virtana

2026 Research Report

AI FACTORY REALITYCHECK

How UK enterprises are scaling AI faster than the systems managing it

United Kingdom Edition

Executive Summary

In the UK, AI factories are scaling faster than the systems built to govern and operate them.

The UK enterprise AI market is entering a new phase. Organizations are moving beyond experimentation and deploying AI into production while operating in one of the world's most demanding governance environments. UK enterprises must balance accelerating AI adoption against obligations spanning UK GDPR, sector-specific regulation, emerging AI governance frameworks, and increasing expectations around data, cloud, and AI sovereignty.

At the same time, AI infrastructure is becoming more distributed across on-premises environments, sovereign cloud deployments, and hyperscale platforms. As organizations expand AI factories, the challenge has expanded from simply building AI capability to also proving control, accountability, performance, and compliance across increasingly complex environments.

This study examines how UK enterprises are navigating that transition and where operational readiness is failing to keep pace with AI ambition. It found that the shortfall in visibility, attribution, and diagnosis runs the full length of the AI factory stack. More than half of UK enterprises cannot automatically identify root cause across all infrastructure domains when an AI workload alert fires, and only about a quarter describe AI workload performance as highly predictable. The teams closest to the factory floor report materially lower diagnostic capability than the leaders authorising infrastructure expansion.

Security and compliance reviews are being deprioritised by 39% of UK enterprises, a material exposure for organisations operating under UK GDPR, emerging AI Act obligations, and sector-specific regulatory frameworks. In the UK, operational observability and regulatory accountability are the same question. The systems that prove an AI factory is working are the systems that satisfy an auditor, and the same systems that turn a cloud or AI sovereignty commitment into something an organisation can evidence across every environment it runs in.

A vast majority of UK enterprises (70%) name the same two priorities ahead of everything else: a unified platform with visibility across all AI and infrastructure layers and AI-powered root cause analysis that eliminates the need for manual correlation. What remains is closing the distance between the AI factories UK enterprises have built and the systems they need to operate, govern, and prove them.

Key Findings

59% of UK enterprises are already scaling AI across teams, and the absence of operational controls is causing operational problems.

- A further 13% are running live production workloads while expanding their infrastructure at the same time. These enterprises carry the highest exposure when controls lag, because they cannot pause either the workload or the build-out to catch up.
- Scaling jumps sharply once organisations pass $500M in revenue, peaks at 70% among $1B to $3B organisations, and holds in a 59% to 65% range across the largest UK enterprises.

UK enterprises rank cost and efficiency metrics as their hardest monitoring problem, and 54% are cutting cost optimisation as AI demands grow.

- The capability organisations name as their most acute problem is the one they are spending less to fix, which leaves them less able to govern cost as the hardware bill climbs.
- Almost half of UK enterprises have deprioritised legacy infrastructure modernisation (48%), team training (43%), and security and compliance reviews (39%), each one a discipline an AI factory leans on more heavily as it grows.

More than half of UK executives (59%) say their organisation automatically identifies root cause across all domains; among the engineers who field those alerts, only 34% report the same capability. Budgets gets approved against the executives' more confident view.

- That 25-point divergence means the people authorising infrastructure spend are working from a more confident picture of readiness than the people who answer the alerts at 2am.
- Only 26% of UK enterprises say their AI workload performance is highly predictable, meaning nearly three-quarters are planning capacity, SLAs, and cost against behaviour they cannot reliably anticipate.

70% of UK enterprises want the same two things first: unified visibility across the full stack and automated root cause analysis.

- UK enterprises rank infrastructure visibility as their second prerequisite for scaling, which tracks with the regulatory accountability UK organisations carry, where visibility is a condition of operating.

Section 1

Enterprise AI in the UK has entered its operational phase, and the organisations with the most AI activity are the ones asking their governance tooling to work the hardest. As AI factories scale, adding GPU clusters, expanding data pipelines, and onboarding concurrent workloads, the operational surface grows faster than the observability infrastructure managing it.

Most UK enterprises are past piloting and are running AI in production. That combination is where the absence of operational controls turns into a present risk to performance, cost, and reliability.

59%
Already scaling AI across teams

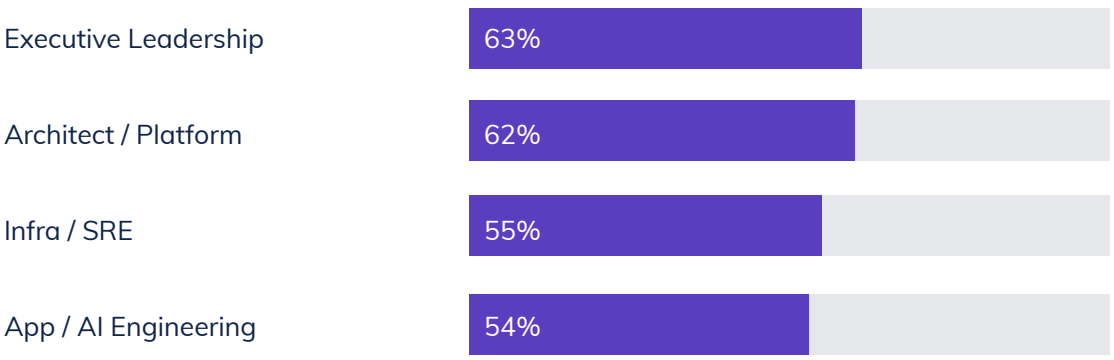
13%
Running production workloads while expanding infrastructure

17%
Operating early production workloads

Leadership confidence runs ahead of practitioner experience

The data reveals a stark divergence between leadership confidence and practitioner experience. Infrastructure and reliability engineers, the teams closest to daily workload operations, report scaling at 55%, eight points behind UK Executive Leadership at 63%. Leaders making AI infrastructure investment decisions are operating with a more confident view of organisational maturity than the teams managing those environments day to day. Investment decisions made against an overstated picture of operational readiness produce AI factories that are harder to instrument, govern, and prove.

UK AI journey stage, by role (% scaling across teams)



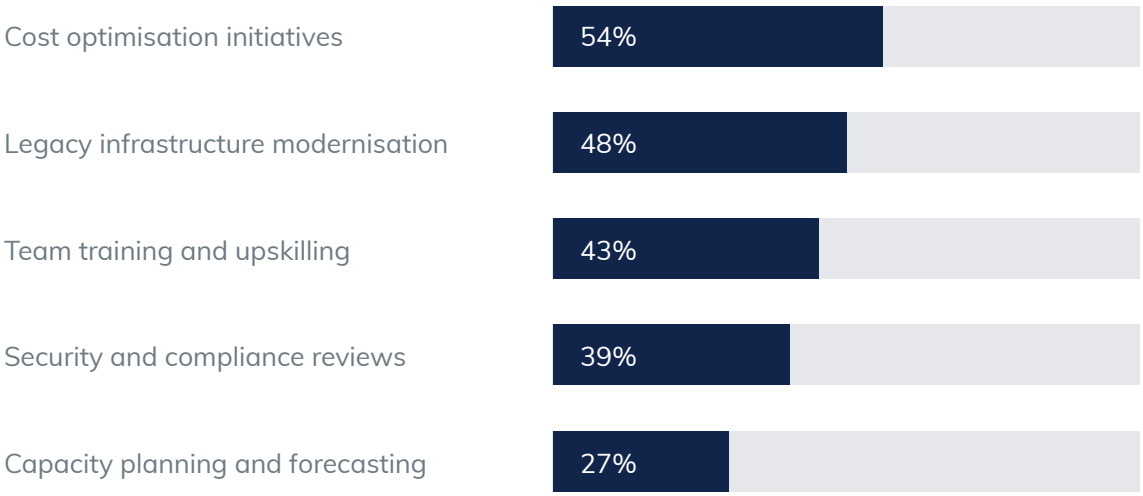
Percentages indicate respondents reporting their organisation is scaling AI across teams.

Scaling jumps once organisations pass $500M in revenue and holds in a 59% to 65% range across every larger band, peaking at 70% in the $1B to $3B range.

As AI factories scale, governance and controls are being deferred

As AI factory demands grow, the work required to establish visibility, accountability, and control is being sidelined. UK enterprises are racing to build out AI factories while reducing the governance and instrumentation work that would let them operate those factories with confidence. The deprioritisation of cost optimisation is especially consequential, because nearly half of UK enterprises name cost and efficiency metrics as their hardest monitoring challenges. Organisations are reducing their capacity to govern cost at the same moment they identify cost visibility as their most acute operational problem.

What UK teams are deprioritising as AI factory demands grow



Responses vary by role. Executives are mostly likely to deprioritise cost optimisation at 62% while the majority (59%) of infrastructure and SRE teams report they are deprioritizing. Executives are also more likely to report deprioritising security and compliance reviews at 46% followed by App/AI Engineers at 41%.

As AI regulations are expanding, security and compliance reviews are being deprioritised by 39% of UK enterprises.

Section 2

The Investment Picture Behind the Build-Out

The AI factory build-out is happening against a financial backdrop that's shaping every infrastructure decision. Premium AI hardware is in short supply and high demand, and UK enterprises are paying close attention to how every pound 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.

Hardware economics have become the gravitational centre of UK infrastructure strategy. The cost of premium compute is no longer a procurement line item handled in isolation; it is reshaping how organisations place workloads, justify spend, and sequence expansion. That pressure is now near-universal across the UK market.

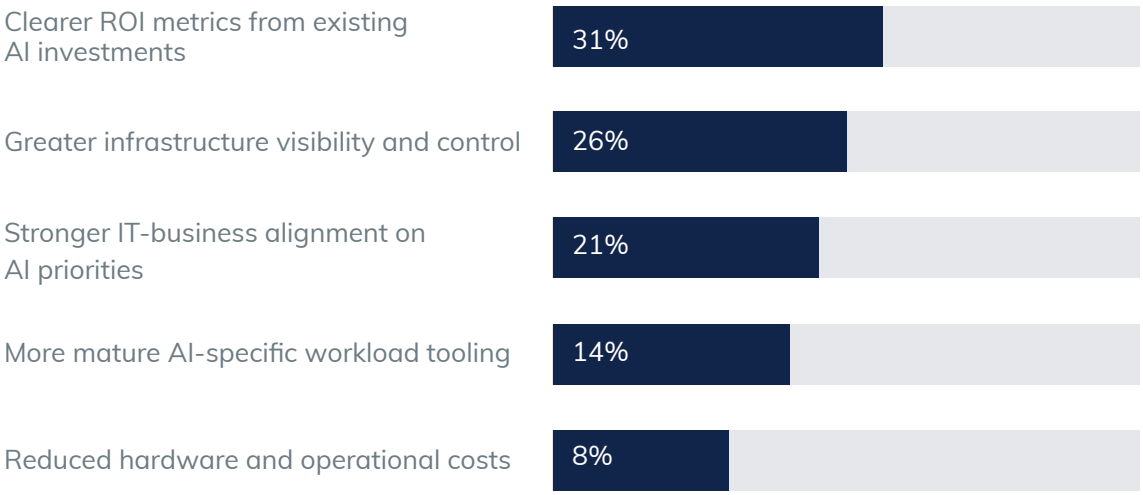
66% **of UK enterprises say the cost of premium AI hardware has changed how they approach investment decisions**

The ROI prerequisite

More than any other factor, UK enterprises identify clearer ROI as the primary prerequisite for scaling AI further, cited by 31% of respondents, ahead of infrastructure visibility, IT-business alignment, and mature tooling. What enterprises want is proof that the factory they have already built is working.

Infrastructure visibility ranks second at 26%, reflecting both the full-stack monitoring shortfalls and the regulatory accountability requirements that make operational visibility structurally necessary for UK organisations. Yet this requirement runs directly into a paradox. The organisational capabilities most required to generate ROI proof, namely cost visibility, performance predictability, and cross-domain accountability, are precisely the areas where UK enterprises are deprioritising investment. Organisations cannot prove AI outcomes because they are systematically reducing their capacity to observe and measure them.

What UK enterprises need to confidently scale AI in the next 12 months



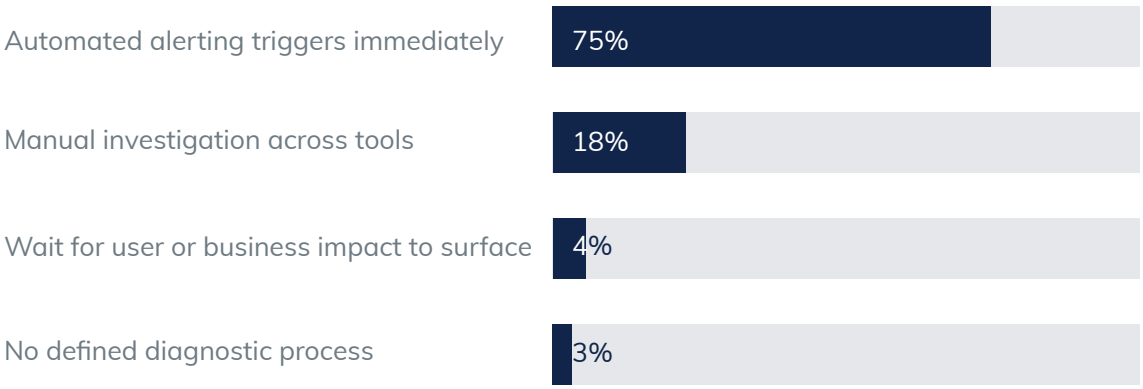
Section 3

Inside the UK AI Factory, Where Control Breaks Down

The operational consequences of scaling AI factories faster than the systems to operate them are visible in how UK AI factories operate at scale. They emerge in how teams respond to failures, predict workload behaviour, understand resource utilisation, and maintain visibility across the full execution stack. For organisations running dedicated GPU clusters and high-performance AI infrastructure, every hour of unresolved workload degradation is an hour of underperformance against the hardware's potential.

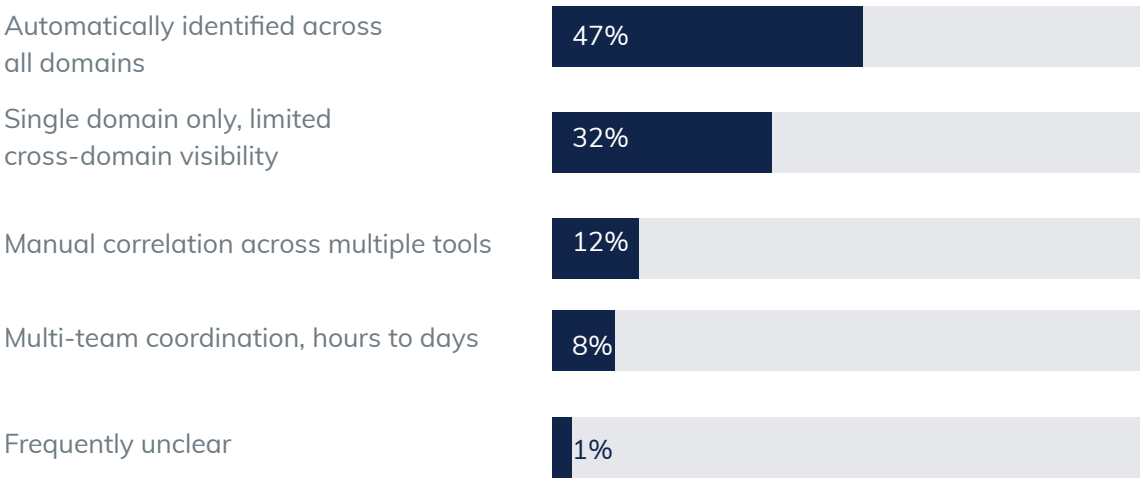
Most UK enterprises have automated alerting as their first response when an AI workload underperforms or fails. A meaningful minority still rely on slower paths.

When an AI workload underperforms or fails, what is your team's first response?



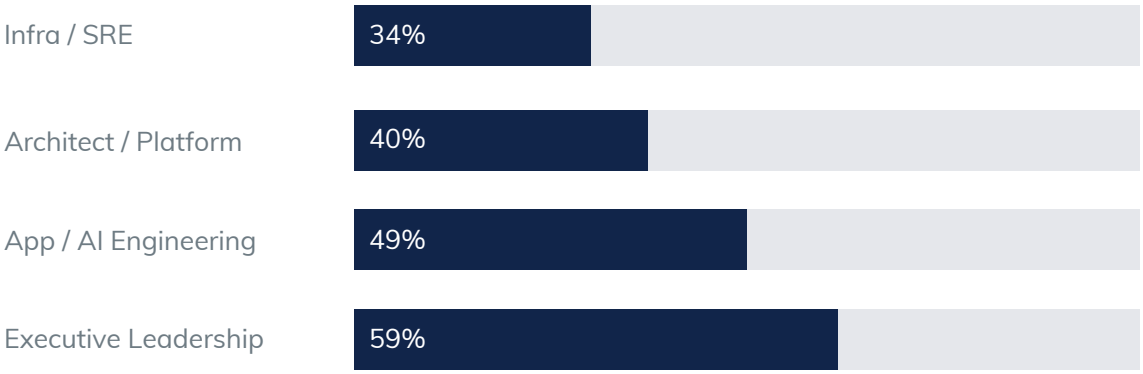
However, detecting a problem is not the same as explaining it. More than half of UK enterprises cannot automatically identify root cause across all infrastructure domains when an alert fires. 12% require manual correlation across multiple tools, and 8% require multi-team coordination taking hours to days, this is roughly 20% of UK enterprises that are operating with significant exposure when AI workloads fail.

Reported root cause capability when an alert fires



Underneath that average, the overall picture splits sharply by role. UK infrastructure and SRE practitioners, the engineers who field alerts, report automated root cause capability at 34%, while UK executives report 59%. This 25-point difference reflects what it means to be the person on call when a GPU workload degrades with no automated diagnosis to work from.

UK: 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 lies the more persistent challenge of not knowing how AI workloads will behave from one run to the next. Performance variability makes capacity planning approximate, SLA commitments unreliable, and cost forecasting a guessing exercise. Predictability is what turns a hardware investment into a measurable operational capability.

26%

Describe AI workload performance as highly predictable

23%

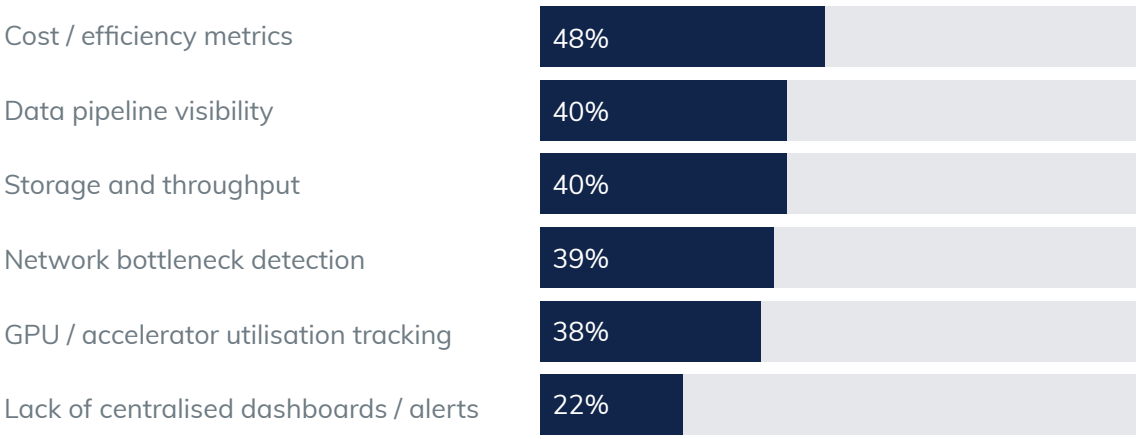
Highly predictable among Infra/SRE practitioners

Only 26% of UK enterprises describe AI workload performance as highly predictable overall, leaving nearly three-quarters without predictable performance to plan against. Infra/SRE practitioners, responsible for day-to-day workload operations, report predictability of just 23%, among the lowest of any role group.

What teams can and cannot see when it comes to AI factory monitoring

Cost and efficiency metrics, data pipeline visibility, and storage and throughput are the top monitoring challenges across the UK dataset. Visibility problems reach every layer of the stack, which marks this as a structural condition that spans the whole AI factory.

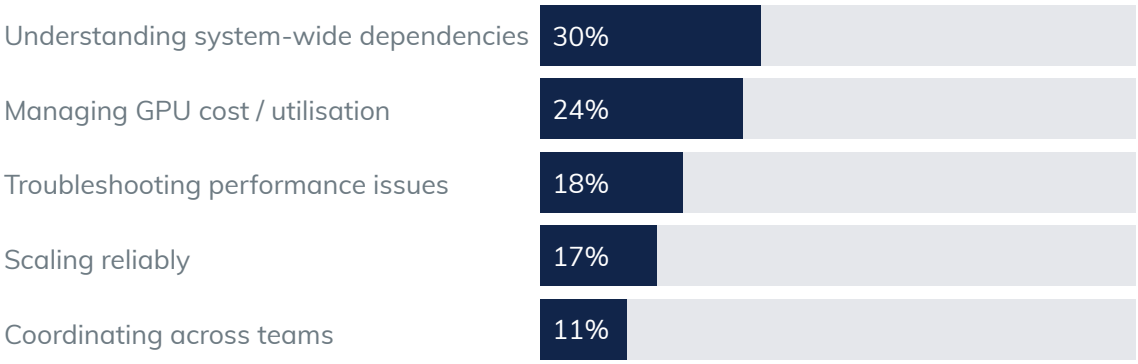
Monitoring challenges across the AI factory stack



Each team encounters a different face of the same full-stack problem. UK Infra/SRE teams lead on cost and efficiency metrics (59%) and network bottlenecks (50%), reflecting where daily operational pressure lands. Architects lead on storage and throughput (48%) and data pipeline visibility (48%), reflecting the supply chain view of the AI factory. Executives and App/AI Engineering both put cost and efficiency metrics at the top (44% and 46%), the one challenge no role escapes, even as their secondary pressures diverge by where they sit in the stack.

When asked what the single hardest aspect of operating AI infrastructure today is, UK enterprises rank understanding system-wide dependencies first followed by managing GPU cost and utilisation.

Hardest aspect of operating AI infrastructure today



Section 4

The Regulatory Dimension: A UK-Specific Governance Exposure

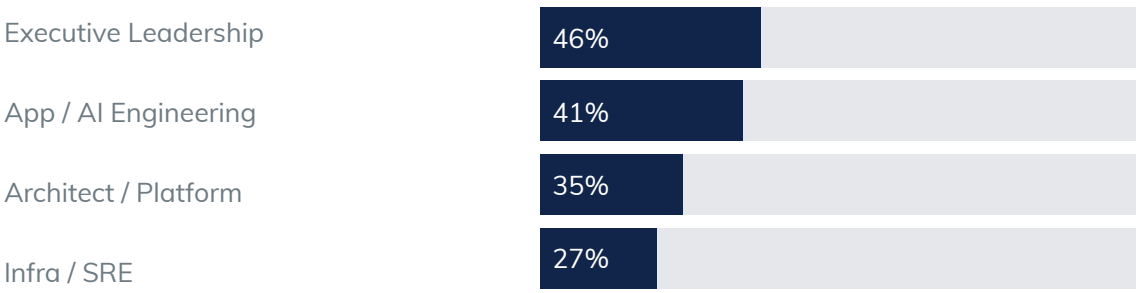
In the UK, the system that proves an AI factory works is the same system that satisfies a regulator.

UK enterprises operate within a regulatory environment that amplifies the cost of deferred governance. UK GDPR establishes accountability and audit obligations over how personal data flows through AI systems. Emerging AI Act obligations, extending to UK-adjacent supply chains and internationally operating organisations, introduce requirements around transparency, risk classification, and operational documentation that AI infrastructure observability directly supports. Sector-specific frameworks in financial services, healthcare, and essential national infrastructure compound these requirements further.

Despite increasing regulations, 39% of UK enterprises are deprioritising security and compliance reviews as AI factory demands grow. Executives are the most likely to report deprioritizing security and compliance at 46% contrasting sharply with only 27% of infrastructure and SRE teams the least likely to deprioritise at 27%. For organisations subject to these obligations, the consequence is a material compliance exposure that accumulates as AI factories scale and the operational surface subject to regulatory scrutiny expands.

The instrumentation required to demonstrate regulatory compliance, including audit trails, cost attribution, and cross-domain accountability, is the same instrumentation required to operate AI factories at scale. UK organisations operating under financial conduct, data protection, or critical systems regulation cannot separate the question of operational observability from the question of regulatory accountability. The systems that provide operational confidence are the same ones that satisfy regulatory obligations. That makes a lack of observability a compound risk, with operational and regulatory exposure accumulating in parallel.

Security and compliance reviews deprioritised, by role



This is where sovereignty becomes operational. Cloud and AI sovereignty are often framed as questions of where data and models reside. For UK enterprises, sovereignty resolves into a more immediate question of whether an organisation can see, attribute, and prove what its AI systems are doing across every environment it runs in. Sovereignty without observability is a claim that cannot be evidenced. Unified visibility across on-premises, virtualised, and public cloud environments is what turns a sovereignty commitment into something an organisation can demonstrate to a regulator, an auditor, or its own board.

Section 5

Observability and Automation Are the Operational Foundation for AI Infrastructure at Scale

When asked what would have the most immediate impact on their ability to scale AI operations, UK enterprises converge on 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%). Together, these account for 70% of UK respondents.

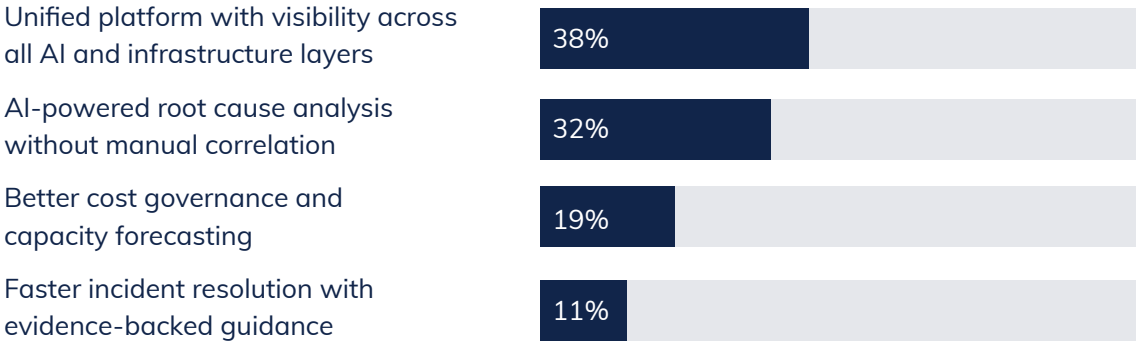
70%

of UK enterprises identify unified visibility plus automated root cause analysis as their most immediate operational priority

Infrastructure visibility ranks second as a scaling prerequisite in the UK at 26%. This reflects both the full-stack monitoring shortfalls organizations are experiencing and the regulatory accountability requirements that make operational visibility structurally necessary for UK organisations.

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 optimise.

Most immediate operational priorities



Without unified visibility across the full stack and AI-powered root cause analysis that eliminates manual correlation, UK organisations cannot:

- Demonstrate whether AI-driven services are meeting performance, reliability, and SLA commitments.
- Control and optimise the cost and utilisation 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 and operationally sound.
- Satisfy the governance, accountability, and audit obligations that UK regulatory frameworks require.

Until UK 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 than they can operate them or prove their ROI.

Conclusion

UK enterprises have built their AI factories, and the work remaining is the systems to operate, govern, and prove them. In the UK, those are one system, where operational proof and regulatory accountability meet. The same instrumentation that explains a 2am failure to the engineer on call is the visibility that satisfies an auditor and evidences a sovereignty commitment. Until UK organisations can see, attribute, and prove what their AI systems are doing across every environment they run in, they will keep scaling faster than they can operate, govern, or account for what they have built.

How Virtana Helps

Virtana's AI Factory Observability provides the visibility, governance, and accountability layer for the AI factories UK enterprises are building today. By correlating model behavior, token consumption, GPU utilization, infrastructure health, service dependencies, and cost across on-premises, virtualized, and public cloud environments, Virtana delivers system-aware observability for AI at scale. Its agentic AI identifies constraints, proves root cause, and surfaces opportunities to improve performance, efficiency, and ROI, enabling operations teams to optimize AI factories while giving business and technology leaders the insight needed to govern them effectively.

METHODOLOGY

Demographics

This study is based on responses from 238 UK-based professionals at enterprise organisations 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 organisations.

By role

Executive Leadership	71
Application, Service & AI Engineering	63
Architect & Platform Design	60
Infrastructure, Cloud & Reliability Engineering	44

By annual revenue

$500M – $1B	68 (29%)
$3B – $5B	37 (16%)
$1B – $3B	33 (14%)
$5B – $7B	31 (13%)
$10B+	24 (10%)
Under $500M	23 (10%)
$7B – $10B	22 (9%)

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**

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