Akamai

The State of AI Inference 2026

A letter from Akamai

This report is built on the direct experience of architects and engineers managing live AI inference workloads in production today

As AI moves from experimental labs into latency-sensitive customer journeys, it's exposing the mismatch between the requirements of real-time AI and the limitations of infrastructure built for a previous era.

Inside, you'll see how teams are using rollbacks, rerouting, and re-engineering as inference workarounds, and what's required to move toward systems that truly scale.

Mo Tabares
VP of Engineering, Akamai

Executive summary

Inference has moved from experimentation into live business operations. Success is no longer just about model quality. It's about delivering responses fast enough, reliably enough, and consistently enough to support real customer experiences and operational decisions.

Organizations increasingly recognize that meeting those demands will require inference to run closer to users, data, and decision points. But infrastructure is changing more slowly than AI itself. As a result, many teams are operating in the gap between what real-time AI requires and what current architectures can support.

To compensate, teams are relying on tighter operational controls. They're steering traffic, managing failover paths, and building more resilience into runtime behavior. These measures help keep inference moving, but they also add complexity. This makes cost, performance, and governance harder to manage at scale.

This report examines this transition by tapping into the experiences of technical leaders running inference today. It shows where production inference is creating new pressure, how teams are adapting, and what the next phase of AI infrastructure will demand.

Contents

Methodology

This report is based on a global survey fielded in March 2026 and designed by Adience and Akamai. Two hundred respondents participated. All were individuals who influence or make decisions about the deployment or operation of AI inference in their organizations.

The audience was primarily technical, and 76% served as primary or secondary decision-makers for AI deployment. While this initial study focuses on North American and European markets, the technical requirements for real-time AI remain a global priority. All statistics referenced in this report are derived from the March 2026 survey, unless otherwise noted.

Region	%	N
United States	**35%**	**70**
United Kingdom	**25%**	**50**
Europe (excl. UK)	**23%**	**45**
Canada	**18%**	**35**

Size (Employees)	%	N
10–99	**23%**	**45**
100–499	**51%**	**102**
500 or more	**27%**	**53**

Job (Title)	%	N
Engineer	**67%**	**133**
Architect	**18%**	**36**
VP/Director/Head	**17%**	**31**

Percentages may not total 100% due to rounding

adience

Adience is a dedicated B2B market research agency that rebels against boring, cookie-cutter traditional research. Drawing on decades of experience across SaaS markets, IT, and other B2B sectors, Adience delivers trusted answers to the questions that matter, so decision-makers can act with confidence.

01

The requirements of real-time AI

From experiments to experiences

For most enterprises, AI is past the experimental phase. Models that were recently being piloted — or simply evaluated — are now embedded in live customer journeys, business workflows, and time-sensitive decisions.

That changes the standard.

When inference sits inside a live business moment, success is no longer about generating the right output eventually. It's about generating the right output fast enough, reliably enough, and consistently enough to keep the experience intact.

When AI was just a side project, a slow or localized tool was acceptable. Today, AI is being woven into the core of live operations. In this context, answers are only as valuable as the speed at which they're delivered.

Among the organizations we surveyed, 82% say their most important use cases require end-to-end response times of 500 ms or less. For 64%, even that is too slow. They need end-to-end response times of 250 ms or less.

64%

require end-to-end response times of 250 ms or less

Under these new standards, speed has to be paired with consistency.

Today's leaders are moving beyond average performance and focusing on what users actually experience under real conditions. Thirty-nine percent prioritize p95, and another 17% prioritize p99. More than three quarters expect 99.9%+ availability.

For advanced operators, these are not separate goals; they're a combined baseline. More than a third of global enterprises now demand the three Rs of real-time AI.

The three Rs of real-time AI [1]

1. Responsive

End-to-end response times under 250 ms

2. Reliable

Performance maintained at the 99th percentile

3. Resilient

No more than 0.01% downtime

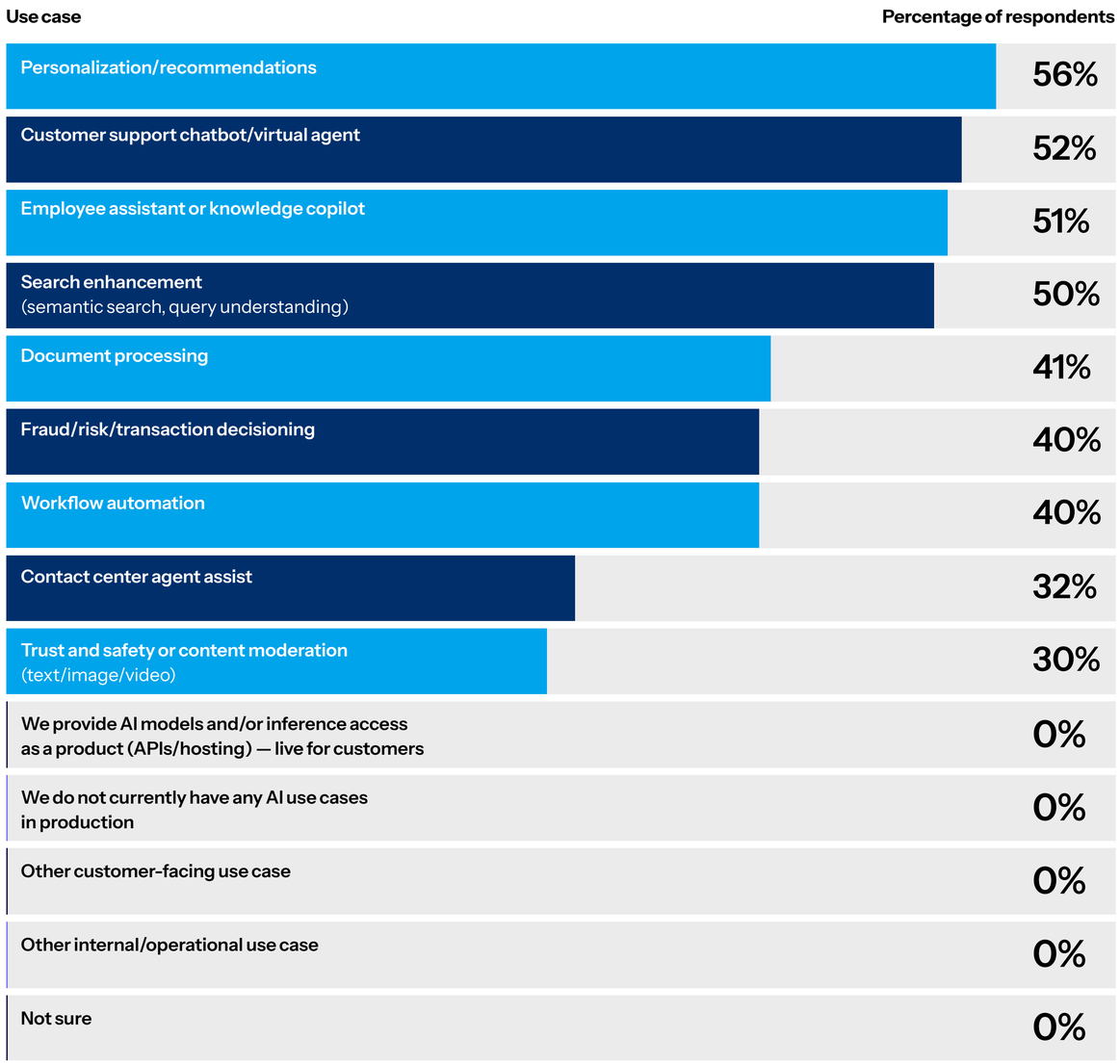
With inference now driving business metrics, lags become lost opportunities.

Inference is spreading across the enterprise

The debate over AI's utility is settled. Every organization in this survey has AI in production. What's really striking is the breadth of that adoption. Whether it's personalization, customer support, internal copilots, or search, inference is no longer a specialized experiment. It's become a standard toolkit for both customer-facing and operational work.

1. These benchmarks represent what the specified respondents identified as requirements for their most important use cases

Which of the following AI use cases are currently in production at your organization?



Performance drives profitability

When AI performance lags, profitability suffers across three critical fronts: it erodes customer satisfaction, inflates cost to serve, and suppresses revenue per visitor. In short, if your inference isn't performing, your business isn't either.

The business metrics most affected by inference performance, according to respondents, are shown below.

Business metrics most affected by inference performance

Metric	Percentage of respondents
Customer satisfaction	49%
Cost to serve	40%
Revenue per visitor	39%

Where speed matters most

Performance matters across the board, but it moves the needle most aggressively in real-time workflows. When AI sits behind a live customer interaction or a split-second decision, the margin for error vanishes. This is why organizations demand rapid responses for personalization, fraud or transaction decisioning, search, customer support, and trust and safety.

In these high-pressure moments, success is binary. Inference is either fast enough to capture the opportunity or it isn't. And if it's not, both trust and revenue will be lost.

This risk only grows as companies mature. Compared with organizations that have not yet deployed GenAI in production, businesses with GenAI already in production are 45% more likely to face these speed-critical scenarios. Organizations with 500+ employees are 44% more likely to run latency-sensitive workloads than smaller competitors. Scale does not just bring complexity — it makes speed essential.

Of the AI use cases you have in production, which require a rapid response from inference to work well?

Use case	Percentage of respondents
Personalization	54%
Fraud, risk, or transaction	49%
Search	44%
Customer support	41%
Trust and safety, or content moderation	40%

> **The majority of AI spending in 2026 will be on inference workloads**[2]

2. www.gartner.com

Supporting the rise of inference

There's a lot on the line. Inference is essential, but supporting these workloads is not straightforward.

Current market discourse often frames AI around ROI skepticism. But among the technical and operational leaders surveyed here, the debate regarding value is effectively over. Only 2% say doubts over ROI are holding AI back. The real barriers are infrastructure limitations, integration complexity, and risk.

Compared with peers that do not yet have GenAI in production, businesses with active production deployments are:

- 36% more likely to track unit economics for some use cases
- 14% less likely to be constrained by vendor lock-in
- 31% more likely to classify traffic steering as very important or critical
- 38% more likely to require rollback times of 15 minutes or less

That is the first clear sign of operational maturity. The challenge is no longer proving the model. It's building an environment that can support rising inference requirements without breaking.

02

The reality of centralization

Proximity takes on more importance

The destination for the next era of AI is clear: the edge.

As inference begins to power more high-stakes business moments, organizations are reaching a solid consensus. They know that to win on speed, they have to move the engine closer to the action. A commanding 60% of respondents now view proximity to the end user and decision point as a make-or-break factor.

Architecture has barely moved

There's a striking tension between where businesses want to be and where they're currently stuck. On the surface, the average organization appears to be standing still, with nearly half remaining tethered to a single, centralized cloud region for the next two years.

Despite the growing consensus around proximity, most organizations remain anchored to centralized deployment patterns. Today, 46% run inference in a single centralized cloud region. Looking at plans for the immediate future, that figure barely changes, landing at 45%. Multiregion cloud shifts from 22% to 20%, while hybrid cloud, on-premises, and edge remain flat at 19%.

That is the core contradiction in our data. The requirement is getting stricter, yet architecture is barely changing.

60%

say proximity to end users and decision points is important or critical

Where do you run inference today? Where will your most important use cases run in 1–2 years?

Deployment model	Today	In 1–2 years
A single, centralized cloud region	46%	45%
Multiregion cloud	22%	20%
A hybrid of cloud, on-premises, and edge	19%	19%

The average hides the divide

The figures above do not apply to all businesses equally.

When we look at global organizations, the picture shifts dramatically. These leaders are 22% more likely to view proximity as a critical requirement, and they are the ones breaking away from the status quo. For global enterprises, a centralized cloud isn't so much an anchor as a bottleneck. While the rest of the market waits, the most mature players are already architecting for a distributed future.

Global organizations are more than 2x as likely to be moving toward multiregion cloud or hybrid distribution

The business case is settled

The business case for distributed inference is no longer up for debate. A full 98% of organizations recognize its value. Yet most remain stuck in the middle, held back not by a lack of vision but by technical debt, tooling immaturity, and operational fear.

The case is clear:

98%

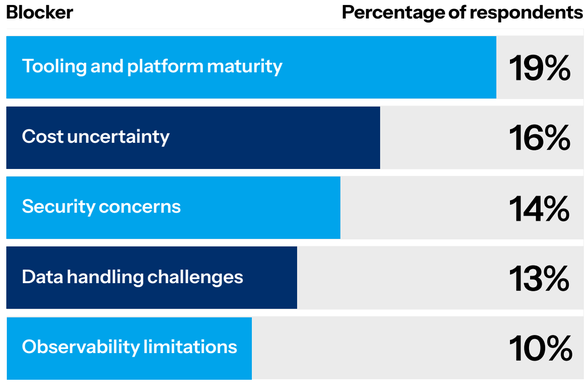
of organizations recognize the business value of distributed inference

The blockers are practical

The jump from centralized to distributed infrastructure is a transformation many teams don't feel ready for.

One in five respondents say they lack the platform and tooling maturity to make the move. Others cite cost uncertainty, security concerns, data handling challenges, and observability limitations. These are not ideological objections but practical barriers to change.

What is the single biggest blocker preventing more distributed/edge inference in your organization today?



The comfort of centralization is fading

Centralization remains the safe choice because it is familiar. But that comfort is cooling fast.

Already, 20% of businesses say they lack confidence that centralized inference will continue to meet future requirements. That is still a minority, but it is a meaningful one. Business needs are increasingly latency sensitive. Deployment remains cloud-centric. And the gap between those two facts is getting harder to ignore.

03

The rollbacks and reroutes

The control layer takes over

Business requirements have changed for good. Architecture has yet to shift. That means the adaptation is happening somewhere else: in operations.

Instead of relying on a distributed architecture to guarantee performance, teams are protecting runtime through more active control. They are steering traffic, tightening rollback windows, retrying degraded requests, and sending workloads to fallback paths when performance slips.

Speed of correction becomes critical

If teams cannot always rely on static infrastructure to perform, they need to move fast when things start to break.

More than 70% say they need to enact rollback within an hour. Nearly half push that much further and name 15 minutes as the acceptable window. Production inference is no longer being measured only by how well it performs when everything goes right. It's also being measured by how fast teams can recover when something goes wrong.

Traffic steering becomes mandatory

When infrastructure is not inherently optimized for every user and every workload, traffic has to be directed more intelligently across providers, regions, and models. That's why 64% of respondents classify traffic steering as very important or critical.

64%

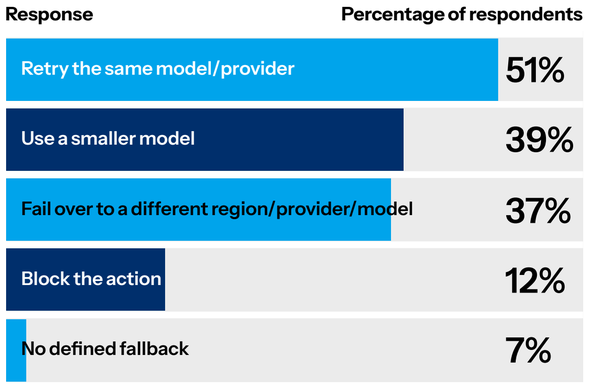
of businesses classify traffic steering as very important or critical

Teams are patching what architecture cannot fix

Organizations with predominantly centralized infrastructure are increasingly running into an unfortunate fact: Those systems were not built for rapid or real-time AI inference.

Rather than rebuilding immediately, teams are generally plugging gaps through more active control. When inference degrades, respondents say their systems most often retry the same model or provider; use a smaller model; fail over to a different region, provider, or model; block the action; or have no defined fallback.

When inference degrades, what do your systems do?



The majority are not redesigning their inference architecture yet. Instead, they're reinforcing its weak points.

Mature operators depend on control even more

Businesses that already have GenAI in active production are not bucking this trend. They're at the forefront of it.

Compared with trailing competitors, they are 31% more likely to classify traffic steering as very important or critical and 38% more likely to require rollback times of 15 minutes or less. This shows that the more inference matters, the more organizations depend on runtime control.

04

The cost of complexity

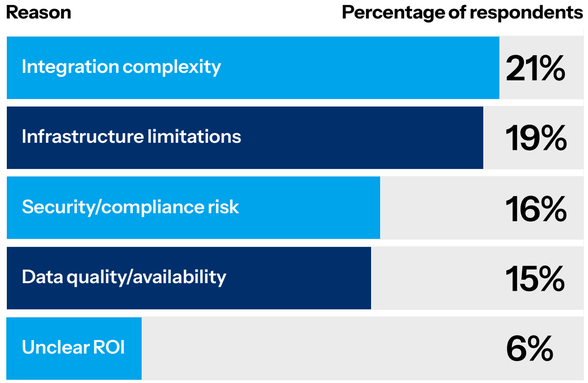
Control buys time, but it's becoming a barrier

Routing, rollback, retries, and fallback all make systems more resilient in the short term. But they also make them more complex.

That complexity is already showing up as a leading barrier to scale. When respondents were asked for the single biggest reason AI has not scaled further in their business, the top answer was integration complexity (21%), followed by infrastructure limitations (19%), security and compliance risk (16%), and data quality and availability (15%). Unclear ROI trailed far behind (6%).

The market is not stalling because the value of AI is unclear. It's stalling because running inference at scale is operationally hard.

What is the single biggest reason AI has not scaled further in your business?



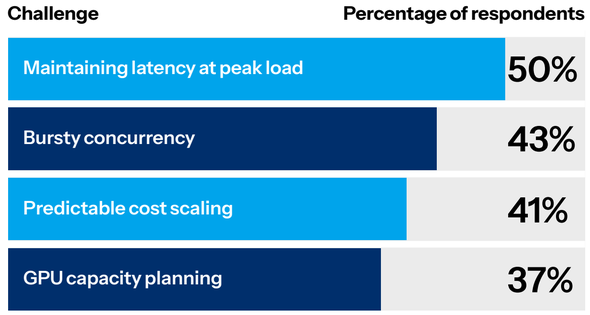
The challenges of meeting inference demand

A full half of respondents say maintaining acceptable latency at peak load is one of the hardest scaling challenges in production inference today. Meanwhile, 43% struggle to deal with sudden spikes in traffic.

50%

of organizations see maintaining latency at peak load as one of their biggest scaling challenges

Which scaling challenges are hardest for production inference today?



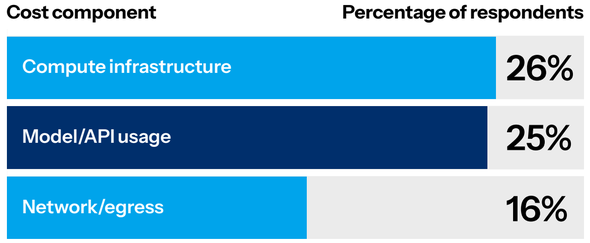
Flexibility creates unpredictability

As we have seen, organizations are not yet moving their infrastructure to the edge. Instead, to compensate for the distance between compute and users, they've built operational resilience based on flexibility.

That helps them meet strict performance requirements. But if execution paths are not fixed, neither are costs.

When respondents were asked which cost components are hardest to predict, the leading answers were compute infrastructure (26%), model or API usage (25%), and network or egress (16%). The result is an uncomfortable trade-off: more operational flexibility, less economic clarity.

Which cost components are hardest to predict?



Most teams still lack cost visibility

Scaling a business should not be a guessing game. Yet 77% of organizations are effectively flying blind, lacking the consistent unit economics tracking needed to see where their money is actually going as they grow.

Without mature unit economics tracking, cost outcomes can be understood retrospectively — but not understood or controlled at runtime. The organization may be improving performance through dynamic controls, but without visibility, it cannot tell whether it's scaling efficiently or just expensively.

77%

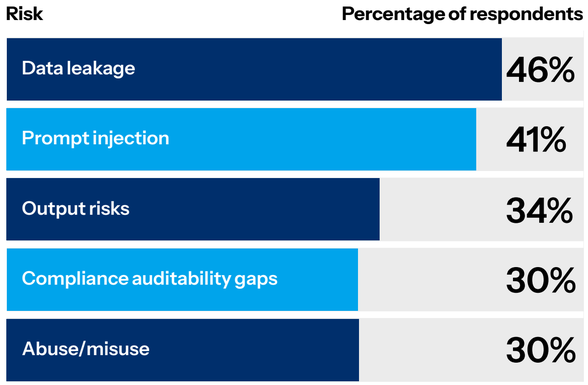
lack consistent unit-level economics tracking

Variability expands the governance burden

Variability is not just an issue for price. It is also an issue for governance. Given that a model's output isn't set, inference requires continual governance at runtime. That's a major challenge.

Respondents name the top perceived security and governance risks for production inference as data leakage (46%), prompt injection (41%), output risks (34%), compliance auditability gaps (30%), and abuse or misuse (30%). As inference gets routed across more providers, regions, and models, the need for policy, observability, and enforcement grows with it.

What are your top perceived security and governance risks for production inference?



05

Systems that scale

Finding a way forward

At this point, the market's center of gravity has shifted.

The question is no longer whether inference matters but how to support it without breaking under pressure. With only 2% of respondents skeptical of the business case, that debate is closed. What organizations need to decide on now is their infrastructure strategy.

Maturity will not look like edge everywhere overnight

Architectural transitions do not happen at the speed of model releases. They happen gradually, and usually only when operational pain becomes impossible to ignore.

Mature operators are pairing a gradual distribution of compute with tighter controls over both performance and cost. That is the shape of the next phase. It's not a dramatic overnight migration but instead a steady move away from centralized, single-region inference — one that offers better control over runtime, economics, and governance.

The leaders are already signaling what comes next

The most mature organizations are building toward a different operating model:

- More unit-economics tracking
- Less vendor lock-in
- More traffic steering
- Faster rollback
- Tighter operational control

This is what maturity looks like in inference today.

The next wave is selective distribution

As inference use cases continue to spread, the need to place compute closer to users and decisions will become harder to avoid. But the organizations that win will not be the ones that simply distribute everything. They'll be the ones that know what to distribute, where to place it, how to route it, and how to govern it.

This will be the defining feature of the next phase of systems architecture: Better-performing AI will be the result of better-placed inference.

The case is closed

98% recognize the value of distributed inference. The question now is how fast the infrastructure can catch up.

You have hard AI problems. We have people who know AI. Let’s talk.

Talk to architects and engineers who have done this before. We’ll help you run your AI in production.

BOOK A CONSULT


Akamai