are live, under load, and continuing to expand. This operational change is happening in motion and not in controlled conditions.

80% of enterprises say the cost of premium AI infrastructure has changed how they approach investment decisions

Eight in ten enterprises say the cost of premium AI hardware has changed how they approach infrastructure investment decisions. Among those affected, 60% are optimizing workload placement across the hybrid environments they already operate, and 58% are accelerating consolidation to improve per-unit efficiency. A further 35% have paused select new investments until pricing stabilizes, and 33% report greater pressure to justify ROI on existing assets. These are the responses of organizations feeling the weight of AI infrastructure spend and actively managing it.

How rising hardware costs are shaping infrastructure decisions

Among the 80% of enterprises affected by rising hardware costs (n=630). Respondents could select all that apply.

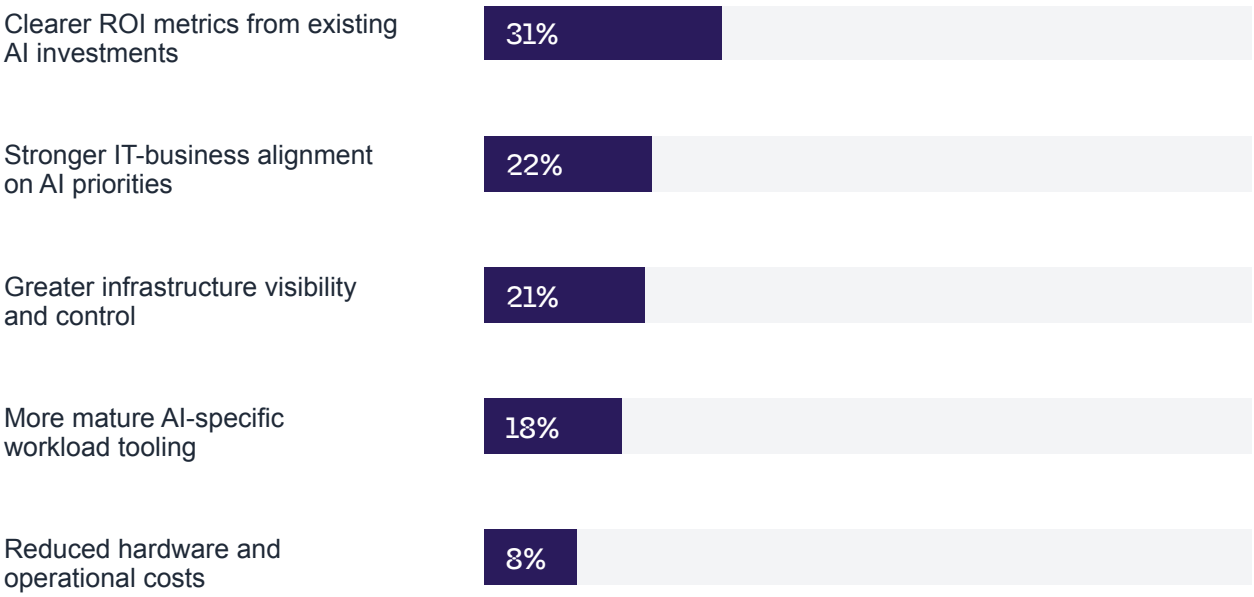


These responses are not exclusive of one another. Enterprises operating AI workloads at scale are running on-premises GPU clusters, virtualized environments, and public cloud concurrently, and the cost pressure is prompting them to use each more deliberately rather than to shift wholesale from one to another.

The ROI prerequisite

More than any other factor, enterprises identify clearer ROI as the primary prerequisite for scaling AI further, cited by 31%, ahead of IT-business alignment (22%), infrastructure visibility (21%), and mature tooling (18%). The demand is for proof that the factory is working, not for more capability.

What needs to be true to confidently scale AI in the next 12 months



ROI clarity rises to 36% as the top prerequisite among the largest enterprises ($10B+ revenue).

Yet this requirement is running directly into a paradox. The organizational capabilities most critical to generating ROI proof (i.e, cost visibility, performance predictability, and cross-domain accountability) are the exact areas where enterprises are deprioritizing investment.

- 57% of teams struggle with cost and efficiency metrics visibility
- 54% are deprioritizing cost optimization as AI demands grow

Organizations cannot prove AI outcomes because they are systematically reducing their capacity to observe and measure them.

Enterprises are accelerating AI infrastructure investments, but the systems needed to instrument those environments and prove business value remain fragmented, manual, and under-resourced.

> Visibility is the prerequisite for operational control. Enterprises need evidence that their AI factories are working at the scale they have built. The instrumentation required to produce that evidence is what is being deferred.

SECTION 3

Inside the AI Factory, Where Control Breaks Down

The operational consequences of this imbalance are increasingly visible in how AI factories operate at scale and how effectively organizations can manage performance, reliability, and cost. These challenges emerge in how teams respond to failures, predict workload behavior, understand resource utilization, and maintain visibility across the full execution stack.

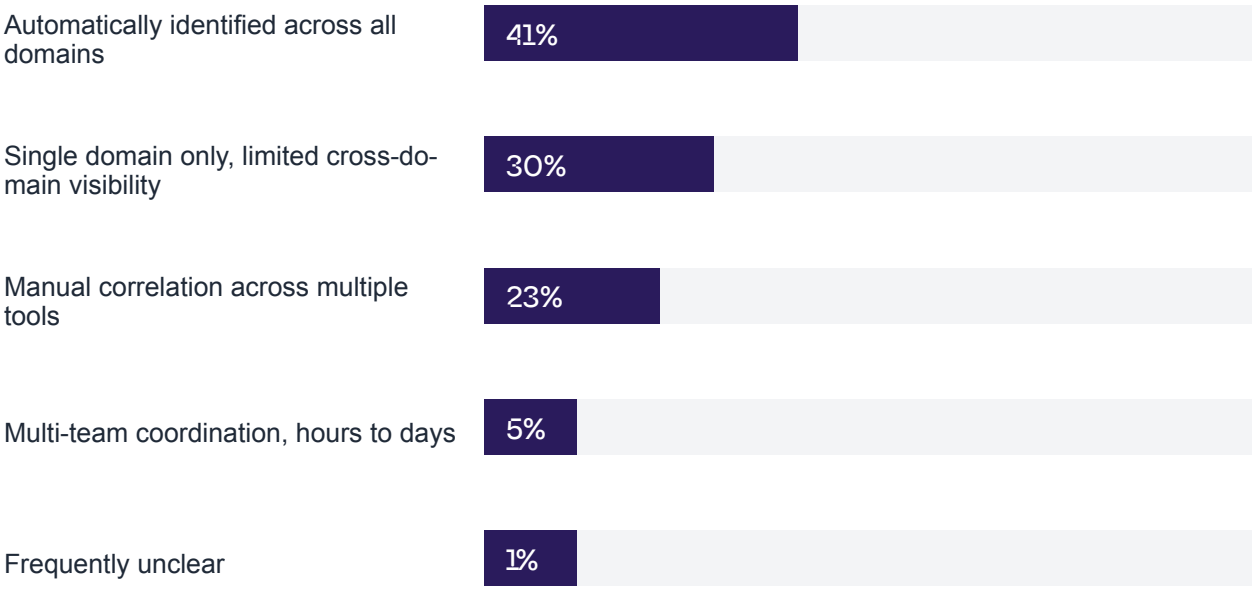
Without system-level observability, organizations struggle to understand how AI outcomes are produced, where constraints originate, how infrastructure resources are consumed, and whether system behavior can be consistently trusted. This section examines where visibility gaps emerge across incident response, performance forecasting, and full-stack observability, and how those gaps limit organizations' ability to operationalize and evaluate the effectiveness of their AI infrastructure investments.

When something breaks

From a business outcome perspective, manual investigation creates a proof problem. When an AI-driven service fails or underperforms, the organization cannot quickly establish why, and therefore cannot prove whether the failure originated in infrastructure, data pipelines, the model itself, or operational decisions. Instead, teams are forced to manually reconstruct events across disconnected tools while expensive GPU and infrastructure resources sit underutilized waiting for resolution.

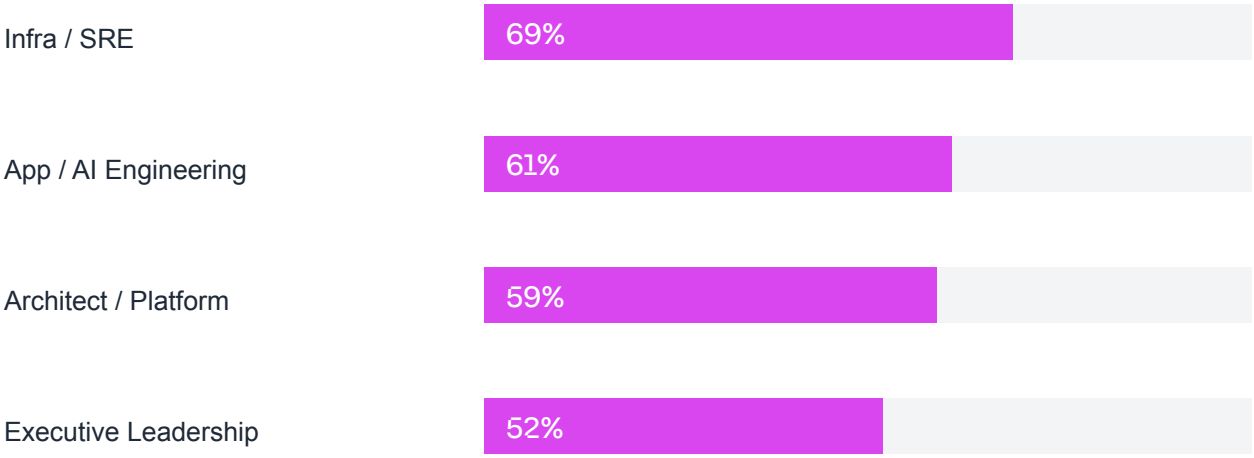
Without automated, system-aware diagnosis, the organization loses the ability to confidently explain outcomes, accelerate remediation, or demonstrate operational control over the AI environment. The result is prolonged incidents, inefficient infrastructure utilization, unproven AI ROI, and declining confidence in the reliability and efficiency of the AI factory.e tooling cannot accelerate.

After an alert fires, how effectively can your organization determine root cause?



The divide between how executives and practitioners experience this is the sharpest in the dataset. More than half (69%) of Infra/SRE practitioners, the engineers who field those alerts, report their organizations cannot automatically identify root cause across all domains. Among executives, 52% say the same. That 17-point difference reflects what it feels like to be the person on call when a GPU workload degrades at 2am with no automated diagnosis to work from.

Cannot automatically identify root cause across all domains, by role



The distance between what leadership believes about diagnostic capability and what practitioners experience is widest where the operational stakes are highest.

Predicting what the factory will do next

Beyond incident response, there is the n
workloads will behave from one run to the next. Performance variability makes capacity planning approximate, SLA commitments unreliable, and cost forecasting a guessing exercise. Teams managing AI factories at scale are responsible for infrastructure whose behavior they cannot consistently anticipate.

34%

Describe AI workload performance as highly predictable

26%

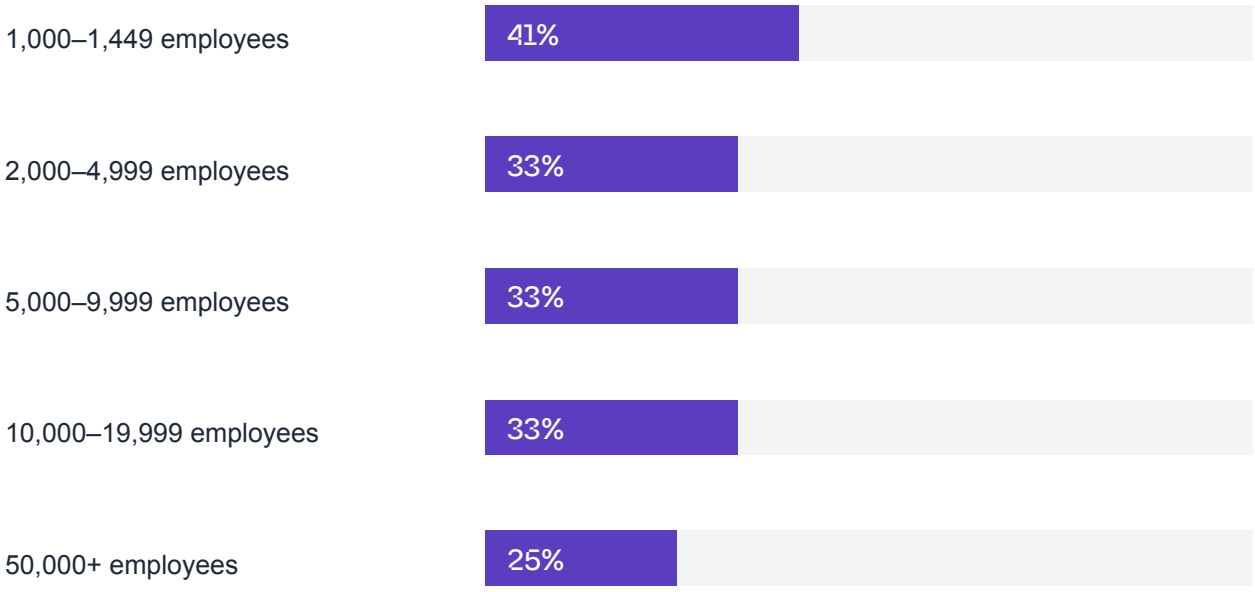
Highly predictable among Infra/SRE practitioners

25%

Highly predictable at organizations above 50,000 employees

Predictability stands at 41% among organizations with 1,000–1,449 employees and falls to 25% above 50,000. Teams managing the largest AI factories have the least ability to anticipate what those factories will do next.

AI workload performance predictability by company size

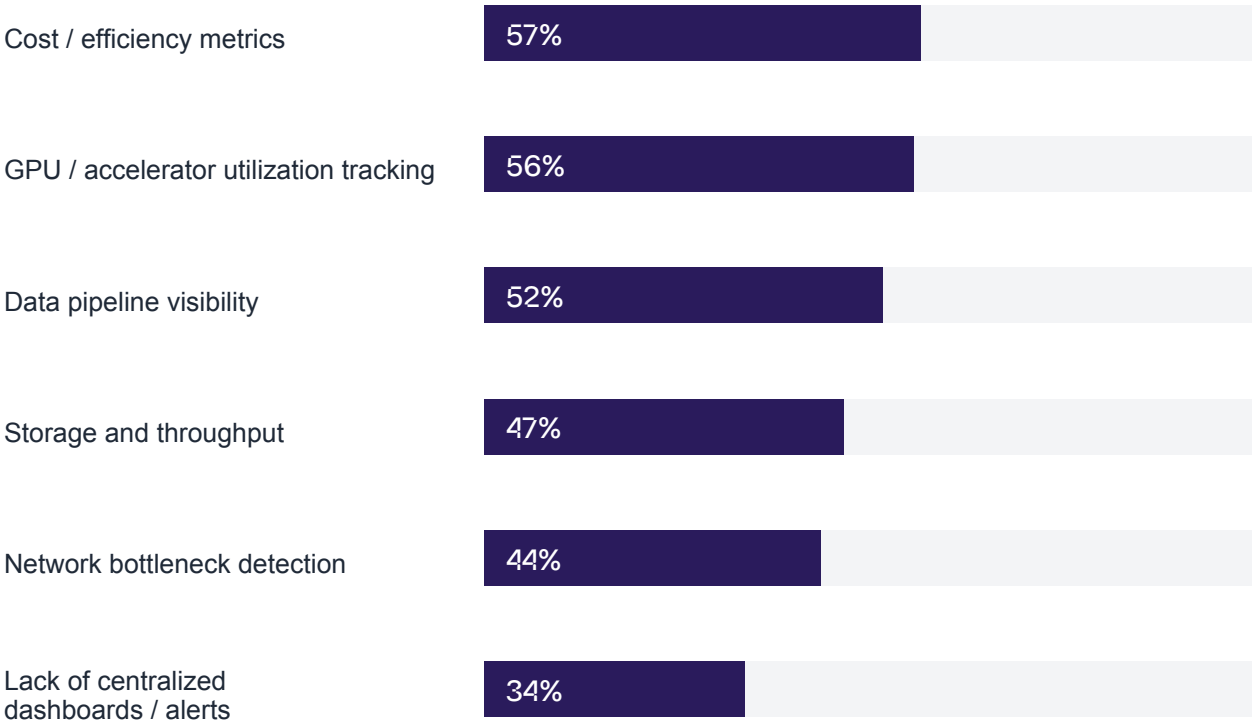


Percentages indicate respondents describing AI workload performance as highly predictable.

What teams can and cannot see

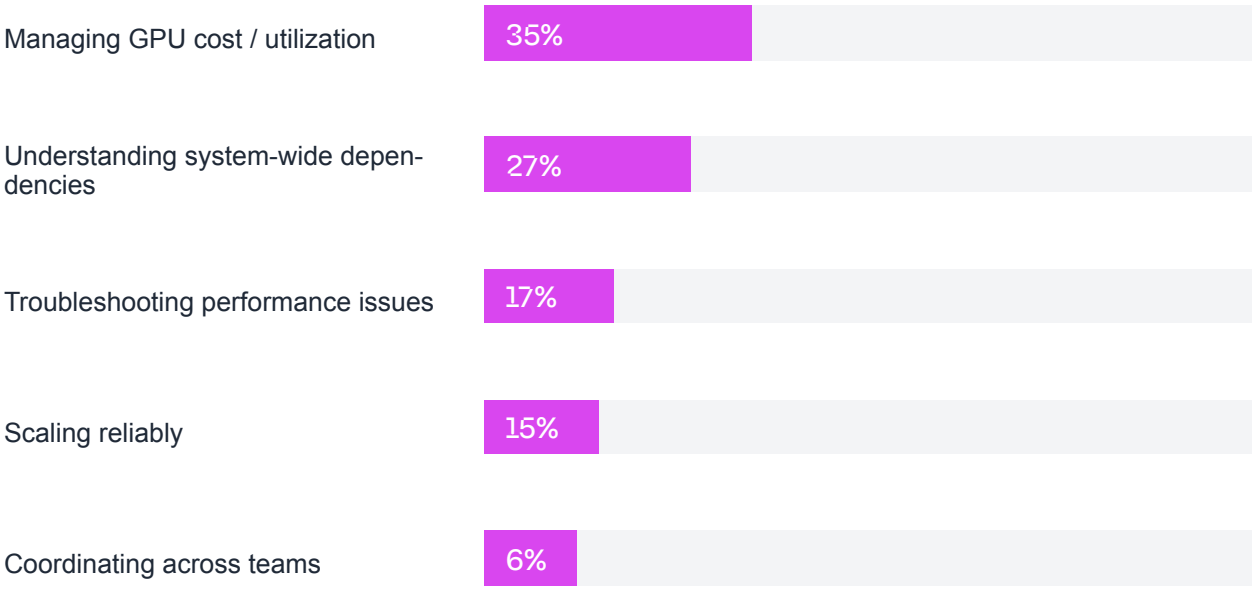
Cost and efficiency metrics (57%) and GPU utilization tracking (56%) are the top two monitoring challenges. GPU infrastructure is being acquired and deployed, but the instrumentation to know whether it is running efficiently is a challenge across every role group. The breadth of what teams cannot see confirms that the visibility problem runs the full length of the AI factory stack.

Top monitoring challenges across the AI factory stack



Each team encounters a different face of the same full-stack problem. Infra/SRE teams lead on network bottleneck detection (48%), reflecting where they operate in the stack. Architects lead on data pipeline visibility (62%), reflecting the supply chain view of the AI factory. Executives and App/AI Engineering both flag GPU utilization tracking at 61% and 60% respectively, the asset accountability layer where financial and operational pressure converges. Managing GPU cost and utilization is the hardest operational challenge for 35% of enterprises overall; executives feel it as financial accountability pressure (39%), architects as architectural complexity (36%), and Infra/SRE teams as the daily challenge of scaling reliably (22%).

Hardest aspect of operating AI infrastructure today



GPU cost and utilization is hardest at <$500M (45%), declining to 21–22% at $7B+, where system-wide dependencies become the top concern.

SECTION 4

What Enterprises Need

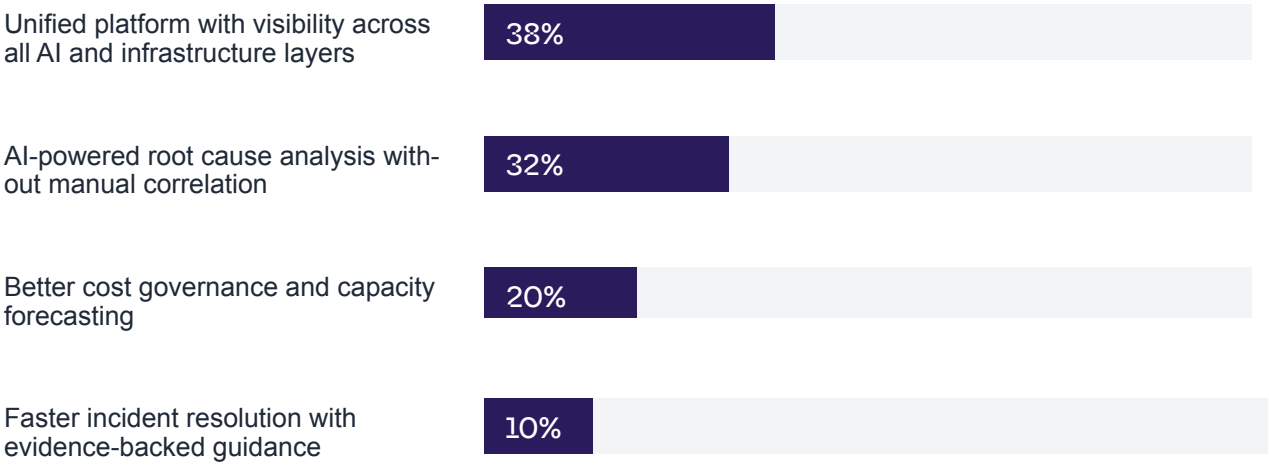
When asked what would have the most immediate impact on their ability to scale AI operations, 70% of enterprises converge on the same two priorities: a unified platform with visibility across all AI and infrastructure layers (38%) and AI-powered root cause analysis that works without manual correlation (32%). The consistency of this finding across role groups and revenue bands reflects a shared operational condition.

What unites these priorities is the need for operational accountability across the AI factory. Teams cannot consistently explain what their AI systems are doing, why performance degraded, or where constraints originated across infrastructure, data pipelines, models, and orchestration layers. At the same time, leaders are under pressure to justify growing investments in GPU infrastructure and AI operations that remain difficult to fully instrument and optimize.

- Without unified visibility across the full stack and AI-powered root cause analysis that eliminates manual correlation, organizations cannot:
- Demonstrate whether AI-driven services are meeting performance, reliability, and SLA commitments
- Control and optimize the cost, utilization, and efficiency of GPU-intensive AI operations
- Rapidly identify and prove the root cause of failures across infrastructure, data, models, and orchestration layers
- Establish confidence that AI outcomes are reliable, trustworthy, and operationally sound
- Maximize ROI from AI infrastructure investments by reducing operational inefficiency and accelerating remediation

70% of enterprises identify unified visibility plus automated root cause analysis as the most immediate operational priority

Most immediate impact on the ability to scale AI operations



Until enterprises can see inside their AI factories in real time, correlate signals across the full stack, and resolve what goes wrong without depending on manual effort across fragmented tools, they will continue scaling AI systems faster that they can operate them, and faster than they can prove they work.

HOW VIRTANA HELPS

Virtana's AI Factory Observability platform delivers the visibility, governance, and accountability layer for the AI factories enterprises are building today. It provides cross-domain visibility, automated root cause analysis, and cost and efficiency instrumentation across on-premises GPU infrastructure, virtualized environments, and public cloud, built for the teams managing AI factories at the scale this study documents, and for the leaders accountable for what those factories produce.

METHODOLOGY

Demographics

This study is based on responses from 788 US-based professionals at enterprise organizations actively running, piloting, or planning AI workloads in production. All respondents have decision-making authority or significant influence over IT infrastructure, AI strategy, or technology investment within their organizations.

By role

Application, Service & AI Engineering	307
Executive Leadership	270
Infrastructure, Cloud & Reliability Engineering	120
Architect & Platform Design	91

By company size

2,000–4,999 employees	302 (38%)
5,000–9,999 employees	141 (18%)
1,000–1,449 employees	103 (13%)
10,000–19,999 employees	72 (9%)
1,500–1,999 employees	68 (9%)
More than 50,000 employees	36 (5%)
30,000–50,000 employees	25 (3%)
Under 1,000 employees	21 (3%)
20,000–29,999 employees	20 (3%)

By annual revenue

$500 million – $1 billion	30%
$1 billion – $3 billion	21%
Less than $500 million	14%
$3 billion – $5 billion	11%
More than $10 billion	10%
$5 billion – $7 billion	9%
$7 billion – $10 billion	5%

By environment size

1,000–10,000 VMs	51%
100–1,000 VMs	39%
10,000+ VMs	7%
Under 100 VMs	3%

Investment authority

IT / infrastructure leadership	85%
Finance / CFO office	9%
Shared or unclear process	6%

ABOUT VIRTANA

Virtana delivers the deepest and broadest observability platform for hybrid and multi-cloud, with full-stack AI observability spanning applications, services, data pipelines, GPUs, CPUs, networks, and storage. Powered by high-fidelity data and agentic AI, Virtana provides unmatched visibility across end-to-end IT services and AI workloads, correlating health, performance, cost, and user impact in real time.

With advanced event intelligence and autonomous insight generation, Virtana delivers clarity no other provider can match. Trusted by Global 2000 enterprises and public sector organizations, Virtana helps IT operations and DevOps teams reduce risk, strengthen resilience, improve efficiency, and modernize with confidence across multi-cloud, on-premises, and edge environments.

Learn more at **virtana.com**

info@virtana.com | **+1 408-579-4000** | **virtana.com**